Collecting and organizing systematic sets of protein data

John G. Albeck[*‡], *Gavin MacBeath*[§], *Forest M. White*[*], *Peter K. Sorger*[*‡], *Douglas A. Lauffenburger*[*] *and Suzanne Gaudet*[*‡]

Abstract | Systems biology, particularly of mammalian cells, is data starved. However, technologies are now in place to obtain rich data, in a form suitable for model construction and validation, that describes the activities, states and locations of cell-signalling molecules. The key is to use several measurement technologies simultaneously and, recognizing each of their limits, to assemble a self-consistent compendium of systematic data.

Systematic set
A data set in which all data are collected from the same experimental system in such a way that all data can be directly compared, regardless of when measurements were made.

Signal
Any type of biomolecule that transfers information in a signalling network.

*Center for Cell Decision Processes, Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.
‡Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA.
§Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA.
Correspondence to D.A.L.
e-mail: lauffen@mit.edu
doi:10.1038/nrm2042

The goals of systems biology are orthogonal to large-scale efforts to catalogue genomes, proteomes and interactomes. Rather than seeking a broad knowledge of biological components and their functions, systems biology seeks a deep understanding of biological processes in quantitative terms (FIG. 1). Therefore, 'omics' and systems biology complement each other, but data collection strategies in each field are entirely different. Biological processes are not static or homogeneous in space. Systems biology therefore requires dynamic, spatially resolved data on gene and protein function in specific cell types in response to specific stimuli. The complexity of this type of 'rich' data is extremely large and it cannot simply be collected by steady, hypothesis-independent accumulation (as was possible in sequencing the human genome). Instead, data must be collected with reference to specific, quantitative models.

This article is a guide to collecting and organizing systematic sets of rich, multivariate data that characterize how signals in regulatory pathways change in time and in space. For simplicity, we limit our discussion to biochemical changes (particularly protein phosphorylation) found in eukaryotes. For the detection and characterization of nucleic acids and metabolites, readers are referred to other reviews[1–8].

Rarely can a model be constructed solely from data that are reported in the literature. Here, we assume that readers are interested in making their own measurements and assembling new data sets. We review recent developments in three measurement technologies that are important for systematic data collection: affinity-based assays, physical assays and enzymatic assays. We also discuss approaches to quantifying signals in single cells and the necessity of combining population-based and single-cell data. Last, we explore methods to transform raw data into reliable measurements, organize and check the quality of data during and after data collection, and fuse heterogeneous measurements into a compendium.

Considerations in biological measurement

Although we always strive to collect more comprehensive data, five practical considerations dominate the design of actual data-collection efforts.

First, matching data and a model is rarely simple. It is usually said that sensitivity analysis can be used to prioritize parameters in a mechanistic model for measurement (see the accompanying Review by Aldridge, Burke, Lauffenburger and Sorger in *Nature Cell Biology*), but many of the most sensitive parameters cannot be measured using existing reagents and technology. When modelling, we must be conscious of what can and cannot be measured, and with what degree of precision. Modelling yields a wealth of hypotheses, but only a subset of these — and not necessarily the most interesting — might be testable. Because measurement is the most limiting aspect of a systems analysis, approaches to modelling are usually designed to accommodate possible measurements, and not the other way around.

Second, practical trade-offs must be considered when selecting what to measure and how frequently to measure it (see Supplementary information S1 (box)). Reagents are expensive, validation is time-consuming and experiments can only be repeated a limited number of times. We must choose between performing replicates of a single experiment, sampling more densely in time, exploring multiple perturbations, or looking at more proteins. Even a twofold increase in the number of measurements is significant when hundreds of tissue-culture plates are involved. One effective strategy is to do a preliminary experiment in which multiple axes in data space (for example, signals, perturbations and

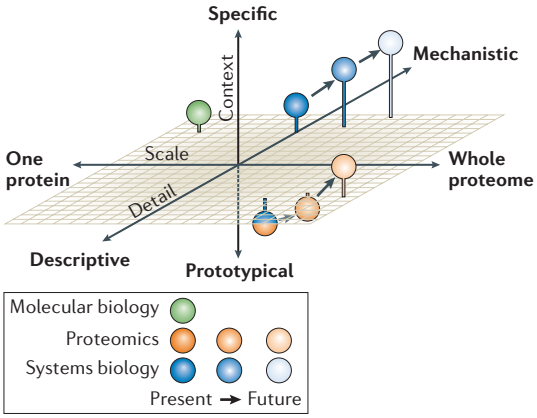


Figure 1 | **The scope of systems biology.** The approaches of molecular biology, proteomics and systems biology are compared with respect to three factors: scale (ranging from a single protein to the whole proteome); level of detail (ranging from descriptive to mechanistic); and context (ranging from a prototypical cell to a specific cell type). Molecular biology studies have traditionally focused on one or a few proteins with a highly mechanistic level of detail, often in specific cellular contexts. By contrast, proteomics seeks to catalogue many proteins, with the eventual goal of covering the entire proteome. So far, proteomic studies have largely been descriptions of the prototypical cell, but they are now moving towards more specific cellular contexts. Systems biology seeks a mechanistic understanding of phenomena that involve many proteins, but not the entire proteome; these phenomena are often specific to a cellular context. As time progresses, systems biology studies will encompass larger numbers of proteins with greater mechanistic detail in increasingly specific cellular contexts.

time points) are explored sparsely so as to estimate their information content (either by inspection or by formal analysis (see below)) and the overall performances of the measurement techniques are determined. Subsequent experiments can then focus mainly on the axes with the highest information content, with more limited sampling of other axes to ensure that nothing important is missed.

Third, protein-based data compendia are inevitably composed of heterogeneous measurements that require fusion. Whereas a single methodology, such as oligo-based microarrays, is, in principle, sufficient to measure every mRNA species in a cell, no single method can measure the full diversity of protein signals. Technical trade-offs exist among measurement approaches, including: breadth versus depth, low versus high throughput, small versus large sample size, ease versus difficulty, fixed versus live cell and, last, single-cell versus population measures (FIG. 2; Supplementary information S2 (table)). The heterogeneity of protein assays demands an efficient approach to data fusion. Results that arise from several different measurement technologies must be combined, and data that have been acquired over time must be merged into self-consistent sets for which one cannot simply rely on direct comparison to contemporaneous controls.

Fourth, the collection of signalling measurements must be hypothesis driven. The idea that complex, context-sensitive, biological measurements can steadily be accumulated without regard for their eventual use is wrong in principle and in practice. The scope of a set of measurements must match the hypotheses being tested. For example, a statistical model that is aimed at uncovering new relationships among disparate signalling modules requires sparse but broad sampling, whereas a physicochemical model requires denser and narrower sampling. The process of measurement is itself often dependent on specific hypotheses. For example, we might be able to measure the levels of phosphorylation on a critical kinase, but not the biochemical activity of the kinase; activity could then be inferred, but only in the context of a specific hypothesis about the mechanism of regulation.

Fifth, population-based and single-cell measurements are highly complementary. Most mathematical models of pathways are representations of reactions that take place in a single cell (see below). Single-cell methods provide important information on the mean and variance of cell responses, but allow for far fewer signals to be measured. The most effective strategy is to combine single-cell and population-based measurements using quantitative data models.

Affinity-based assays in systems biology

Affinity-based methods, most commonly using antibodies, are a mainstay of protein measurement. Immunoblots are the standard method for determining protein abundance and state of modification[9,10], but they are low throughput and difficult to automate (FIG. 2). Several other affinity-based assays have been commercialized, however, each has its advantages and disadvantages. Affinity-based assays are typically capable of measuring low-abundance proteins in small samples (typically ~10^4 cells) with high throughput (hundreds of samples per day). Antibodies and other affinity reagents are used in either of three formats: affinity-based capture, direct affinity-based detection and sandwich detection (BOX 1). The first two methods require a single antibody, whereas sandwich methods require separate capture and detection antibodies. Regardless of format, chemiluminescence, fluorescence or colorimetric methods can be used as readouts.

