BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 9 2005, pages 2027–2035
doi:10.1093/bioinformatics/bti278

Systems biology

Modeling of signal–response cascades using decision tree analysis

Sampsa Hautaniemi[1,2,*], Sourabh Kharait[3], Akihiro Iwabu[3], Alan Wells[3] and Douglas A. Lauffenburger[1]

[1]Biological Engineering Division, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, [2]Institute of Signal Processing, Tampere University of Technology, 33101 Tampere, Finland and [3]Department of Pathology, University of Pittsburgh, Pittsburgh, PA 15261, USA

Received on November 21, 2004; revised and accepted on January 13, 2005
Advance Access publication January 18, 2005

ABSTRACT

Motivation: Signal transduction cascades governing cell functional responses to stimulatory cues play crucial roles in cell regulatory systems and represent promising therapeutic targets for complex human diseases. However, mathematical analysis of how cell responses are governed by signaling activities is challenging due to their multivariate and non-linear nature. Diverse computational methods are potentially available, but most are ineffective for protein-level data that is limited in extent and replication.

Results: We apply a decision tree approach to analyze the relationship of cell functional response to signaling activity across a spectrum of stimulatory cues. As a specific example, we studied five intracellular signals influencing fibroblast migration under eight conditions: four substratum fibronectin levels and presence versus absence of epidermal growth factor. We propose techniques for preprocessing and extending the experimental measurement set via interpolative modeling in order to gain statistical reliability. For this specific case study, our approach has 70% overall classification accuracy and the decision tree model reveals insights concerning the combined roles of the various signaling activities in governing cell migration speed. We conclude that decision tree methodology may facilitate elucidation of signal–response cascade relationships and generate experimentally testable predictions, which can be used as directions for future experiments.

Contact: sampsa@mit.edu

1 INTRODUCTION

Physiological cell behavioral functions, such as proliferation, death, differentiation and migration, are governed to a large degree by networks of signaling proteins whose activities are influenced by a variety of extracellular cues: environmental agents such as chemical ligands, mechanical forces, radiation, toxins, pathogens and so forth. Dysregulation of these networks is often associated with inappropriate cell and tissue behavior, so that signaling cascades are considered to be promising therapeutic targets for complex pathologies such as diabetes, cancer, and inflammatory diseases (Lodish *et al*., 2004).

Quantitative experimental measurement of cell signaling protein properties—i.e. their levels, states (phosphorylation, cleaveage, etc.), activities, locations—is more challenging, relative to gene-level measurements, to undertake in highly multivariate fashion. Consequently, while measurement of mRNA expression for hundreds and thousands of genes across a spectrum of conditions has become common place, analogous measurement of protein properties as listed above remains limited to the order of tens at best. A critical consequence of this situation is that many of the informatics methods by which computational analyses of genomic data are now being typically pursued are not readily applicable to proteomic (if that term can be used properly for coverage of only about tens of proteins) data. This is the problem that our effort here is directed toward addressing: finding appropriate computational techniques to elucidate useful models of the relationships between protein signals and cell functional responses to extracellular cues given the quantitative data across diverse conditions.

As a motivating case study, we consider cell migration, which is a central biological process in several pathological states such as tumor invasion as well as physiological ones such as wound healing (Ridley *et al*., 2003). Migration can be strongly influenced by both soluble environmental cues (e.g. growth factors and cytokines) and insoluble substratum cues (e.g. extracellular matrix proteins). In our specific experimental problem, we are studying the migration of tissue fibroblasts in response to four levels of surface fibronectin (Fn) concentrations in the absence or presence of epidermal growth factor (EGF), offering eight cue conditions. Fn is a ligand for integrin adhesion receptor-mediated signaling pathways and it has been shown to significantly effect migration of fibroblasts as well as other cell types including many tumor cells (Wells *et al*., 2002). EGF also exhibits a strong influence on migration of normal tissue cells, including fibroblasts, and various types of cancer cells, via signaling pathways mediated by EGF receptor (EGFR) (Maheshwari *et al*., 1999). Indeed, the EGFR system has been associated with the development and progression of a large number of tumors and is one of the most prominent pathways for therapeutical targets in human cancers. Furthermore, integrin and EGF pathways have been identified to crosstalk during cell migration, so it is highly relevant to study them together. While a very large number (in the dozens, easily) of signaling proteins downstream of integrins and EGFR potentially involved in regulation of migration can be identified, our experimental measurements focus here on the following five which have been shown to be among the key molecular switches in the motility signaling

*To whom correspondence should be addressed.

cascades: EGFR itself, extracellular-regulated kinase (ERK), myosin light chains (MLC), protein kinase Cδ (PKCδ) and phospholipase Cγ (PLCγ). These signaling proteins play significant roles in driving major biophysical processes, such as lamellipod protrusion, cell/substratum attachment and detachment, and cell contractile force generation and transmission, which underlie the net cell migration behavior (Lauffenburger and Horwitz, 1996). Our experimental measurements are accomplished by quantitative immunoblotting, a standard but laborious procedure that typically limits the number of proteins and conditions which can be examined for any given situation under normal (at least academic laboratory) circumstances.

Many data-driven modeling approaches aim at finding correlations or cause–effect relations between genes or proteins. The resulting model is usually validated by comparing selected parts of the modeled relations with the literature or with additional biological experiments without considering how good the model is for predicting outcomes of biological processes. In contrast, we are seeking to achieve two objectives in our analysis of signal transduction cascades. The first objective is to build a model from which the most relevant signaling proteins in regard to response can be identified. The second objective is to assess prediction accuracy of the model. The algorithmic methodology with which we propose to accomplish these goals is decision tree modeling. Often, as in the present case, the experimental data are noisy and the amount of observations is inadequate for dependency modeling or prediction. Therefore, before analyzing the data with decision trees, the data should be preprocessed and, if the amount of the data is insufficient for robust analysis, interpolative simulation of additional, internally consistent data points might be considered as a computational aid.

The order of this study is as follows. First, we discuss an analysis of variance (ANOVA)-based quality control approach, a minimum description length (MDL)-based polynomial fitting method to simulate data points and prediction with decision trees. Second, we apply these approaches to a case study, where we aim at classifying cell migration speed using phosphorylation levels of five signaling proteins.

