BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 14 2012, pages 1902–1910
doi:10.1093/bioinformatics/bts290

Systems biology

Advance Access publication May 15, 2012

Identifying mechanistic similarities in drug responses

Chen Zhao[1,2], Jianping Hua[2], Michael L. Bittner[2], Ivan Ivanov[3], and Edward R. Dougherty[1,2,4,*]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843, [2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, 85004, [3]Department of Veterinary Physiology and Pharmacology, Texas A&M University, College Station, TX, 77845 and [4]Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston, TX, 77030, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In early drug development, it would be beneficial to be able to identify those dynamic patterns of gene response that indicate that drugs targeting a particular gene will be likely or not to elicit the desired response. One approach would be to quantitate the degree of similarity between the responses that cells show when exposed to drugs, so that consistencies in the regulation of cellular response processes that produce success or failure can be more readily identified.

Results: We track drug response using fluorescent proteins as transcription activity reporters. Our basic assumption is that drugs inducing very similar alteration in transcriptional regulation will produce similar temporal trajectories on many of the reporter proteins and hence be identified as having similarities in their mechanisms of action (MOA). The main body of this work is devoted to characterizing similarity in temporal trajectories/signals. To do so, we must first identify the key points that determine mechanistic similarity between two drug responses. Directly comparing points on the two signals is unrealistic, as it cannot handle delays and speed variations on the time axis. Hence, to capture the similarities between reporter responses, we develop an alignment algorithm that is robust to noise, time delays and is able to find all the contiguous parts of signals centered about a core alignment (reflecting a core mechanism in drug response). Applying the proposed algorithm to a range of real drug experiments shows that the result agrees well with the prior drug MOA knowledge.

Availability: The R code for the RLCSS algorithm is available at http://gsp.tamu.edu/Publications/supplementary/zhao12a.

Contact: edward@ece.tamu.edu

Received on January 21, 2012; revised on May 4, 2012; accepted on May 10, 2012

1 INTRODUCTION

The ability to measure the abundance and degree of modification of many macromolecules in cells has allowed researchers to examine cells for molecular characteristics that indicate susceptibility to particular drugs (Chen *et al.*, 1991; Scherf *et al.*, 2000; Sirota *et al.*, 2011). This approach has produced a number of very useful guidelines to the use of therapeutics; however, it has failed in instances where the drug induces changes in the type and/or abundance of proteins that either pump drugs out of the cell or allow the drug-targeted activity to be provided in an alternative way that is not affected by the drug (Sergina *et al.*, 2007; Yusuf *et al.*, 2003). For these reasons, the ability to examine the molecular dynamics of cells' responses to drugs becomes of primary interest. In addition, identifying dynamic patterns will help to detect whether drugs targeting a particular gene produce the desired response. One possible way to achieve this goal would be to develop a way to quantitate the degree of similarity between the responses that cells show when exposed to drugs, so that consistencies in the regulation of cellular response processes that produce success or failure can be more readily identified.

1.1 Analysis of gene transcription dynamics

A considerable amount of research using fluorescent proteins as transcription activity reporters has examined transcription in living cells in both single-cell and multicellular organisms (Chalfie *et al.*, 1994; Hunt-Newbury *et al.*, 2007). Since fluorescent imaging can be performed in ways that do not destroy cells, fluorescent reporters are very effective tools when studying the time evolution of gene expression. Inserting DNA cassettes with a particular promoter driving expression of a fluorescent protein into egg cells or partially differentiated intermediate cells in ways that allow the cassette to become incorporated in the cell's genome allows one to follow that cell and its daughter cells' developmental course. This kind of information can be used to specify in which cell types and in how many cells a specific gene is active throughout the stepwise course of development and provides clues to gene function. By using a very similar approach on cells with reporters responding to drugs, it is possible to determine which and how many cells are altering the transcription level of a given gene during the course of the cell population's response to the drug. As cells' responses to drugs can take days to run their course, it is expected that a drug that mobilizes a change in the transcriptional regulation of a gene or genes in a cell will produce a distinguishable temporal trajectory of change in both the level of transcriptional change in cells and the number of cells showing altered expression level in a population of treated cells. It is further expected that sets of drugs that induce the same or a very similar alteration in transcriptional regulation will produce similar temporal trajectories on many of the reporter proteins, allowing them to be identified as having similarities in their mechanisms of action (MOA).

*To whom correspondence should be addressed.