Classic enzyme-linked-immunosorbent assay (ELISA) techniques rely on the adsorption of antigens or antibodies to solid phase substrates, followed by enzyme-based detection. However, sandwich methods that use fluorescent detection in 96-well or 384-well plates (for which ELISA is really a false moniker) are the primary means for quantifying signalling proteins and their modifications in microplate format. Sandwich assays for many proteins are now available, and typically have good selectivity and sensitivity (BOX 1). Sample sizes are small and throughput is good, but the degree of multiplexing is low. Detection of very low-abundance proteins can be enhanced by complex detection schemes such as rolling circle amplification[11] and specialized plate materials (such as electroluminescent detection[12]).

Perturbation
Any experimental condition that is applied to a cell that causes a shift in the cell's behaviour away from the basal state. This includes extracellular stimulation by physiological ligands, inhibition of protein activities by small molecule inhibitors, or alterations in protein-expression levels by RNA interference or overexpression.

Immunoblot
Also known as a western blot. Following gel-based separation by mass, charge or both, proteins are transferred to a membrane and probed with target-specific antibodies.

Enzyme-linked-immunosorbent assay
(ELISA). ELISAs involve adsorbing or coupling capture antibodies to a 96-well plate. Following protein capture, a target protein is detected, either directly (if it was labelled in the sample) or indirectly, through a labelled detection antibody.

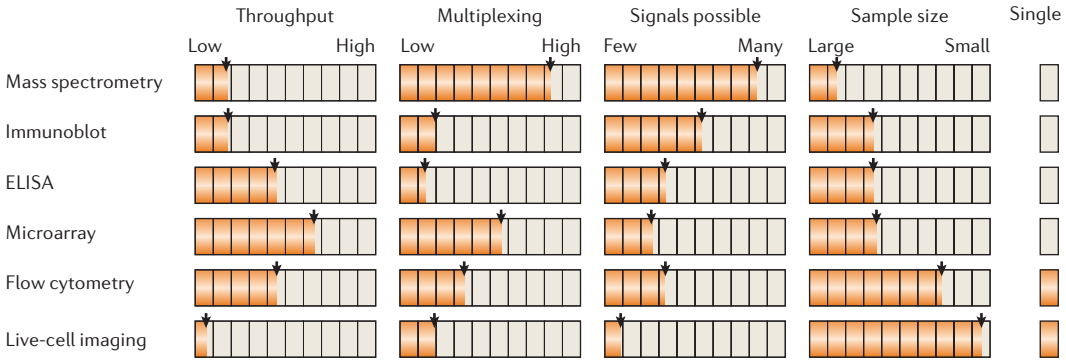


Figure 2 | **Merits of different assay technologies.** A comparison of the experimental techniques discussed in this review with respect to several important considerations. Orange bars indicate the strength of the assay with regard to each criterion, with longer bars being more favourable. For more details, see Supplementary information S2 (table). ELISA, enzyme-linked-immunosorbent assay.

Flow cytometry
A method in which fluorescence-intensity data are recorded from particles in solution as they flow past a detector.

Protein profiling
A method that assesses the expression level of a large set of proteins in a specific tissue or cell type. It is analogous to transcriptional profiling by DNA microarrays.

Protein-interaction microarray
A protein microarray that is used to assay protein interactions. In such arrays, the capture reagents are purified proteins or protein domains, and the analyte solution contains a potential binding partner. Detection strategies are the same as in antibody microarrays (direct labelling or sandwich).

Substrate-protein microarray
A protein microarray that is used to identify substrates of enzymes, such as kinases. In this format, the array consists of potential substrates, and the analyte contains a purified enzyme. Modification of the substrates on the array (for example, phosphorylation) by the analyte is detected by radiolabel incorporation or other labelling strategies.

Microfluidic device
A device for fluid handling in which the smallest dimensions of the features (channels, valves and so on) are on the scale of a few to a few hundred micrometers.

Stable-isotope labelling with amino acids in culture
(SILAC). This method labels proteins from different samples with heavy atoms, yielding mass differences of several Daltons between the same peptide from different samples.

Building on the ELISA's sandwich format are bead-based arrays, in which capture antibodies are bound to beads rather than immobilized on plates and flow cytometry is used to detect antigen binding. Up to 100 differently coloured types of bead can be mixed together and later distinguished by their unique fluorescence signatures, although antibody cross-reactivity typically limits the number of antigens that can be multiplexed to between 10 to 20 (reviewed in REF. 13). Bead-based assays for proteins and their modifications show considerable promise for systems biology.

A variant on conventional immunoassays is the 'in-cell western', a low-resolution form of immunofluorescence microscopy in which cells are grown and fixed in 96-well plates and then probed with target-specific antibodies. The method is rapid and allows many conditions to be assayed[14], but, as in reverse-phase antibody arrays (see below), accuracy demands highly specific antibodies.

Microarray technology. Microarrays — in which printers are used to spot and immobilize cell extracts, antibodies or recombinant proteins on glass slides — are basically miniaturized assays based on the three basic immunoaffinity formats (BOX 1) combined with arrays to enable simultaneous and repeated analysis[15]. Early suggestions that a complete proteome could be profiled on a chip have not proven to be realistic. However, many useful protein chips have been developed, typically with 10–100 features[16]. Microarrays based on a sandwich format are best suited to protein profiling[16], particularly in the case of cytokines[17,18]. For example, an array of anti-cytokine antibodies has been used to determine, in a single experiment, the abundance of 51 cytokines in the supernatants of cultured dendritic cells[67].

Profiling sets of binary interactions among members of multiprotein families relies on protein-interaction microarrays. The equilibrium binding affinities of virtually every Src-homology-2 (SH2) and phosphotyrosine-binding (PTB) domain encoded by the human genome for 31 sites of tyrosine phosphorylation on the 4 human ERBB receptors were recently determined using arrays[19]. The resulting interaction network revealed an unexpected systems-level property: ERBB receptors differ in the extent to which SH2 binding becomes more promiscuous upon overexpression, a property that correlates with the oncogenic potential of various receptors[19].

Last, substrate-protein microarrays make it possible to study transient interactions between protein-modifying enzymes and their substrates on a large scale. In a recent example, microarrays comprising ~4,400 yeast proteins uncovered *in vitro* substrates for 87 protein kinases[20]. Collecting enzyme-substrate data in an unbiased and systematic fashion and integrating the data with protein–protein interaction and transcription-factor-binding data made it possible to derive features of yeast physiology that had eluded conventional analysis.

Advances in affinity-based assays. The mainstay of affinity-based methods are antibodies, although other types of biomolecule show long-term promise[21,22]. Most antibodies are developed and produced in animals[23], although recent improvements in recombinant antibody cloning, display and production techniques have made it easier to generate complex collections of antigen-binding sites *in vitro* and thereby generate high-affinity synthetic antibodies[24]. Microfluidic devices will also improve affinity-based assays by decreasing the volumes of sample and reagent that are required and by increasing throughput and accuracy[68].

Physical methods for protein measurement

Mass spectrometry is a physical method for protein measurement that can, in contrast to affinity-based techniques, potentially detect and measure any soluble apopeptide or modified peptide[25]. An important feature of mass spectrometry is its capability to identify peptides; however, technologies for peptide quantification — which is important for modelling — have only recently been developed.