2 APPROACH

In this section, a strategy to analyze a signal transduction cascade in regard to a cellular outcome is presented. Depending on how the signaling protein activation levels are measured, the resulting dataset is practically always noisy. Therefore, data quality control and normalization are imperative along the course of signal transduction cascade analysis. In Section 2.1 we present a quality control protocol for replicate measurements. In Section 2.2, we create additional data points via interpolating polynomial models that are chosen according to the MDL principle and undertake validation efforts in Section 2.3. A procedure for constructing decision tree models is described in Section 2.4 and the experimental methods including data preprocessing are summarized in Sections 2.5 and 2.6.

2.1 Quality control

A topic that has been somewhat neglected in several systems biology studies is data quality control. The objective of the quality control step is to identify samples that are aberrant due to non-biological reasons (e.g. technical or measurement errors). If such outliers are not identified, they may confuse the analysis method and result in wrong conclusions. On the other hand, a stringent quality control criterion and discarding of outliers without careful consideration can cause loss of valuable information. Therefore, measures taken after identifying an outlier sample should be dependent on the reasons for the sample's aberrance.

Here, we present a statistical quality control algorithm for datasets consisting of multidimensional samples with replicates. Let vector $\mathbf{p}_i^j \in \mathbb{R}^{n\times 1}$ contain i-th set of replicate measurements for j-th sample and let r_j denote the number of the replicate measurement sets for j-th sample. Further, we define μ_i^j as the expected value of $\mathbf{p}_i^j$. Here n is the same across all the samples but the algorithm below allows missing values. In our case study, there are four fibronectin levels for each EGF level and since the EGF levels are dealt with separately until decision tree analysis, $n = 4$ for each $\mathbf{p}_i^j$. Now, outlier replicate samples can be found using the following ANOVA-based algorithm:

For j-th sample

(a) Test $H_0 : \mu_1^j = \mu_2^j = \cdots = \mu_{r_j}^j$ with a one-way ANOVA and perform a multiple comparison for the ANOVA results using the Tukey-Kramer test (Hochberg and Tamhane, 1987) with significance level α.

(b) If any of the replicate samples is aberrant, flag it according to the following rules:

R1 If a sample is statistically different from two or more samples, flag the sample.

R2 If there are several samples that could be flagged with R1, or two samples are statistically different, flag the sample whose deletion gives the minimum standard deviation for the means of the remaining samples.

Repeat (a) and (b) until H_0 is not rejected.

Repeat until all samples are processed.

The crux of the above algorithm is the ANOVA with the Tukey–Kramer multiple comparison test (Hochberg and Tamhane, 1987). In general, the following assumptions are needed for the ANOVA:

(1) Samples are independent.

(2) Variances are constant across the samples.

(3) Observations are approximately Gaussian distributed.

As the quality control algorithm is applied to identify outliers among replicates and the replicates are usually measured with the same or similar kind of apparatus, it is reasonable to assume that variances are approximately the same. Further, except in cases of failures to clean or calibrate the measurement apparatus after use, samples should be independent. The ANOVA is not very sensitive to violations of the normality assumption, so the normality assumption is not a major one. Moreover, often several independent sources affect the measurements and inline with the central limit theorem the data tend to be approximately normally distributed. The assumptions behind the ANOVA are usually fulfilled in biomedical research, so ANOVA-based quality control algorithm could be applicable to many experimental setups.

2.2 Parametric model for the data

An insufficient number of data points relative to the number of variables and interaction processes may impede or prevent identification of dependencies among the variables. A solution to this problem is to create a parametric mathematical model based on the data at hand, which is then used to interpolatively simulate additional data points

so that dependencies between variables can be modeled and used in prediction. It is imperative to emphasize that the objective of this approach is to merely generate multiple realizations of pseudomeasurements that are internally consistent with the statistical distribution of the actual measurements, rather than creating new information in an extrapolative manner.

If a preprocessed dataset consists of several replicates, it may be worthwhile replacing replicate observations with a single value that is the most plausible value given the data. This value is referred to as a point estimate. Traditionally, the point estimator is chosen to be (arithmetic) sample average because it is the best linear unbiased estimator for Gaussian distributed data and error estimates are straightforward to calculate. However, the breakdown point for the sample average estimator is $1/n$, where n denotes the number of data points, meaning that even one outlier might drastically affect the point estimate. This is highly undesirable and therefore we use the median, which has a breakdown point of $1/2$, as a point estimator.

One drawback with the median is that deriving error estimates analytically may be difficult. This drawback can be overcome with bootstrapping (Efron and Tibshirani, 1994): First, create B bootstrapping samples and compute median value for each bootstrapping sample. Error estimate for a point estimate is standard deviation of the bootstrapped medians.

Common trends for biological processes include linear, biphasic or asymptotically plateauing dependencies on a given variable. Several of these trends can be captured using polynomial models that have several benefits:

- Reliable polynomial modeling can be done with a relatively small sample size while still capturing highly non-linear trends.
- Polynomial modeling is not confounded by a few missing values.
- Discontinuous trends can be modeled with piecewise polynomials.
- Polynomial fitting procedures, such as least squares and maximum-likelihood methods, are included in practically every statistical modeling software.
- Simulation of the polynomial model is straightforward and fast.

For data simulation we assume the following model for i-th value of j-th observed variable such as the migration speed or a signaling protein:

$$g_2(y_{i(j)}) = f_{p(j)}(g_1(x_{i(j)})) + \epsilon_j, \tag{1}$$

where $x_{i(j)}$ denotes i-th experimental condition (e.g. the level of fibronectin and the absence or presence of EGF) for j-th variable, $f_{p(j)}(\cdot)$ is p-th order polynomial for j-th variable with parameters $\beta_{p(j)}^T = [\beta_{p(j)}, \beta_{p-1(j)}, \ldots, \beta_{0(j)}]$, $g_{1,2}(\cdot)$ are transformation functions and ϵ_j is an error term. For example, when $p = 2$ and g_1 and g_2 are identity functions, $y_{i(j)} = f_{2(j)}(x_{i(j)}) + \epsilon_j = \beta_{2(j)} \cdot x_{i(j)}^2 + \beta_{1(j)} \cdot x_{i(j)} + \beta_{0(j)} + \epsilon_j$. Although transformation functions are usually identity functions, sometimes it is beneficial to perform the fitting in log-space ($g_1(x_{i(j)}) = \log(x_{i(j)})$) or in log–log space ($g_1(\cdot) = g_2(\cdot) = \log(\cdot)$). In general, experimental conditions may vary between the variables and, therefore, quantities in Equation (1) depend on j. In the subsequent discussion, however, the subscript j is dropped for notational convenience.