Note that this trajectory is conditioned by the presence and activity status of the gene product that the drug is targeted to interact with, as well as those gene products affected by the activity of the targeted gene. This contextual conditioning insures that the temporal trajectories induced by a drug at a specific dose will only be highly similar across cells having highly similar contexts of the drug responsive genes. Similarly, a second drug targeted to the same gene would be expected to provoke a highly similar response at a dose of similar potency in any of the set of cell lines that respond similarly to the first drug if the interaction between the second drug and the gene was very similar to the interaction of the first drug with the gene. Should this be true, the two drugs could be said to have the same MOA for cells bearing the shared context. This kind of analysis is not intended for the type of large screens associated with drug discovery, where a single type of outcome such as reduced cell growth rate or cell death at one or two time points is used. It is intended to be used on candidate and existing drugs to develop a more specific characterization of the processes affected by the drug and the particular molecular context that allows the drug to produce the desired cellular consequences.

The idea of using gene expression data to explore drug relationships was also explored in Lamb *et al.* (2006). Using Gene Set Enrichment Analysis, compounds whose gene expression patterns correlate strongly with a query signature are considered to be similar or to have strong connections in the connectivity map, a collection of messenger RNA (mRNA) expression data from multiple cell lines treated with bioactive small molecules. Iorio *et al.* further developed this idea by building a drug similarity network and identifying network communities using graph theory. Hu and Agarwal and Sirota *et al.* extended the idea for drug repositioning by paring drugs and diseases whose gene expression patterns are negatively correlated. There are several major differences between their work and our approach. First, the mRNA data used in their studies are only collected at one time point (6 or 12 h) after the compound treatment. There is no dynamic or temporal data. In our experience, temporal data are critical in comparing the detailed MOA of different drugs. Many drugs do not even show a distinguishable response until 15 h into the experiment (Fig. 7). Therefore, the compound connections discovered in Hu and Agarwal (2009); Iorio *et al.* (2010); Lamb *et al.* (2006); and Sirota *et al.* (2011) could be strongly time dependent. Second, similarity requirements are different. For example, in Lamb *et al.* (2006), two compounds with distinct MOAs could be viewed as similar if they share similar clinical indications; however, in our case, two compounds are considered similar only if they interact with a similar set of pathways. And the similarity is cell line dependent. In this sense, our requirement for similarity is much more stringent and specific. Finally, we note that to use the proposed method, a set of putative pathways affected by the testing drugs should be formulated first and subsequently a set of reporter proteins should be selected to monitor the corresponding pathway activities. The method is hypothesis-driven and requires prior knowledge.

In our adaptation of this methodology, imaging is performed using a robotic imaging device (ImageXpress Micro, Molecular Devices). Multichannel imaging of sets of adherent cells with various reporters cultured in 384 well plates can be performed every hour, and at typical initial cell loading levels, these cultures can be followed for 50 h. Two typical fluorescent images are shown in Figure 1a and b, where nuclei are detected in the blue channel and promoter reporters are detected in the green channel. The objective of image processing is to extract gene-expression levels from the population of cells in the fluorescent image and follow these intensity distributions over time. To do so, we use morphological image processing, in particular, the watershed transformation (Dougherty and Lotufo, 2003). Although the image processing is rather involved, overall it breaks down into three major components: nuclei channel segmentation reporter channel segmentation, and measurement of cell-by-cell promoter activity levels. Briefly, to extract the promoter activity level for each cell, one needs to first identify the position of cells in the image and then identify the area of the image covered by each cell's cytoplasm. As the cells can be either compact or spread out, we estimate the promoter activity by measuring the sum of the green fluorescent protein (GFP) intensity (arbitrary camera intensity units) in all the pixels within the cell area and report the log2 transformed, summed intensity values for each cell. To achieve this, we first process the nuclei (blue) channel to locate all nuclei present in the image, and then process the reporter (green) channel to determine the activity level of the reporter for each cell. We refer to Hua *et al.* (2012) for full imaging details.

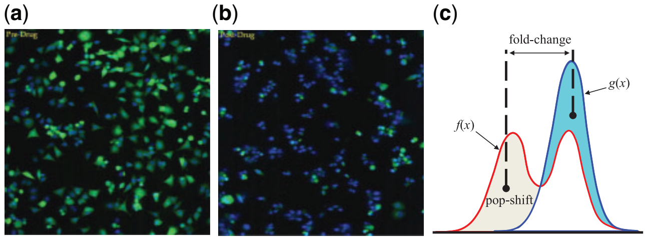


Fig. 1. Two typical fluorescent images for cell-line HCT116 with a promoter reporter for the gene MKI67: (**a**) before any drug is applied (control or pre-drug case); (**b**) 43 h after the drug Lapatinib was added (post-drug case); (**c**) calculation of population change/shift and fold change: $g(x)$ and $f(x)$ represent the log2 GFP intensity distributions for the cells in (a) and (b), respectively