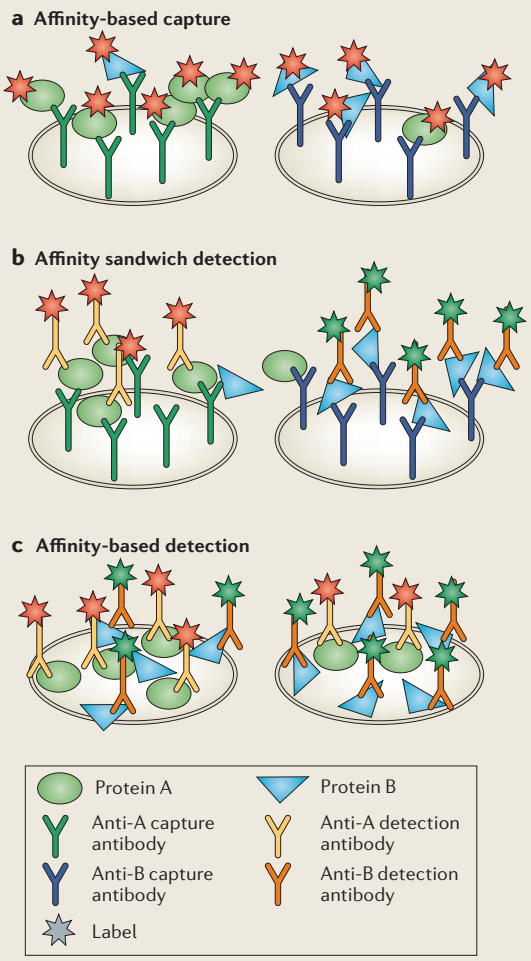
The key to relative peptide quantification by mass spectrometry is differential tagging with a mass label. Stable-isotope labelling *in vivo* relies on metabolic incorporation (for example, using stable-isotope labelling with amino acids in culture (SILAC)[26]), and *in vitro*

Box 1 | **Affinity-based approaches**

Three strategies exist to measure the abundance and state of modification of proteins. Affinity-based capture uses protein-capture reagents (typically antibodies) that are coupled to a solid matrix to pull target proteins out of solution[15] (panel **a**). The captured proteins are visualized by labelling them with a fluorophore[15,61] or another small molecule (such as biotin)[62] before capture. Although this method is straightforward and easily scaled to analyse many target proteins, it is often difficult to find highly selective capture reagents. As such, antigen-capture assays are valuable for initial characterization, but they are not always appropriate for obtaining accurate, quantitative data. Quantification is affected by cross-reactivity of the capture reagents with non-target proteins.

A more accurate, but less scalable, strategy is to use a sandwich immunoassay[63,64] (panel **b**). Following protein capture, a cocktail of detection antibodies is used to detect and quantify the captured proteins. Each protein must now be recognized by two distinct antibodies, a capture antibody and a detection antibody, and therefore interference from antibody cross-reactivity is minimized. Identifying appropriate antibody pairs is difficult and time-consuming. However, as antibody pairs become available for more proteins, microarrays of sandwich assays will prove invaluable for obtaining accurate, quantitative data on tens to hundreds of proteins.

Last, in affinity-based detection or 'reverse-phase' assay[65,66] (panel **c**), the samples (for example, cellular lysates) are spotted directly onto a protein-binding membrane (typically glass-supported nitrocellulose for microarrays). The immobilized proteins are subsequently detected by probing with different antibodies. This assay also suffers from inaccuracies that are introduced by antibody cross-reactivity, but powerful signal-amplification methods can be used to detect and quantify even low-abundance proteins.



involves protein or peptide derivitization (for example, using isobaric tags for relative and absolute quantification (iTRAQ)[27]). In both SILAC and iTRAQ methods, control and experimental samples with different mass labels are mixed prior to analysis on a mass spectrometer. By comparing the relative amounts of each marker ion on a peptide, relative changes in the peptide levels can be determined for large numbers of peptides. SILAC permits multiplex analysis of three samples, and iTRAQ of four, and 8-plex iTRAQ is currently under development. Multiplexing can be used to compare different time points in a time course or different cell populations at a single time point. For example, 4-plex iTRAQ has been used to characterize changes in the levels of phosphotyrosine at more than 100 sites at 4 time points following the stimulation of human mammary epithelial cells with epidermal growth factor (EGF)[28]. Clustering the resulting data yielded preliminary functional assignments for many proteins not previously known to be EGF targets.

One of the great strengths of mass spectrometry is its capability to distinguish among closely related protein species. Mass spectrometry offers exquisite specificity because it can identify proteins using both mass and peptide-fragmentation patterns (as well as chromatographic retention time for liquid chromatography–tandem mass spectrometry)[25]. This makes it possible to distinguish among very similar isoforms of a single protein (for example, isoform-1 and -2 of STAT3 (signal transducer and activator of transcription-3))[28], or to distinguish among phosphorylation states of the same protein (such as the singly and doubly phosphorylated mitogen-activated-protein kinases extracellular signal-regulated kinase-1 (ERK1) and ERK2 (REF. 28)).

A challenge that is faced by all physical approaches to protein analysis, including mass spectrometry, is that most biological samples are very complex — they consist of thousands of different proteins, each of which exists in multiple modified forms. Analysis therefore requires either great power of separation in the instrument itself, or prior fractionation of the sample to reduce complexity. Various affinity-based fractionation and enrichment methods have been developed (see Supplementary information S3 (box); reviewed in REF. 29), making it possible to detect hundreds or thousands of low-abundance phosphopeptides in a single sample without interference from abundant non-phosphorylated proteins[30–32].

Isobaric tags for relative and absolute quantification (iTRAQ). iTRAQ labels are initially isobaric, ensuring that the same peptides from different samples behave identically in the full mass spectrum (MS) mode, but they fragment to generate marker ions that differ by a single Dalton in tandem MS mode during peptide identification.

Marker ion
An ion that carries the isotope label in the breakdown of a peptide during tandem mass spectrometry analysis.

Despite its advantages, mass spectrometry in its current form has several weaknesses. First, throughput is low (FIG. 2). Second, only three to four samples can be compared at a time, and each run typically requires several weeks to analyse. Here, the biggest limitation is not the instrumentation, but rather inadequate software and the need to confirm peptide assignments by hand. Last, samples for protein profiling must be about 200 μg of protein (typically 10^5–10^7 mammalian cells, depending on their size). Nonetheless, the capability of mass spectrometry to provide data on hundreds or thousands of peptides makes it the pre-eminent technology for the analysis of protein modifications when many signals are to be assayed in a few large samples.

Enzymatic activity assays

Activity-based assays are as diverse as the protein functions they measure. Many enzyme activities can now be assayed with reasonable throughput and multiplex measurement. As a first example, G proteins can be assayed in a 96-well format using sandwich detection. A recombinant fragment of an effector molecule (a protein regulated by the GTPase) is used as an affinity capture reagent. Sandwich assays are an indirect measure of GTPase activity and rely on the assumption that only GTP-bound (active) GTPase binds to the effector. G-protein activity assays have been used, for example, to characterize and model the roles of Ras and Rap1 on ERK kinase activation[10].

Second, radioactive immunocomplex kinase assays are an excellent method to monitor the activities of protein kinases in high throughput, and they provide a complementary and more direct view of activity than quantifying activating phosphomodifications. Capture antibodies that do not interfere with catalytic activity and efficient peptide substrates have been identified for several kinases in mammalian cell signalling[33], and the value of high-throughput *in vitro* kinase-activity assays has been shown in an analysis of cellular responses to insulin, tumour necrosis factor and EGF[9,34]. Highly sensitive and selective fluorescence-based chemosensors[35,36] have the potential to replace radiochemical approaches for at least some protein kinases[37]. A third example of an activity assay measures proteases, including the caspases that mediate cell death. Proteases can be assayed in a plate-based format using fluorogenic peptide substrates. As in the case of kinase assays, several companies provide immunocomplex protease assays with good molecular specificity.

The assays described above are quantitative, sensitive and can be done in parallel in a 96-well format, but they are not multiplex. By contrast, activity-based protein profiling (ABPP) allows for the quantification of many enzymes using probes that specifically label proteins with shared catalytic features in single samples (reviewed in REF. 38). In an elegant study, ABPP and the quantification of coupled enzymes by mass spectrometry were used to classify 22 enzymes in 7 groups based on their distinct activity profiles in mouse xenograft models of breast cancer[39]. This study showed that, following implantation in nude mice, human breast cancer cells have dramatically elevated serine protease activities that might contribute to enhanced metastatic potential.