The challenge with polynomial modeling is to choose the order of the polynomial (p) that describes the data best without overfitting. In order to solve this problem we use the MDL principle (Rissanen, 1978). The basic idea behind the MDL principle in model selection is to find the model that gives the minimum stochastic complexity relative to the model class (Rissanen, 1998). Stochastic complexity can be understood as a measure of the goodness-of-fit of a model based on the model's ability to compress the data, given a model class. As statistical inference is viewed as a data compression problem, there is no need to assume underlying, 'true' data generating distributions. Therefore apart from choosing the model class, there is no need to make subjective assessments.

To be more precise, we use normalized maximum-likelihood (NML) approach (Rissanen, 2000), which follows when the MDL principle is applied to the maximum-likelihood estimation. Let $\gamma \in \Omega$ be a restricted set of indices for the current polynomial order k and $X_\gamma \in \mathbb{R}^{n \times k}$ be a matrix of predictor values with indices γ. For example, when $k = 2$ (line fitting) in our case study, the first column of X_γ is $[0.1\ 0.3\ 1\ 3]^{\mathrm{T}}$ and the second, $\mathbf{1}$.

We assume that $\epsilon_j \sim \mathcal{N}(0, \tau)$, so the response data ($\mathbf{y} = y_1, \ldots, y_n$) are also Gaussian distributed with density function $f(\mathbf{y}; \gamma, \beta, \tau) = 1/(2\pi\tau)^{n/2} \exp(-1/(2\tau) \sum_i (y_i - \beta^{\mathrm{T}}\mathbf{x}_i)^2)$. Thus, maximum-likelihood solutions for a fixed γ are

$$\hat{\beta}(\mathbf{y}) = Z^{-1} X_\gamma^{\mathrm{T}} \mathbf{y}, \tag{2}$$

$$\hat{\tau}(\mathbf{y}) = \frac{1}{n} \sum_i (y_i - \hat{\beta}(\mathbf{y})^{\mathrm{T}} \mathbf{x}_i)^2, \tag{3}$$

where $Z = X_\gamma^{\mathrm{T}} X_\gamma = n \sum_\gamma$. In the subsequent discussion the subscript γ is dropped.

The maximum-likelihood estimates are used to obtain the NML density function

$$\hat{f}(\mathbf{y}; \gamma) = \frac{f(\mathbf{y}; \gamma, \hat{\beta}(\mathbf{y}), \hat{\tau}(\mathbf{y}))}{\int_{Y(\tau_0, R)} f(\mathbf{z}; \gamma, \hat{\beta}(\mathbf{z}), \hat{\tau}(\mathbf{z})) d(\mathbf{z})}, \tag{4}$$

where $\mathbf{y}$ is restricted to the set $Y(\tau_0, R) = \{\mathbf{z} | \hat{\tau}(\mathbf{z}) \geq \tau_0, \hat{\beta}(\mathbf{y})^{\mathrm{T}} \sum \hat{\beta}(\mathbf{y}) \leq R\}$. Parameters τ_0 and R are determined so that the maximum-likelihood estimates are within $Y(\tau_0, R)$.

The NML density function is unique solution to the minmax problem

$$\min_q \max_{\mathbf{y}} \ln \frac{f(\mathbf{y}; \gamma, \hat{\beta}(\mathbf{y}), \hat{\tau}(\mathbf{y}))}{q(\mathbf{y})}, \tag{5}$$

where q range over any distributions (Rissanen, 2000). Therefore, solving $\hat{f}(\mathbf{y}; \gamma)$ results in the best model for the data relative to the chosen model class. Evaluation of Eq. 4 (for details, see Rissanen, 2000) gives the final decomposition for finding the best polynomial order:

$$\min_{\gamma \in \Omega} \Big\{ (n-k) \ln(\hat{\tau}(\mathbf{y})) + k \ln(n\hat{R}) + (n-k-1) \ln\left(\frac{n}{n-k}\right) - (k+1) \ln(k) \Big\}, \tag{6}$$

where $\hat{R} = \hat{\beta}^{\mathrm{T}}(\mathbf{y}) \Sigma \hat{\beta}(\mathbf{y})$. In our case study, the NML criterion [Equation (6)] is used to find the best polynomial order (p).

After finding the best polynomial order, the parameters for that model are derived from Equation (2) and these are taken to estimate β. In addition to estimating the parameters for f_p in Equation (1), it is necessary to have an estimate for the standard deviation for ϵ_j. One way to get this is to first pool individual bootstrap error estimates:

$$s_j^{(\text{pooled})} = \sqrt{\frac{\sum_i^n (r_i - 1) \cdot s_i^2}{\sum_i^n r_i - n}}, \tag{7}$$

where s_i is a bootstrap error estimate and then either use $s_j^{(\text{pooled})}$ directly or squared.

2.3 Validation of the parametric models

After a parametric polynomial model is constructed with the NML procedure, it is useful to check how good the model is for the original measurements. Since we assume the data to be approximately Gaussian, the goodness of the model can be checked by considering a Gaussian distribution whose mean is the simulated value and the standard deviation is obtained via Equation (7). If each point estimate is located close to the mean and, for example, not above or below 2.5% of the right and left tails, the model can be considered statistically feasible.

It may also be useful to perform statistical tests such as the Z-test to test whether the point estimate (or original measurements) could originate from the model. If several point estimates belong to the extreme ends of the distribution, doubt may be cast over the validity of the model.

2.4 Finding dependencies between variables with the decision tree analysis

The majority of the studies in the field of systems biology aims at finding dependencies between variables. These models are, however, rarely used to predict the outcomes of cellular processes. In this section, we provide means to achieve both of these objectives with decision trees (Breiman *et al*., 1984). Decision trees have several advantages used in biomedical research:

(1) Decision trees can be effectively applied to any data structure, in particular to discrete, continuous or mixed data.

(2) Decision trees are capable of resulting in good prediction accuracies for highly non-linear prediction problems.

(3) Prediction rules are easy to interpret.

(4) Decision trees perform a stepwise variable selection and complexity reduction.

(5) Decision trees are very robust against outliers.

The basic idea behind the decision trees is to first identify prediction rules from the data and then illustrate them as a binary tree where each terminal node (leaf) corresponds to a class and the other nodes represent measured variables. An example of a rule is 'IF the phosphorylation level of ERK is high AND the phosphorylation level of MLC is high THEN cells migrate at medium speed.' This rule can be readily seen in Figure 5. The rules are constructed by recursively splitting the data into smaller and smaller regions so that after each split the new data subset is 'purer' than the old data subset (Breiman *et al*., 1984). A pure decision tree predicts all the classes in the training set correctly. In real world applications a pure (or close to pure) decision tree is very large and almost surely suffers from overfitting. Thus, a decision tree is usually constructed in two phases. The first phase, tree growing, is done until splitting does not significantly improve the measure of purity. The second phase, tree pruning, is done in order to avoid overfitting. Here, we use the cost–complexity pruning approach (Breiman *et al*., 1984) because we are able to create a separate pruning dataset. Briefly, the tree pruning phase starts with a very large (overfitted) decision tree. The cost–complexity pruning method selectively produces a sequence of subtrees until only the root node is included in the subtree. In the cost–complexity pruning approach the sequence of subtrees is achieved by minimizing the sum of misclassification cost and the complexity of the tree. For detailed discussion on the tree growing and the cost–complexity pruning methods we refer to Breiman *et al*. (1984).

In general, decision trees suffer from two drawbacks: masking and instability (Breiman *et al*., 1984). Masking may occur if the relation between class (migration speed) and measured variables (signaling proteins) is very complex. In this case a variable may be partially duplicated by another variable and if two variables result in almost equally pure subsets, the level of noise may govern which variable is used in the splitting. If this happens in the early phase of tree growing, the two decision trees may look dissimilar potentially impeding the interpretation of the results. In addition, masked variables may not show in the decision tree thereby hindering the understanding of the results. These drawbacks are further discussed in Section 3.2.

2.5 Signaling protein experiments

We used NR6 mouse fibroblasts for our studies. These cells are derived from the 3T3 lineage and are devoid of endogenous EGFR. We have overexpressed human EGFR in these cells, hence referred to as NR6 wild type (NR6 WT), and they provide an excellent model system to study EGFR mediated signaling events as well as cellular biophysical processes like migration. An equal number of NR6 WT cells were plated on fibronectin coated surfaces and allowed to grow in alpha modified eagle's medium containing 7.5% fetal bovine serum (FBS) for 24 h, by which time cells reached about 90% confluence. Fibronectin coating concentrations of the surfaces were 0.1, 0.3, 1 and 3 μg/ml (Fn $\in \{0.1, 0.3, 1, 3\}$). Subsequently, cells were quiesced in a medium containing 0.5% dialyzed (with minimum growth factors) FBS for another 24 h, to remove the effect of exogenous growth factors present in the serum. Cells were either lysed in the quiescent medium without any exogenous human EGF or stimulated with 10 nM (saturating concentration) of human EGF for 5 min. In the subsequent discussion, 0 nM EGF and 10 nM EGF conditions are denoted with EGF $= 0$ and EGF $= 1$, respectively. After stimulation, cells were washed once with ice-cold PBS, and then lysed in lysis buffer containing 50 mM HEPES, pH 7.4, 150 mM NaCl, 1% Triton X-100, 1 mM Na Vanadate and 10% glycerol supplemented with protease inhibitors including 1 μg/ml Leupeptin, 1 μg/ml Aprotinin and 1 mM Phenylmethylsulfonylfluoride (PMSF). Cell lysates were quantified using Biorad protein assay. An equal amount of total proteins were mixed with the loading buffer containing 4% SDS (w/v), 0.1 M Tris–HCl, pH 6.8, 20% glycerol, 0.2% Bromophenol blue and 5% β-mercaptoethanol, boiled for 5 min and then loaded on either 7.5% (for analysis of pPKCδ, pERK, pEGFR, pPLCγ) or 15% (for pMLC) SDS polyacrylamide gels. Cell lysates were resolved by electrophoresis and subsequently transferred onto nitrocellulose membranes, after which membranes were immunoblotted with specific antibodies to detect the specific proteins or their activated phospho-protein forms.

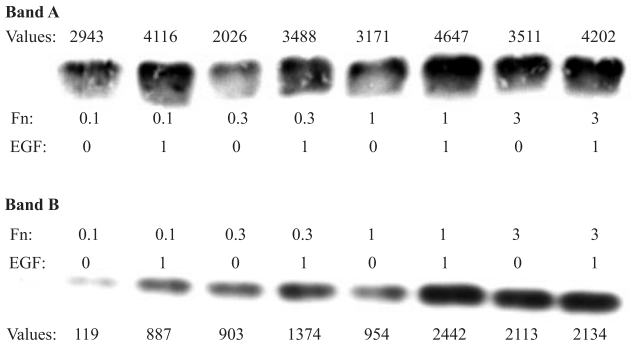

Fig. 1. Two immunoblotting bands for MLC. All Fn and EGF conditions with the non-normalized values are shown.

2.6 Data preprocessing

Immunoblots were quantified with the NIH image analysis densitometry software. The software generates a plot area for each protein band, the density of which represents the amount of the protein in each lane. In the signaling protein experiments, the quantitative values generated represented the activated status of a protein since the proteins detected were in their activated or phosphorylated state. Examples of two immunoblotting bands are given in Figure 1.

For the NML and decision tree analysis, the band densities were normalized by the value of the first lane (Fn $= 0.1$ and EGF $= 0$) for each immunoblot (between-band normalization). After this normalization, results become comparable between immunoblots since the experimental conditions in each of the experiments were kept constant. For quality control, the bands were within-band normalized: all protein conditions in a band without exogenous EGF were normalized by the value with EGF $= 0$ and Fn $= 0.1$, while all protein conditions in a band with exogenous EGF were normalized by the value with EGF $= 1$ and Fn $= 0.1$. The within-band normalization ensures that proteins under the same EGF condition within a band are comparable. Prior to normalization all basal values <250 were converted to 250 in order to prevent division by a small value that is likely due to noise. After normalization, all the values were $\log_2$-transformed.

Normalization was followed by the ANOVA-based quality control approach (Section 2.1) with $\alpha = 0.05$. An example of a quality control plot for MLC is given in Figure 2. Replicate 8 (marked with a star) is aberrant from the seven other replicates and is discarded. Also replicate 2 is discarded due to rule R1 given in Section 2.1. Replicates 1 and 8 correspond to bands A and B in Figure 1, respectively. The numbers of the replicates before and after the quality control are given in Table 1.