Systems biology at the single-cell level

Cytometry and imaging techniques quantify fluorescent signals at cellular or subcellular resolution and they are the primary means for monitoring single cells. In flow cytometry, subcellular resolution is not possible, but up to 17 fluorophores, potentially representing 17 different antibodies, can be quantified at the same time[40]. Fluorescent data on single cells can also be collected with moderate subcellular resolution using image cytometry[41]. Last, high-resolution fluorescence imaging provides data with detailed subcellular resolution from both live and fixed cells. Both cytometry and imaging technologies are amenable to high-throughput analysis of cells grown in 96-well or 384-well plates, making it possible to measure hundreds to thousands of samples per day. However, significant challenges in data analysis remain to be solved before the full potential of high-throughput high-resolution imaging can be realized (see accompanying Review by Swedlow and Goldberg in *Nature Cell Biology*).

Cytometry and imaging techniques show that signal-transduction events and cellular responses are heterogeneous among cells; however, this heterogeneity is obscured in population-based assays (for example, immunoblotting of cell extracts)[42–44]. Many mathematical models of signalling pathways are representations of reactions taking place in a single cell[44–46], making single cells the ideal source of data for modelling. However, it is impractical to insist that all data should be collected using single-cell methods. Sensitivity of detection is one limiting factor (because many proteins of interest are present at low copy number), as are low throughput and the limited possibilities for multiplexing. Simultaneous measurement of multiple signals is limited by the fact that single-cell techniques usually rely on fluorescence, making detection of ~10 simultaneous signals the practical upper limit for fixed cells (perhaps 20 signals can be measured using the most advanced instruments) and 2–3 signals for live cells. Specificity is also an important challenge because *in vitro* characterization cannot always predict the behaviour of fluorescent probes in the complex intracellular environment.

The most productive approach in practice is to combine population-based and single-cell measurements. Even a small amount of single-cell data can be informative for the interpretation of a data set that was collected using population-based assays. In the simplest case, such as a graded transcriptional response, signals from individual cells are normally distributed around a mean value so that knowledge of the mean and variance of each signal is sufficient to link single-cell and population measurements[47]. In other cases, such as all-or-none enzyme activation[48], signal distributions can be bimodal and population-average measurements are then poor indicators of single-cell behaviour.

Quantification of signals in fixed cells. Although flow cytometry has long been used to assay surface receptors and marker proteins, cytometry is now equally useful for quantifying intracellular signals[49]. For example, multicolour flow cytometry and Bayesian network inference[50] have been used to assemble a model of the

Chemosensor
In the context of kinase assays, a chemosensor is a substrate peptide that contains the non-natural amino acid Sox, which displays chelation-enhanced fluorescence when the peptide is phosphorylated.

Activity-based protein profiling
(ABPP). A method that uses reactive probes carrying a label that will covalently bind specifically to active enzymes of a certain class. The label is often a fluorophore, enabling visualization and quantification of coupled enzymes on gels, antibody microarrays or in cells. Recently, reactive probes have been labelled with an affinity tag for capture of the coupled enzymes, quantification and identification by mass spectrometry.

Image cytometry
A method that uses microscope optics to collect low-resolution data from cells that are adhered to a slide.

Bayesian network inference
A statistical method for inferring the probable relationships between measured variables.

interactions among 11 signalling proteins in human cells. Multicolour cytometry has also been used for comparative phosphoprofiling of stimulated blast cells from healthy and leukaemia-afflicted individuals[51]. Immunostaining techniques are also a mainstay of microscopy, which provides the advantage that protein localization can be quantified. For example, high-throughput microscopy has been used to characterize the response of cancer cells to small molecules[52] and to examine the differentiation of primary neural stem cells in defined microenvironments[53]. These studies show the capability of fixed-cell immunofluorescence cytometry and microscopy to collect rich quantitative data sets from small numbers (10^2–10^4) of cells. Taking full advantage of this feature, several of these studies systematically probed signalling in primary human cells that had been exposed to various stimuli, rather than simply assaying the basal state of these rare and hard-to-obtain cells[50,51,53].

Protein expression and signalling dynamics in living cells. Another useful application of microscopy and cytometry is the quantification of proteins in living cells by use of genetically encoded fluorescent proteins (FPs). One outstanding study combined FP-tagged endogenous genes with calibration by quantitative immunoblotting to determine the absolute expression levels of 28 proteins that are involved in cytokinesis[54]. This study showed that flow cytometry and microscopy can provide similarly accurate measurements of protein concentration, although microscopy had lower variance and allowed for the measurement of subcellular protein concentrations. The inherent cell-to-cell variability in gene expression has also been examined using flow cytometry[55,56]. Together, these studies show that FPs can be used to determine the levels and variability in protein abundance, data that are critical for physicochemical models of biological networks.

FPs can also be used to observe the dynamic behaviour of a biological network by following single cells over time. For example, asynchronous oscillations in protein signals, such as the transcription factors p53 and nuclear factor (NF)-κB, in individual cells are often averaged out in population-level measurements, or frozen in time by immunofluorescence-based cytometry [42,44]. Another feature of live-cell microscopy is that patterns of behaviour can be identified within subpopulations by measuring many cells (10^2–10^3) in parallel. In a study of the p53 signalling network, individual cells exposed to UV radiation were observed to respond with 1 to greater than 10 digital pulses of p53 nuclear translocation, an oscillatory behaviour that persisted for 2–3 days after the initial stimulus[46]. Further analysis showed that the pulses varied significantly in amplitude but not in period; these data were then used for the development of a biochemical model by identifying the best-fitting model topology[46].

The studies discussed above show the power of single-cell assays in providing rich data that can be used in modelling. Nonetheless, single-cell techniques will never equal biochemical methods in the breadth of signals that can be assayed (FIG. 2). The key is integrating population and single-cell data, for example, by using quantitative models to represent the behaviours of populations of single cells[45].

Integration of models and data sets

Following the development of a data-collection strategy, several steps are involved in transforming raw measurements into systematic data that is useful for quantitative mathematical modelling. These include: data validation and error estimation; data normalization and fusion; data scaling; derivation of computed metrics; and comparison of model and data.

Data validation and error estimation. The specificity and linearity of an assay are the first concerns during data validation. In affinity-based approaches, problems often arise from cross-reactivity of antibodies with proteins other than the target. Immunoblots provide useful information on antibody specificity because they separate proteins by mass, but it can be difficult to validate the specificity of protein arrays or in-cell westerns that rely on affinity capture or direct affinity-based detection. In these cases we typically compare results obtained with different antibodies, use RNA interference or chemical inhibitors to deplete signals, or cross-validate measurements with an independent technique (often immunoblotting). The relationships between measured and actual signal values must also be determined. Fortunately, this can usually be accomplished by appropriate dilution of samples, although absolute calibration requires recombinant proteins and careful titration. Typically, good assays are monotonic and stable over a 10–100-fold range of signal, and within this range measured values can be converted into true values using a standard curve.