3 RESULTS

Cell migration is a crucial cellular function that contributes, for example, to wound healing and normal immune responses. On the other hand, migration drives progression of diseases such as tumor invasion and metastasis (Ridley *et al*., 2003). In general, migration consists of a complex assembly of five biophysical processes: polarization, protrusion, adhesion, contractility and retraction. While the effects of biophysical processes to migration speed are somewhat well-known, the effects of signaling proteins that govern these processes and their dependencies are less so. In this section, we explore

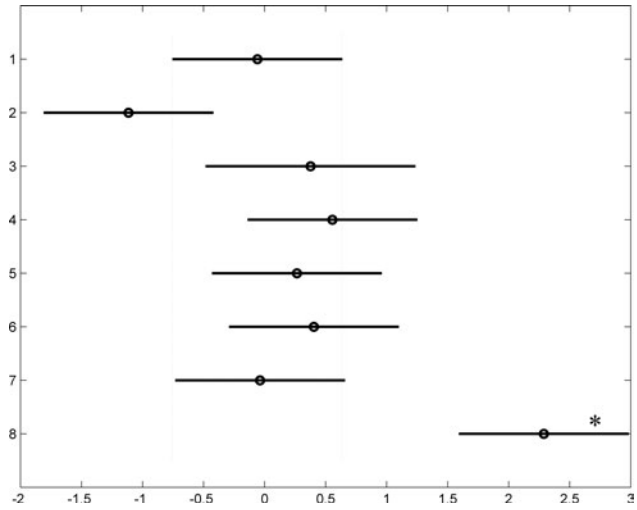

Fig. 2. An example of quality control plot for MLC (EGF $= 0$). A circle denotes a mean and a line corresponds to a comparison interval. Replicate 8 (marked with a star) is aberrant from all the other replicates and is discarded. Replicates 1 and 8 correspond to bands A and B, respectively, in Figure 1.

Table 1. The number of the replicates for signaling proteins before and after the quality control

Protein	EGF $= 0$	EGF $= 1$	Total
EGFR	7	7	8
ERK	5	8	8
MLC	6	5	8
PKCδ	7	6	8
PLCγ	5	5	5

how five signaling proteins (EGFR, ERK, MLC, PKCδ and PLCγ) affect cell migration speed under combinations of two extracellular cues using the methods discussed in Section 2.

The cues used here are four different surface fibronectin concentrations with or without additional stimulation with EGF. Earlier studies (Maheshwari *et al*., 1999) have shown that if migration speed is measured as a function of fibronectin levels, presence or absence of EGF has a dramatic impact on migration speed. If EGF is present, migration speed is biphasic, while in the absence of EGF, cells migrate at a constant speed. The data in Maheshwari *et al*. (1999) consist of four fibronectin-levels for 0 and 25 nM EGF, resulting in eight measurements. Accordingly, we measured the phosphorylation levels of the five signaling proteins using the same condition for $EGF = 0$ (0 nM EGF) as in the study by Maheshwari *et al*. (1999). For condition $EGF = 1$, we used 10nM EGF for the signaling proteins, while it was 25 nM (Maheshwari *et al*., 1999). Since both 10 and 25 nM EGF are identical in motility and both are saturating, the data for signaling proteins and cell migration speed are comparable.

3.1 Parametric model for the signaling proteins and migration speed

Having four observations per one EGF-level is not enough for reliable identification of dependencies between the signaling proteins and migration speed. Therefore, we applied the procedure given in

Table 2. Polynomial estimates ($\hat{\beta}$) and standard deviation estimates ($\hat{\sigma}$) for signaling proteins under presence and absence of EGF using the NML criterion

Protein	$\hat{\beta}_{\mathrm{EGF}=0}$	$\hat{\sigma}_{\mathrm{EGF}=0}$	$\hat{\beta}_{\mathrm{EGF}=1}$	$\hat{\sigma}_{\mathrm{EGF}=1}$
EGFR	$0.24x - 0.02$	0.10	1.9	0.15
ERK	0.51	0.25	4.1	0.20
MLC	0.20	0.04	$0.08x + 0.48$	0.04
PKCδ	0.06	0.06	0.32	0.04
PLCγ	$0.36x - 0.86$	2.6	$0.22x + 2.38$	0.38

Section 2.2 to generate additional, interpolative data using simulation. Before applying the NML approach, we computed median phosphorylation levels for the signaling proteins using the replicate measurements. Each median value was accompanied with an error estimate that was computed with bootstrapping ($B = 5000$).

Due to the small number of the data points we restricted the maximal polynomial degree in the NML approach to 2, i.e. $\Omega = \{0, 1, 2\}$. Further, the polynomial models were constructed separately for the values under the conditions of EGF $= 0$ and EGF $= 1$. Polynomial orders for the signaling proteins using the NML criterion are given in Table 2. An example of the polynomial models and associated point and error estimates for MLC and ERK is given in Figure 3.

Standard deviations and point estimates for migration speed are given in Maheshwari *et al*. (1999). We computed pooled standard deviation with Equation (7), where we made a conservative approximation that $r_j = 70$ since the estimates were based on 70–100 cells. The pooled standard deviations were 3.0 for EGF $= 1$ and 3.4 for EGF $= 0$. Polynomial fitting for migration speed was done in log–log space based on the model validation procedure depicted in Section 2.3. For EGF $= 1$ the polynomial order was two ($-0.49x^2 + 0.07x + 5.8$), while for $EGF = 0$ it was zero (4.1). As our objective is to predict slowly, at medium speed or fast migrating cells, the values were further discretized into three categories ({slow, medium speed, fast}) using the Lloyds algorithm (Lloyd, 1982), where the training data were obtained from the noiseless polynomial model. The model, discrete categories, simulated data and the original measurements for migration speed are shown in Figure 4.

3.2 Decision tree for migration speed

Using the polynomial models we simulated observations between Fn $= 0.1$ and Fn $= 3$ using $\Delta = 0.0001$ resulting in 58 002 observations per variable. The protein phosphorylation values were discretized with the Lloyds algorithm so that the number of the discrete categories equaled the number of parameters in the polynomial models for each protein: EGFR $= \{0, 1, 2\}$, ERK $= \{0, 1\}$, MLC $= \{0, 1, 2\}$, PKCδ $= \{0, 1\}$ and PLCγ $= \{0, 1, 2, 3\}$. With this discretization approach the proteins that are affected more by the extracellular stimuli, and thereby considered more informative, are described with more discretization categories than the proteins with lower polynomial degrees. Discrete categories reflect relative phosphorylation levels. For example, ERK and MLC have two and three discrete categories, respectively. Thus, ERK $= 1$ denotes that ERK is highly phosphorylated, while MLC $= 1$ means that the phosphorylation level of MLC is medium.

In order to overcome the instability problem with the decision trees we first constructed 10 000 decision trees without pruning. The parameters for growing the decision trees were as follows. Splitting criterion was Gini-index (Breiman *et al*., 1984), prior probability for i-th class was obtained by dividing the number of the cases of i-th class by the total amount of observations and the minimum number of observations for impure nodes to split was set to be five. We defined the misclassification costs to be such that for misclassifying a slow (medium) speed to medium speed (fast) the cost is one, but if slow speed is misclassified to fast, the cost is two. This resulted in the following cost matrix

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix},$$

where each row and column corresponds to a migration speed category.

After the tree growing phase, all 10 000 trees were pruned with the cost–complexity pruning method (Breiman *et al*., 1984). For the validation step we created 1000 datasets. For computational reasons Δ was set to be 0.001, so each validation and pruning dataset consisted of 5802 observations. The following criterion was used to choose the best tree model:

$$D = \frac{1}{T} \sum_{i}^{T} |y_i - \hat{y}_i|, \tag{8}$$

where y_i denotes true classes for i-th test dataset, $\hat{y}_i$ denotes predicted classes and T is the number of test datasets (here 1000). After the pruning and validation, there were 23 separate decision tree models and the best decision tree was the one that minimized Equation (8). Although the best decision tree is chosen using Equation (8), we report also the mean classification accuracy (CA), which is the mean of the classification accuracies across 1000 test cases. Classification accuracy corresponds to the number of correct classifications divided by all cases.

The best decision tree (CA $= 70\%$) is given in Figure 5. Round nodes correspond to the signaling proteins and square nodes to the migration speed classes. Classification rules and their relative importance can be seen easily from the decision tree in Figure 5. For example, IF ERK $= 1$ AND MLC $= 1$ THEN cells migrate fast, and 62% of the measurements for the fast migration class (in the training set) can be explained by this rule.

If the signaling proteins were not discretized, the best decision tree consists of only MLC and PLCγ (data not shown) and CA was slightly below 70%. Based on this decision tree graph, it could be argued that cell migration speed is dependent only on MLC and PLCγ and ERK is irrelevant when predicting cell migration speed. However, earlier studies have shown that ERK is one of the key signals governing migration speed (Glading *et al*., 2000; Matsubayashi *et al*., 2004; Webb *et al*., 2004), so its absence in the decision tree model was unexpected. When we looked for explanations for the exclusion of ERK we found that ERK was masked by MLC. This can be seen by comparing data for ERK and MLC in Figure 3. When $EGF = 0$, both ERK and MLC are constant with approximately the same level of phosphorylation. However, when $EGF = 1$, phosphorylation levels for ERK are again almost constant but very high, whereas MLC activity is increasing linearly. Thus, the decision

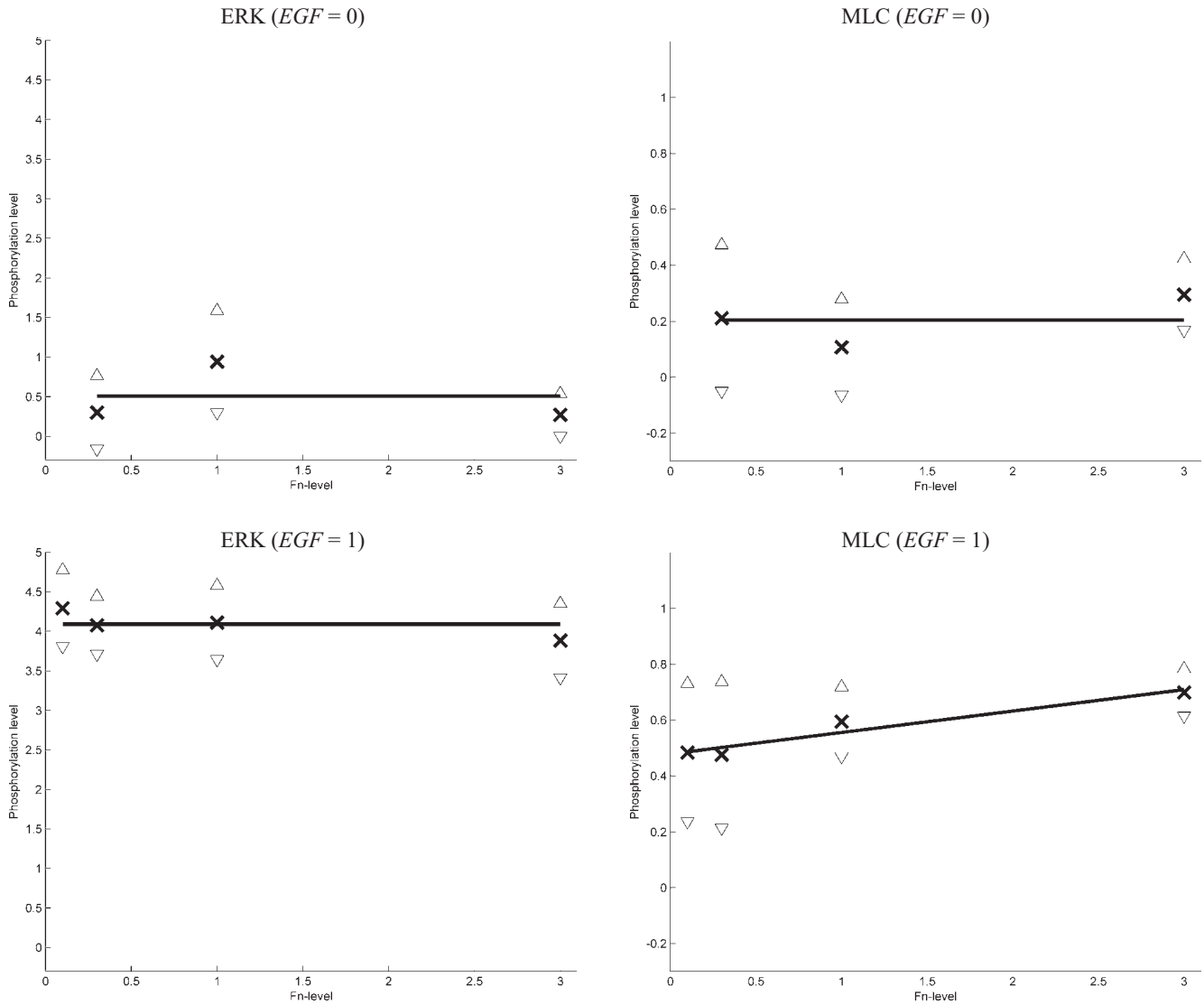

Fig. 3. Point estimates, cross; upper, triangle and lower, inverted triangle error estimates and fitted polynomial for ERK and MLC across the fibronectin levels.

tree growing algorithm considered ERK was not of use in migration speed classification given MLC, which was undesirable. While some guidelines for detecting masked variables are given in Breiman *et al*. (1984), these are not helpful in getting the masked variable into the model.