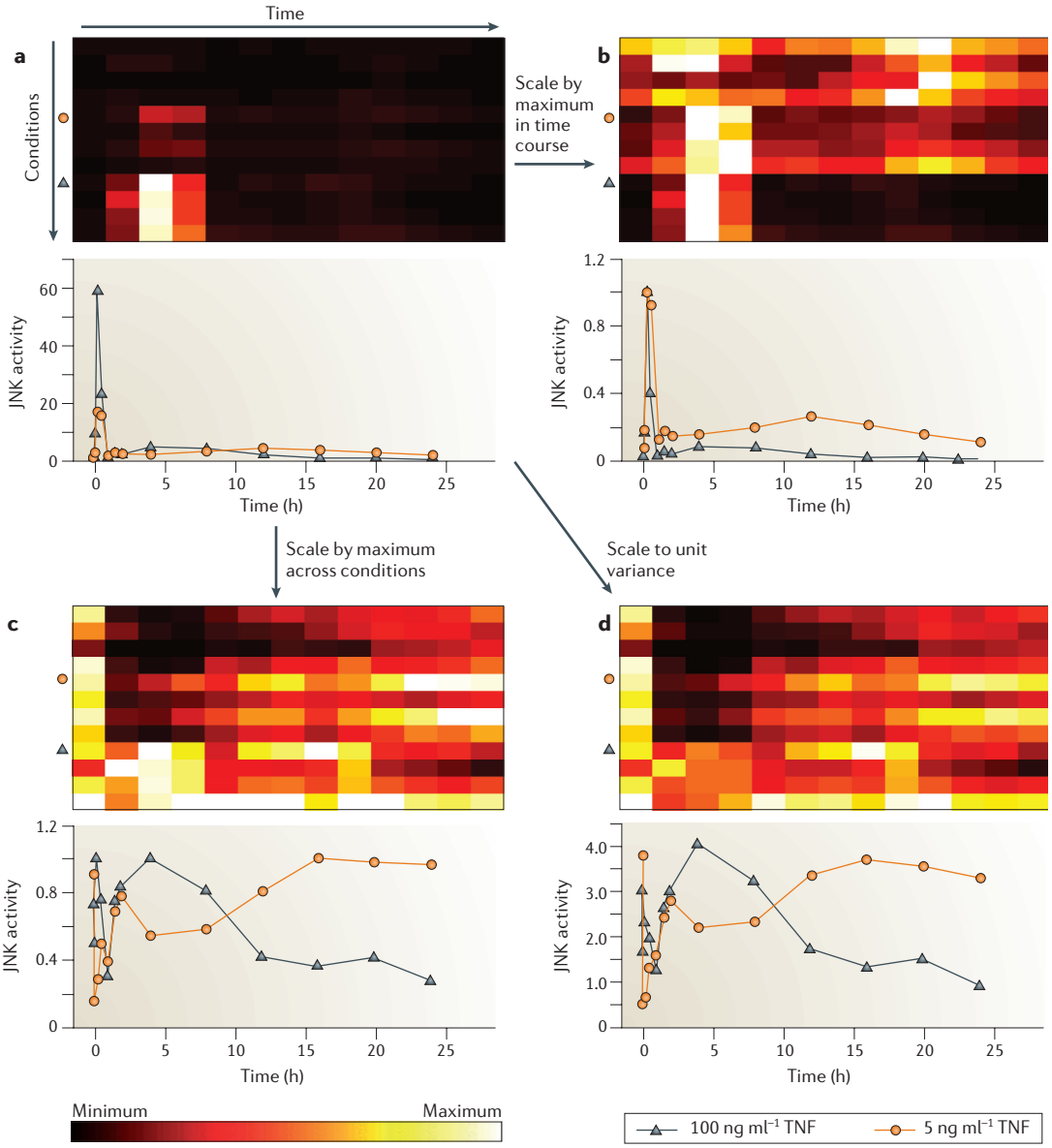
The error that is associated with measurement must also be determined. Fixed error is usually identified when several methods are compared. The most obvious way to estimate variance is to repeat a measurement many times (that is, increase n) and estimate the standard deviation. However, limited amounts of sample, costly reagents and limited time usually make it impossible to obtain enough replicates for standard statistical methods. Duplicate measurements are therefore common, and in this case error must be estimated by other means. One highly effective approach is to develop a quantitative physical model of the data-collection process itself. For example, the effects of cell-to-cell variability, finite optical resolution, fluorophore chemistry, reliability of data-processing algorithms and so on were used to precisely model errors arising in fluorescence speckle microscopy[57,58]. Another less rigorous approach is to use error estimates that are obtained from repeated sampling of a single signal as a best guess of average error. In this case we still need to be concerned with mistakes that are made in the course of a complex experiment, and duplicate measures on different samples (that is, biological replicates) are an absolute minimum requirement. It should be noted that techniques that are potentially applicable to protein signals have also been developed for low-replicate, biological DNA-microarray data[59].

Data validation
The process of verifying assay accuracy.

Data normalization
The adjustment of measured values to account for possible run-to-run and day-to-day variability in the assays.

Fluorescence speckle microscopy
Speckles that form by the random association of fluorophores with macromolecular structures are tracked by live-cell imaging. The information in the dynamic behaviour of these speckles is converted into a quantitative spatio-temporal readout of cytoskeleton-polymer transport and turnover.

Box 2 | **Scaling of data**



Scaling of data is useful to highlight trends or features that might be obscured by dominant signals in the data set. For example, panel **a** shows a heat map that describes Jun N-terminal kinase (JNK) activity over a time course of 24 h (horizontal axis) in response to 12 different stimulation conditions (vertical axis). The time courses for two individual conditions (stimulation with 5 ng ml^{-1} or 100 ng ml^{-1} tumour necrosis factor (TNF)) are shown below the heat map. Within the unscaled data set in panel **a**, it is obvious that high concentrations of TNF induce an early and strong peak of JNK activity, whereas lower concentrations induce a weaker peak.

To highlight the progression in time of the JNK-activity signal in conditions in which maximal JNK activity is much lower (for example, 5 ng ml^{-1} TNF), we must scale the measurements to the maximal value in each time course. That is, we express all values relative to the peak activity in the time course (panel **b**). After scaling, relative signal strengths across conditions are lost, but differences and similarities in timing become apparent. Although early peaks coincide, there is a second wave of JNK activity under several treatment conditions, and, although this second wave is of much smaller amplitude than the initial peak, it is also variable across different conditions.

Third, to highlight variation across different stimuli, data should be scaled according to the maximal value across all treatments on a time point by time point basis (panel **c**). This allows for the identification of conditions that have a significantly higher signal at all time points (bottom row in heat map, panel **c**).

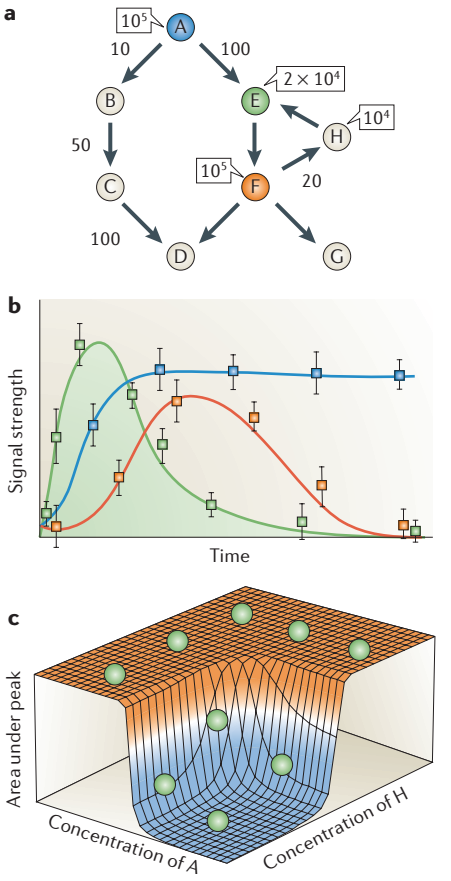
Last, a commonly used data-processing step for regression-based models is unit-variance scaling. Scaling each variable (here, JNK activity at each time point) by the square root of its variance results in all the scaled variables having a variance of one. Absolute differences are lost, but the relative variability across stimulation conditions for each time point is preserved and can be explicitly analysed without bias from differences in signal amplitude (panel **d**).
Data from REF. 9.

Box 3 | **Matching data to models**

Here we provide several illustrations of the ways in which experimental data can relate to models. For a rigourous treatment, see the accompanying Review by Jaqaman and Danuser in this issue. In the most straightforward case (panel **a**), experimental data directly correspond to the structure or the parameters of the model. For example, protein-interaction microarrays and substrate-protein microarrays identify connections between molecules in the network (panel **a**, arrows). Similarly, values for protein concentrations (panel **a**, numbers in callouts) can be determined by protein-profiling microarrays, and values for rate constants (panel **a**, numbers near arrows) can be determined by activity assays.

In another simple case (panel **b**), the dynamic behaviour of the model is directly comparable to an experimental time course (here, using molecules A, E and F). Experimental values for signals at various time points (panel **b**, square data points) can be overlaid on the simulated signals from the model (panel **b**, lines). The extent to which the experimentally observed behaviour matches the model simulation can be assessed by quantitative methods.