This case study provides a proof-of-principle that the proposed data-driven modeling approach is applicable to biomedical research. Accordingly, detailed discussion on biological implications of the results is out of the scope of this study and will be elaborated elsewhere (manuscript in preparation). Briefly, ERK and MLC regulate the adhesion/contraction ratio (Iwabu *et al*., 2004; Webb *et al*., 2004), which is one of the most important biophysical processes during the migration cycle (Lauffenburger and Horwitz, 1996). Thus, it is not surprising that these two proteins together represent useful predictors for migration speed. This result also suggests that in further studies, ERK and MLC should be studied together rather than individually.

4 DISCUSSION

Analysis of signal transduction cascades is an important application in several biomedical research studies. In this study we have presented a data-driven modeling approach to perform such analysis. In our case study we have applied the modeling approach to model and predict whether cells are moving slowly, at medium speed or fast, using a set of intracellular signaling proteins under various levels of fibronectin and EGF cues. The resulting decision tree graph indicates that the phosphorylation level of ERK alone shows whether cells are migrating slowly. In order to obtain higher classification accuracy for the cells migrating at medium speed or at fast speed, MLC, PLCγ and PKCδ are also needed. These results highlight the central idea of systems biology, i.e. complex biological processes cannot be analyzed by perturbing only one component at a time but there is a need to study several components simultaneously. However, usually it is not known what these components are. Our results indicate that

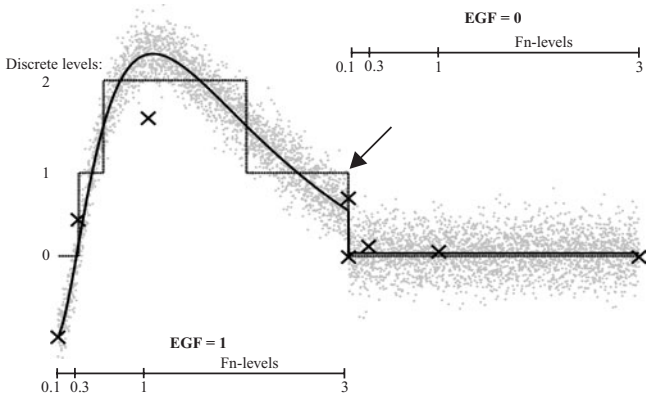

Fig. 4. Polynomial fit (solid line), discretization categories (dotted line), original observations (cross) and simulated, noisy data (dots) for migration speed. The change from $EGF = 1$ to $EGF = 0$ is marked with an arrow.

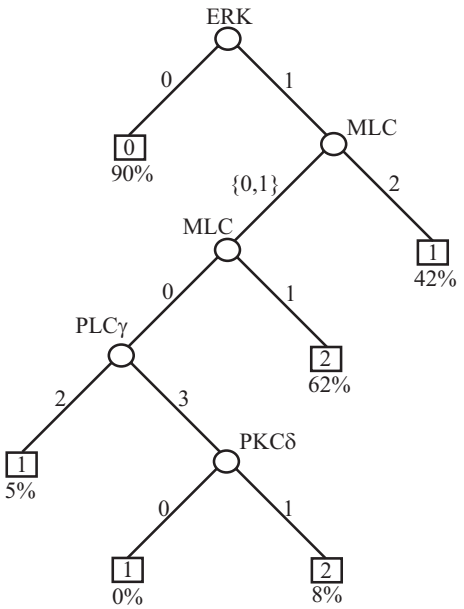

Fig. 5. The best decision tree for classifying migration speed using signaling proteins. Round nodes denote variables, while square nodes represent estimated migration speed categories. Integers attached to the arcs correspond to the splits of the parent node. Under each migration speed category the fraction of cases explained by that classification rule is given. For example, if ERK $= 0$, the migration speed category is 0, and 90% of the observations (in the training set) for migration speed category 0 can be explained by this rule.

the decision tree analysis can be used to suggest what components should be studied together.

Based on Table 2 it can be argued that the information content of the signaling protein dataset is low. Degrees of the polynomial functions, however, do not tell the whole story. For example, ERK has zero order polynomial in the presence and absence of EGF, but the absolute difference between these constants is large (3.6). That is, ERK acts like a switch triggered by the EGF status and clearly brings in information to the analysis. Further, after inducing EGF, cell migration measurements were performed after 8 h in order to observe maximal migratory response for the cell type used in this study (Maheshwari *et al*., 1999). Therefore, it was expected that in 5 min the changes in the signaling proteins phosphorylation levels may not have been visible at the migration speed level. From another standpoint, the conclusion that a dataset is not rich in its information content may be valuable information and the NML modeling with decision tree analysis provides means to assess this issue. One of our future goals is to measure signaling protein activities in regard to cell migration speed at different time points, and the methods given here can be used to approximate the time it takes for extracellular stimuli to have an effect on the signaling proteins. These results may be used when estimating rate constants for temporal mathematical modeling.

Quality control is an inevitable part of biomedical research; studies not performing quality control implicitly state a 100% confidence in the measurements. In this study we applied a statistical quality control approach for replicate measurements with 95% confidence level. We also tested 80% confidence level and no quality control and the resulting decision trees resulted in 62 and 67% classification accuracies, respectively (data not shown). Detailed discussion on quality control issues and choice of the confidence levels is beyond the scope of this study, but it would be an interesting topic for further study. In essence, one of our aims for the future is to first identify quality features, learn a classifier with them and use the trained classifier to assess quality control as described in Hautaniemi *et al*. (2003).