In a more complex case, computed metrics (descriptive) can be used to compare experimentally observed behaviour with the model. For example, it might be desirable to examine the area under a signal peak (panel **b**, shaded green area). To examine the behaviour of this metric, time course data can be collected under conditions in which the network is perturbed (for example, by using RNA interference to change the concentrations of proteins A and H). From each experimental time course, the signal-area metric can be calculated and plotted as a function of the perturbed protein levels (which can also be measured experimentally; panel **c**, green circles). For comparison, model simulations can be run under conditions that correspond to the experimental perturbations and the computed metric can be calculated and plotted for each simulation (panel **c**, surface plot). This type of analysis can provide insight into the mechanisms that underlie higher-order behaviours of the system.



Data normalization and fusion. Assembly of a data compendium requires the fusion of multiple data types, many of which are collected over many weeks or months. Correctly normalized and validated data compendia allow quantitative analysis to be performed over a much larger landscape of experimental conditions than is possible using single experiments. Therefore, effective means to fuse heterogeneous data are essential. Data normalization is a crucial part of data fusion and involves correcting for changes in experimental and assay conditions that can be measured, but not necessarily controlled. Normalization usually involves calibrating measurements against a set of standards that are included within each run and then adjusting raw values to that of the standard, taking into account cell number or total protein concentration. It should be noted that calibration against 'housekeeping' proteins is not necessarily reliable, as the levels of many of these proteins fluctuate under specific experimental conditions, such as apoptosis[60].

Following normalization, one has to verify that combining data from several different assays has generated a single self-consistent data set across which comparisons can be made. If data were acquired in a succession of experiments that were separated in time, one efficient way of verifying the consistency of the fused set is to replicate a subset of measurements in a separate experiment using orthogonal design[9]. The data compendium is self-consistent if the new measurements correlate well with the equivalent measurements in the initial data set. We have observed cases in which it has been more effective to build multiple models from subsets of data, each collected in the course of a single experiment, and then fuse the models rather than the data. Although a rigorous analysis has not yet been done, fusion at the level of models seems to be most effective when biological variability is high but measurements are accurate.

Data scaling. Once a data set has been validated, normalized and fused, it is possible to explore approaches to scaling it (BOX 2). Scaling involves the transformation of axes in such a way as to highlight features or trends in the data. One simple form of scaling is switching from linear to logarithmic axes; more complex forms of scaling involve adjusting the axes relative to some global property of the data. For example, in a time course of protein signals collected from many different perturbations, regression techniques tend to focus on the largest absolute changes in signal strengths, even if these changes occur in every perturbation. However, if unit-variance scaling (BOX 2) is applied, modelling focuses on what is the most characteristic of each perturbation. Data scaling can have profound effects on data-driven models[9] and it is necessary to compare several scaling methods before selecting one.

Orthogonal design
A method of validation in which conditions that were previously varied from experiment to experiment in the course of collecting a full data set are varied in a single experiment such that what was previously separated in time now becomes contemporaneous.

Computed metrics. Computed metrics are data transformations in which single measurements are used to compute a more global property of the signal, such as the frequency of oscillation, the area under the curve of a peak or the rate of rise. In many cases, computed metrics seem to have a higher information content than individual data points alone[9,46]. This arises in part because most regression techniques treat each time point in isolation so that crucial dynamic information is lost. The use of computed metrics captures key features of the time dependence of signalling in a single variable[52,58].

Comparing data and model. It is important to compare the data to the model during model construction, when the model is calibrated against training data, and when the predictions and hypotheses that are generated by the model are tested. In some cases, this comparison is straightforward (BOX 3). For example, model simulations can be used to produce a predicted time course of kinase phosphorylation that can be directly compared to experimental phosphoprotein levels. It is also possible to determine the relative importance of each piece of data to the model. In this way, a quantitative evaluation of the data-collection process itself can be done (see accompanying Review by Jaqaman and Danuser in this issue).

Conclusions

To model signalling and regulatory networks, quantitative data on multiple parts of a protein network are required. However, signalling networks are built from proteins and other biomolecules with diverse functions and properties, and this poses a significant challenge for data collection: several measurement techniques that yield heterogeneous data must be combined to characterize the signals that are transmitted through the networks. Technical advances in biochemical analyses using physical, affinity-based and activity-based methods have created more accurate, more sensitive and higher throughput methods. Similar improvements in fluorescent labelling and simultaneous detection of multiple fluorophores have increased the value of microscopy-based and flow-cytometry-based measurements to systems biology. However, what has been missing is a commitment by systems biologists to use multiple measurement techniques simultaneously and to develop methods for fusing the resulting data into self-consistent, validated data compendium. As the value of these compendia for physicochemical and statistical modelling becomes clear, we can expect much greater attention to issues that surround data fusion, modelling and analysis.