The overall classification accuracy for simulated data, 70%, is quite good given that original signaling protein measurements were done at one time point of only 5 min after stimulation, whereas cell migration was measured at 8 h. Furthermore, cell migration speed data and signaling protein measurements originate from different studies: The cell migration dataset was done in 1999 while the signaling protein dataset was done in 2004. As a consequence, there are some differences between the experimental setups causing noise to the analysis. Classification accuracy can also be used as a yard stick for sufficiency of the measured dataset in regard to modeling a biological process. If classification accuracy is poor, it could be an indication that the dataset does not comprise enough variables or information in order to model the biological process in question. In our case study, the classification accuracies for medium and high migration speeds are fair. The most likely reason for this is that our measurements cover only a limited portion of the signaling network components critically involved in governing migration. Merely as one relevant facet of this highly multi-variate system, for instance, there is accumulating evidence that virtually all of the key MAP kinases influence cell motility in diverse ways (Huang *et al*., 2004). This shortcoming can, of course, be addressed by enlarging the scope of the measured signaling component space to the extent cost-effective. Accordingly, the decision tree results can be helpful to determine what components should be measured in future experiments, and whether there is a need to measure additional components. On the other hand, our results demonstrate that decision trees are applicable to studies where several key components are not observed.

To our knowledge, this is the first study where cell migration speed is quantitatively predicted using phosphorylation levels of signaling proteins. Several other modeling approaches such as Bayesian networks (Pearl, 1988; Woolf *et al*., 2004), neural networks or support vector machines (Haykin, 1999) have their own benefits and drawbacks. The two latter methods are very good classifier approaches in various applications but they suffer from a major

drawback—dependencies between the variables and their relevancies are practically impossible to obtain from the model. In contrast, Bayesian networks have been mainly used to obtain dependencies between variables but it is not self-evident that a Bayesian network that describes dependencies between the variables performs well when predicting cellular outcomes. Moreover, the variables used in learning a Bayesian network are usually required to be discrete or Gaussian distributed, which may be an impractical requirement.

The decision tree based modeling is not supported by a unique and solid mathematical background. Thus, it is imperative to report parameter settings in detail so that the results can be reproduced. Furthermore, decision trees require a relatively large training dataset, which may not be feasible. This requirement, however, is not unique to decision trees but is present with the other classification and modeling approaches as well. Here we have expanded the dataset via interpolating polynomial functions whose order was determined with the MDL principle. Parameters for polynomial functions are straightforward to estimate and several well-established methods exist for this purpose. When polynomial functions do not yield satisfactory results, the alternative might be Monte Carlo based techniques. In some cases, however, it may be difficult to identify probability density functions that describe the system to be modeled. Additionally, Monte Carlo methods are notorious for being computationally demanding. Therefore, we argue that the polynomial models should be applied before considering more complex methods.

When additional data are simulated, it is important to choose the extracellular conditions so that they span a large range because it is safer to interpolate than extrapolate. Another requirement is that there should be enough data points so that non-linear trends can be captured. The methods presented in this study do not pose upper limits for the extracellular cues but in its current form the decision tree analysis can be applied to only one biological process at a time. One of our future goals is to develop a multidimensional decision tree that is capable of predicting several cellular outcomes simultaneously. A multidimensional decision tree would be able to, for example, identify signaling proteins that are associated with high cell migration speed and absence of apoptosis.

5 CONCLUSIONS

We have presented a decision tree-based modeling approach for analysis of complex and multidimensional signal transduction cascades. Our case study demonstrates that decision trees can provide several insights to signal transduction cascades. We conclude that decision tree methodology may facilitate elucidation of signal–response cascade relationships and generate experimentally testable predictions, which can be used as directions for future experiments.

ACKNOWLEDGEMENTS

We thank Dr Fei Hua for constructive suggestions regarding the manuscript. This work was supported by the NIGMS Cell Migration Consortium, NCI grant CA88865 to DAL and the Academy of Finland and Emil Aaltonen Foundation.

REFERENCES

Breiman,L., Friedman,J., Olshen,R. and Stone,C. (1984) *Classification and Regression Trees*. Wadsworth.

Efron,B. and Tibshirani,J. (1994) *An Introduction to the Bootstrap*. Chapman & Hall, London.

Glading,A. *et al*. (2000) EGF receptor activation is required for fibroblast motility and occurs via an ERK/MAP kinase signaling pathway. *J. Biol. Chem.*, **275**, 2390–2398.

Hautaniemi,S. *et al*. (2003) A novel strategy for microarray quality control using Bayesian networks. *Bioinformatics*, **19**, 2031–2038.

Haykin,S. (1999) *Neural Networks: A Comprehensive Foundation*, 2nd edn. Prentice Hall, Inc., Upper Saddle River, NJ.

Hochberg,Y. and Tamhane,A. (1987) *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Huang,C. *et al*. (2004) MAP kinases and cell migration. *J. Cell Sci.*, **117**, 4619–4628.

Iwabu,A. *et al*. (2004) Epidermal growth factor induces fibroblast contractility and motility via a protein kinase cδ-dependent pathway. *J. Biol. Chem.*, **279**, 14551–14560.

Lauffenburger,D. and Horwitz,A. (1996) Cell migration: a physically integrated molecular process. *Cell*, **84**, 359–369.

Lloyd,S. (1982) Least square quantization in PCM. *IEEE Transactions on Information Theory*, IT-**28**, 129–137.

Lodish,H., Berk,A., Matsudaira,P., Kaiser,C., Krieger,M., Scott,M., Zipursky,S. and Darnell,J. (2004) *Molecular Cell Biology*. W.H. Freeman & Co, New York.

Maheshwari,G. *et al*. (1999) Biophysical integration of effects of epidermal growth factor and fibronectin on fibroblast migration. *Biophys. J.*, **76**, 2814–2823.

Matsubayashi,Y. *et al*. (2004) ERK activation propagates in epithelial cell sheets and regulates their migration during wound healing. *Curr. Biol.*, **14**, 731–735.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Interference*. Morgan Kaufmann Publishers Sam Mateo, CA.

Ridley,A. *et al*. (2003) Cell migration: Integrating signals from front to back. *Science*, **302**, 1704–1709.

Rissanen,J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465–471.

Rissanen,J. (1998) *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

Rissanen,J. (2000) MDL denoising. *IEEE Transactions on Information Theory*, **46**, 2537–2543.

Webb,D. *et al*. (2004) FAK–Src signalling through paxillin, ERK and MLCK regulates adhesion disassembly. *Nat. Cell Biol.*, **6**, 154–161.

Wells,A. *et al*. (2002) Growth factor-induced cell motility in tumor invasion. *Acta Oncologica*, **41**, 124–130.

Woolf,P. *et al*. (2004) Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics* (in press).