1. Mockler, T. C. *et al.* Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**, 1–15 (2005).
2. Fan, J. B., Chee, M. S. & Gunderson, K. L. Highly parallel genomic assays. *Nature Rev. Genet.* **7**, 632–644 (2006).
3. Joyce, A. R. & Palsson, B. O. The model organism as a system: integrating 'omics' data sets. *Nature Rev. Mol. Cell Biol.* **7**, 198–210 (2006).
4. Kim, T. H. & Ren, B. Genome-wide analysis of protein–DNA interactions. *Annu. Rev. Genomics Hum. Genet.* **7**, 81–102 (2006).
5. Ness, S. A. Basic microarray analysis: strategies for successful experiments. *Methods Mol. Biol.* **316**, 13–33 (2006).
6. Quackenbush, J. Microarray data normalization and transformation. *Nature Genet.* **32** (Suppl.), 496–501 (2002).
7. Morris, M. & Watkins, S. M. Focused metabolomic profiling in the drug development process: advances from lipid profiling. *Curr. Opin. Chem. Biol.* **9**, 407–412 (2005).
8. Nielsen, J. & Oliver, S. The next wave in metabolome analysis. *Trends Biotechnol.* **23**, 544–546 (2005).
9. Gaudet, S. *et al.* A compendium of signals and responses triggered by prodeath and prosurvival cytokines. *Mol. Cell. Proteomics* **4**, 1569–1590 (2005).
An example of data-compendium assembly from a data set of ~7,000 heterogeneous protein measurements. Shows the critical importance of appropriate data normalization and scaling techniques in building predictive models.
10. Sasagawa, S., Ozaki, Y., Fujita, K. & Kuroda, S. Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nature Cell Biol.* **7**, 365–373 (2005).
A mechanistic modelling effort is driven by an impressive data set of immunoblots and GTPase assays. An excellent example of a model carefully matched to experimental data.
11. Schweitzer, B. *et al.* Immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc. Natl Acad. Sci. USA* **97**, 10113–10119 (2000).
12. Debad, J. D., Glezer, E. N., Wohlstadter, J. N. & Sigal, G. B. in *Electrogenerated Chemiluminescence* (ed. Bard, A. J.) 43–78 (Marcel Dekker, New York, 2004).
13. Vignali, D. A. Multiplexed particle-based flow cytometric assays. *J. Immunol. Methods* **243**, 243–255 (2000).
14. Kortum, R. L. *et al.* The molecular scaffold kinase suppressor of Ras1 (KSR1) regulates adipogenesis. *Mol. Cell. Biol.* **25**, 7592–7604 (2005).
15. Haab, B. B. Advances in protein microarray technology for protein expression and interaction profiling. *Curr. Opin. Drug Discov. Devel.* **4**, 116–123 (2001).
16. MacBeath, G. Protein microarrays and proteomics. *Nature Genet.* **32** (Suppl.), 526–532 (2002).
17. Wang, C. C. *et al.* Array-based multiplexed screening and quantitation of human cytokines and chemokines. *J. Proteome Res.* **1**, 337–343 (2002).
18. Olle, E. W. *et al.* Development of an internally controlled antibody microarray. *Mol. Cell. Proteomics* **4**, 1664–1672 (2005).
19. Jones, R. B., Gordus, A., Krall, J. A. & MacBeath, G. A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* **439**, 168–174 (2006).
Protein microarrays were used to measure the synoptic binding profile of all human SH2 and PTB domains for 61 phosphotyrosine sites in the ERBB1–4 receptors.
20. Ptacek, J. *et al.* Global analysis of protein phosphorylation in yeast. *Nature* **438**, 679–684 (2005).
21. Hermann, T. & Patel, D. J. Adaptive recognition by nucleic acid aptamers. *Science* **287**, 820–825 (2000).
22. Tombelli, S., Minunni, M. & Mascini, M. Analytical applications of aptamers. *Biosens. Bioelectron.* **20**, 2424–2434 (2005).
23. Harlow, E. & Lane, D. *Antibodies: a Laboratory Manual.* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1988).
24. Colby, D. W. *et al.* Engineering antibody affinity by yeast surface display. *Methods Enzymol.* **388**, 348–358 (2004).
25. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
26. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
27. Zieske, L. R. A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. *J. Exp. Bot.* **57**, 1501–1508 (2006).
28. Zhang, Y. *et al.* Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol. Cell. Proteomics* **4**, 1240–1250 (2005).
29. Schmelzle, K. & White, F. M. Phosphoproteomic approaches to elucidate cellular signaling networks. *Curr. Opin. Biotechnol.* **17**, 406–414 (2006).
30. Beausoleil, S. A. *et al.* Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl Acad. Sci. USA* **101**, 12130–12135 (2004).
31. Moser, K. & White, F. M. Phosphoproteomic analysis of rat liver by high capacity IMAC and LC–MS/MS. *J. Proteome Res.* **5**, 98–104 (2006).
32. Nousiainen, M., Sillje, H. H., Sauer, G., Nigg, E. A. & Korner, R. Phosphoproteome analysis of the human mitotic spindle. *Proc. Natl Acad. Sci. USA* **103**, 5391–5396 (2006).
33. Janes, K. A. *et al.* A high-throughput quantitative multiplex kinase assay for monitoring information flow in signaling networks: application to sepsis-apoptosis. *Mol. Cell. Proteomics* **2**, 463–473 (2003).
34. Janes, K. A. *et al.* The response of human epithelial cells to TNF involves an inducible autocrine cascade. *Cell* **124**, 1225–1239 (2006).
35. Shults, M. D. & Imperiali, B. Versatile fluorescence probes of protein kinase activity. *J. Am. Chem. Soc.* **125**, 14248–14249 (2003).
36. Shults, M. D., Pearce, D. A. & Imperiali, B. Modular and tunable chemosensor scaffold for divalent zinc. *J. Am. Chem. Soc.* **125**, 10591–10597 (2003).
37. Shults, M. D., Janes, K. A., Lauffenburger, D. A. & Imperiali, B. A multiplexed homogeneous fluorescence-based assay for protein kinase activity in cell lysates. *Nature Methods* **2**, 277–283 (2005).
38. Evans, M. J. & Cravatt, B. F. Mechanism-based profiling of enzyme families. *Chem. Rev.* **106**, 3279–3301 (2006).
39. Jessani, N. *et al.* Carcinoma and stromal enzyme activity profiles associated with breast tumor growth *in vivo*. *Proc. Natl Acad. Sci. USA* **101**, 13756–13761 (2004).
40. Perfetto, S. P., Chattopadhyay, P. K. & Roederer, M. Seventeen-colour flow cytometry: unravelling the immune system. *Nature Rev. Immunol.* **4**, 648–655 (2004).
41. Ecker, R. C. & Steiner, G. E. Microscopy-based multicolor tissue cytometry at the single-cell level. *Cytometry A* **59**, 182–190 (2004).

42. Lahav, G. *et al.* Dynamics of the p53–Mdm2 feedback loop in individual cells. *Nature Genet.* **36**, 147–150 (2004).
43. Nair, V. D., Yuen, T., Olanow, C. W. & Sealfon, S. C. Early single cell bifurcation of pro- and antiapoptotic states during oxidative stress. *J. Biol. Chem.* **279**, 27494–27501 (2004).
44. Nelson, D. E. *et al.* Oscillations in NF-κB signaling control the dynamics of gene expression. *Science* **306**, 704–708 (2004).
Live-cell imaging and computational modelling are combined to link pulses of NF-κB nuclear translocation to the level of transcriptional activity.
45. Eissing, T. *et al.* Bistability analyses of a caspase activation model for receptor-induced apoptosis. *J. Biol. Chem.* **279**, 36892–36897 (2004).
46. Geva-Zatorsky, N. *et al.* Oscillations and variability in the p53 system. *Mol. Syst. Biol.* **2**, 2006.0033 (2006).
An intensive effort in which live-cell measurements of p53 and MDM2 translocation dynamics in ~1,000 single cells are used to constrain mechanistic network models and identify sources of cell-to-cell variability.
47. Rossi, F. M., Kringstein, A. M., Spicher, A., Guicherit, O. M. & Blau, H. M. Transcriptional control: rheostat converted to on/off switch. *Mol. Cell* **6**, 723–728 (2000).
48. Tyas, L., Brophy, V. A., Pope, A., Rivett, A. J. & Tavare, J. M. Rapid caspase-3 activation during apoptosis revealed using fluorescence-resonance energy transfer. *EMBO Rep.* **1**, 266–270 (2000).
49. Krutzik, P. O. & Nolan, G. P. Intracellular phospho-protein staining techniques for flow cytometry: monitoring single cell signaling events. *Cytometry A* **55**, 61–70 (2003).
50. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529 (2005).
A novel method for using flow cytometry data to automatically generate network topology models.
51. Irish, J. M. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).
52. Perlman, Z. E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
53. Soen, Y., Mori, A., Palmer, T. D. & Brown, P. O. Exploring the regulation of human neural precursor cell differentiation using arrays of signaling microenvironments. *Mol. Syst. Biol.* **2**, 37 (2006).
54. Wu, J. Q. & Pollard, T. D. Counting cytokinesis proteins globally and locally in fission yeast. *Science* **310**, 310–314 (2005).
Fluorescence microscopy and flow cytometry of yellow-FP-tagged genes was used to determine the absolute global and local concentrations of ~40 proteins in the yeast cytokinesis network. This is the largest survey of absolute endogenous-protein concentrations so far.
55. Newman, J. R. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
56. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nature Genet.* **38**, 636–643 (2006).
57. Danuser, G. & Waterman-Storer, C. M. Quantitative fluorescent speckle microscopy of cytoskeleton dynamics. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 361–387 (2006).
58. Ponti, A., Machacek, M., Gupton, S. L., Waterman-Storer, C. M. & Danuser, G. Two distinct actin networks drive the protrusion of migrating cells. *Science* **305**, 1782–1786 (2004).
Statistical modelling of high-resolution live-cell microscopy data reveals remarkable kinetic differences between subsets of the actin network in migrating cells.
59. Sasik, R., Calvo, E. & Corbeil, J. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics* **18**, 1633–1640 (2002).
60. Mashima, T., Naito, M., Fujita, N., Noguchi, K. & Tsuruo, T. Identification of actin as a substrate of ICE and an ICE-like protease and involvement of an ICE-like protease but not ICE in VP-16-induced U937 apoptosis. *Biochem. Biophys. Res. Commun.* **217**, 1185–1192 (1995).
61. Sreekumar, A. *et al.* Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. *Cancer Res.* **61**, 7585–7593 (2001).
62. Knezevic, V. *et al.* Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics* **1**, 1271–1278 (2001).
63. Schweitzer, B. *et al.* Multiplexed protein profiling on microarrays by rolling-circle amplification. *Nature Biotechnol.* **20**, 359–365 (2002).
64. Nielsen, U. B., Cardone, M. H., Sinskey, A. J., MacBeath, G. & Sorger, P. K. Profiling receptor tyrosine kinase activation by using Ab microarrays. *Proc. Natl Acad. Sci. USA* **100**, 9330–9335 (2003).
65. Paweletz, C. P. *et al.* Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**, 1981–1989 (2001).
66. Chan, S. M., Ermann, J., Su, L., Fathman, C. G. & Utz, P. J. Protein microarrays for multiplex analysis of signal transduction pathways. *Nature Med.* **10**, 1390–1296 (2004).
67. Schweitzer, B. *et al.* Multiplexed protein profiling on microarrays by rolling circle amplification. *Nature Biotechnol.* **20**, 359–365 (2002).
68. El-Ali, J., Sorger, P. K. & Jenson, K. F. Cells on chips. *Nature* **442**, 403–411 (2006).

Acknowledgements
This work was funded by a systems biology centre grant from the National Institutes of Health.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
UniProtKB: http://ca.expasy.org/sprot
EGF | ERK1 | ERK2 | p53 | Rap1 | STAT3

FURTHER INFORMATION
Homepage of the Center for Cell Decision Processes:
http://www.cdpcenter.org

SUPPLEMENTARY INFORMATION
See online article: S1 (box) | S2 (table) | S3 (box)
Access to this links box is available online.

S2 (table). Quantitative methods for protein measurement

Method	Throughput	Multiplexing	Required prior knowledge - reagents	Sample size	Comments	Typical CV
Physical methods						
Mass spectrometry	Low	10-1000s of targets	Genome (proteome) sequence	10^5-10^7 cells	Best to identify new molecular players in networks. Requires fractionation or enrichment methods. Difficult to detect low abundance proteins.	~10% (REF 1)
Affinity-based methods						
Immunoblots	Low	1-5 targets	High specificity antibodies	10^4-10^5 cells	Not automated, labor intensive.	~15% (REF 2)
Enzyme-linked-immunosorbent assay (ELISA)	High	Single target	High specificity antibody pairs	10^4-10^6 cells	Exquisite sensitivity possible with rolling-circle method of detection.	~10%
Bead-based ELISA	High	1-10 targets	High specificity antibody pairs	10^4-10^6 cells	Allows multiplexing of the ELISA method.	~10%
In-cell western	High	1-2 targets	High specificity antibodies	10^4-10^6 cells	Rapid, immunofluorescence-based method.	~10%
Microarrays	Excellent	10s of targets	High specificity capture and/or detection reagents	10^3-10^4 cells	Excellent for characterization of protein-protein interactions.	~10% (REF2)
Activity-based methods						
GTPase assay	High	Single target	High specificity capture and detection reagents	10^3-10^4 cells	Commercialized assays.	
Caspase assay	High	Single target	High specificity capture antibodies and efficient substrate	10^3-10^4 cells	Commercialized assays.	
Kinase assay	High	Single target	High specificity capture antibodies and/or efficient substrate	10^3-10^4 cells	Radiolabelled detection assays optimized for a few kinases; availability of chemosensors enhances throughput.	~15% (REF 2)
Activity-based protein profiling (ABPP)	Low	10s of targets	Activity-based reactive probe	10^4-10^6 cells	Best to identify all active enzymes of a class. Difficult to detect low abundance enzymes.	

Single-cell methods						
Flow cytometry	High	1-12 targets	High specificity antibodies or fluorescent reporters	10^2-10^4 cells	Easy quantification, no subcellular resolution.	~10% (REF 3)
Image cytometry	High	1-12 targets	High specificity antibodies or fluorescent reporters	10^2-10^4 cells	Easy quantification, low subcellular resolution.	
Fixed-cell microscopy	High	1-12 targets	High specificity antibodies or fluorescent reporters	10^2-10^4 cells	May require intense computational analysis for quantification. Allows high-resolution data acquisition.	
Live-cell microscopy	Low	1-4 targets	Fluorescent reporters	10^2-10^4 cells	Allows direct observation of signal dynamics. Requires intense computational analysis.	

References

1. Schmelzle, K., Kane, S., Gridley, S., Lienhard, G.E. & White, F.M. Temporal dynamics of tyrosine phosphorylation in insulin signaling. *Diabetes* **55**, 2171–2179 (2006).
2. Gaudet, S. *et al.* A compendium of signals and responses triggered by prodeath and prosurvival cytokines. *Mol. Cell Proteomics* **4**, 1569–1590 (2005).
3. Janes, K.A. *et al.* A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* **310**, 1646–1653 (2005).

S1 (Box). Matching data and objectives in systems biology

Data collection strategies must be designed to maximize the relevance of the data to the objective of the study (Fig. 1A). One objective in systems biology is to determine the overall structure or topology of protein signalling networks, that is, to determine which proteins and pathways are activated by a given stimulus and how they are connected. Using Bayesian inference to determine topology requires data on a broad set of signals and care that no critical pathways are omitted. Applying various stimuli and perturbations helps explicate the network by exposing it to a wide range of inputs. However, data on the time-evolution of signals is less important and therefore the measurements could be done at only one or two time points, leaving out a detailed characterization of signalling dynamics.

Once the structure of a signalling network is known, the next objective is to characterize signal flow in response to stimuli. Do signals peak and fall, rise to a sustained level, oscillate, etc? In this case, it is possible to focus on few stimuli or perturbations but measure as many signals as possible, in each case including sufficient data points, to capture signalling dynamics. When analysing mammalian signal transduction, we have found non-uniform time steps useful. We sample frequently in the first 60 min, during which immediate early signals vary rapidly, and then less frequently over the next 24 hr. In many cases, it can be beneficial initially to stimulate cells with "saturating" conditions, where the response of the system is maximal. Although it is arguably artificial (often far from physiological conditions), a saturating treatment yields the largest dynamic range of signal. Subsequent analysis can focus on more subtle stimuli in the physiological range.

When the general behaviour of a network is known we can focus on how network structure determines behaviour. Which feedback loops, forms of crosstalk

and pathway bifurcations are central to system dynamics? Multiple perturbations are important but it is therefore necessary to limit the number of signals measured. With luck, previous analysis will have revealed which co-vary closely. When choosing signals to measure, it is important to select those that are broadly distributed over the network, sampling from each of several pathway branches. Selecting signals upstream in a pathway shows the immediate effects of stimuli whereas downstream signals are typically more predictive of physiological outcome.

S3 (box). Enrichment of phosphoproteins for mass spectrometry analysis.

When performing mass spectrometry on very complex samples such as whole cell lysates, current instruments can only reliably detect and quantify high abundance biomolecules. To allow for the analysis of lower abundance molecules, the sample complexity must be reduced. For the analysis of phosphoproteins, there are multiple of strategies to reduce the complexity, but they are generally based on two steps of enrichment: 1) sample fractionation and 2) immobilized metal affinity chromatography (IMAC) for phosphoproteins.

Physical approaches such as centrifugation techniques have been used to fractionate lysates, and to isolate particular structures such as mitotic spindles[1] or nuclei[2]. Another common strategy for fractionation is immunoprecipitation. For example, one can use antibodies against a specific phosphorylated amino acid (for example, anti-phosphotyrosine antibodies), a specific phosphorylated motif (for example, anti-AKT substrate antibodies), or a specific protein or complex (for example, the EGF receptor, to purify the receptor and its binding partners).

In the second enrichment step, phosphorylated peptides are isolated from the fractionated sample by binding to a resin with high affinity for phosphoryl groups[3]. If this purification step is performed on intact proteins, digestion of the enriched sample with proteases will yield a mixture of phosphorylated and non-phosphorylated peptides. However, if IMAC is performed after proteolytic digestion, virtually all the peptides remaining in the mixture are phosphorylated, further reducing the complexity of the sample prior to analysis.

References

1. Nousiainen, M., Sillje, H.H., Sauer, G., Nigg, E.A. & Korner, R. Phosphoproteome analysis of the human mitotic spindle. *Proc. Natl Acad. Sci. U S A* **103**, 5391–5396 (2006).
2. Beausoleil, S.A. *et al.* Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl Acad. Sci. U S A* **101**, 12130–12135 (2004).
3. Muszynska, G., Andersson, L. & Porath, J. Selective adsorption of phosphoproteins on gel-immobilized ferric chelate. *Biochemistry* **25**, 6850–6853 (1986).