Live cell imaging analysis provides two distinct types of cell-by-cell information that are not easily measured over long time spans by other means: (1) the extent of change in promoter activity in the treated population relative to that of the untreated, control population and (2) the percentage of cells in the treated population shifted into a position in the expression level distribution not occupied by the untreated control population, as a consequence of drug activity. An example of how these two measurements are calculated is presented in Figure 1c, where $g(x)$ (control or pre-drug case) and $f(x)$ (post-drug case) represent the log2 GFP intensity distributions for the cells in Figures 1a and b, respectively. The percentage of population change can be calculated by the difference in area between $g(x)$ and $f(x)$ (the gray area in Fig. 1c). Similarly, the fold change can be calculated as the mean difference between the shifted cells and the control case. Note that when fold change is positive, the corresponding population change will be denoted as positive and when fold change is negative, the corresponding population change will be denoted as negative.

1.2 What information on mechanistic similarity is available in drug response trajectories?

In order to produce metrics of comparison for the similarity of transcription responses induced by drugs, a model of how cellular responses to a drug will shape the trajectories is required. A conceptual model (Fig. 2) of drug response by transcription reporters in one molecularly homogeneous cell line responding to a series of drugs that target protein regulators acting on pathways of interest facilitates a simplified consideration of the informational content of a trajectory. In this example, we start with a set of four genes (*A*, *B*, *C* and *D*) that are suspected of contributing to the regulation of a cellular process that we wish to inhibit. We know that a set of genes (*E*, *F* and *G*) are all strongly expressed when this process is operational and that suppression of expression of gene *G* produces a reduction in cell proliferation, a desired result of intervention, in these cells. We also have a series of drug compounds, 1, 2, 3 and 4, known to interfere in activation of the transcription factors *B*, *A*, *D* and *C*, respectively. In such a setting, Figure 2a, the important question to address is whether an examination of the dynamics of response to each drug by promoter reporters for genes *E*, *F* and *G* would produce sufficient understanding of the process mechanisms to determine which genes are driven by a similar regulatory mechanism? As changes in the number of cells making a regulatory decision that leads to altered expression levels allow fairly intuitive interpretation of the dynamics, we will examine this aspect of the conceptual model to illustrate the mechanistic characteristics of the applied drugs that can be inferred through this approach.

If a technical replicate (TR) had been run, examining the effect of drug 2 on the gene *F* reporter, we would expect the kind of trajectories labeled D2 and D2′ in Figure 2b. (Levels of similarity for replicates in actual experiments are shown in Fig. 6.) If the cell line that these drugs are being tested on is molecularly homogeneous, repeated testing should produce very similar timings of when the cell line will show a detectable amount of population change, how rapidly the population-change trajectory increases and how many cells respond to the drug eventually. These three characteristics, time of onset of detectable transcription alteration, rate of population change increase and final percentage of responded cells, define a dynamic population response signature that can be systematically compared across a variety of drugs.

If two processes have no overlapping use of components, then the effects of drugs targeting one of the processes should not produce a response from members of the other process. The expected result for this situation is shown in Figure 2c. Drug 4 affects Process 2 (gene *C*), but not Process 1 (genes *D*, *B* and *A*), so drugs acting on Process 1 produce no effects on the Process 2 reporter. Similarly, Figure 2d shows that all the three drugs acting on Process 1 produce changes in the furthest downstream reporter for Process 1 and have no effect on the Process 2 reporter.

When drugs are acting on different parts of the same process, it is possible that the dynamic signatures may be similar. The level of similarity may also vary between reporters placed at different locations along the process, due to differences in both the time required and the step efficiencies in carrying out intervening activation/inactivation, transcription and translation processes. Differences due to the intrinsic properties of the drugs, rates of cellular uptake, efflux and enzymatic transformation to an active form, where required, could also alter the dynamics of cell response. In Figure 2b, we see that the reporter *F*'s responses to Drugs 1, 2, 3 and 4 can be ordered by D2 > D1 > D3 > D4. Furthermore, the onset times of detectable population change also follow the same order. These can be seen if (1) Drug 2's inactivation of gene *A* leads directly to shutting down the production of reporter *F*; (2) Drug 1's inactivation of transcription factor gene *B* adds only a single additional transcription step to achieve down-regulation of gene *A* and (3) Drug 3's rate of inactivation of transcription factor gene *D* is lower than that of Drugs 1 and 2 on genes *A* and *B*, but the inactivation is very rapidly transmitted from gene *D* to gene *B* once achieved. These hypothetical situations and results illustrate the general approaches that could be used to evaluate how the key characteristics of population change relate to cellular drug response dynamics.

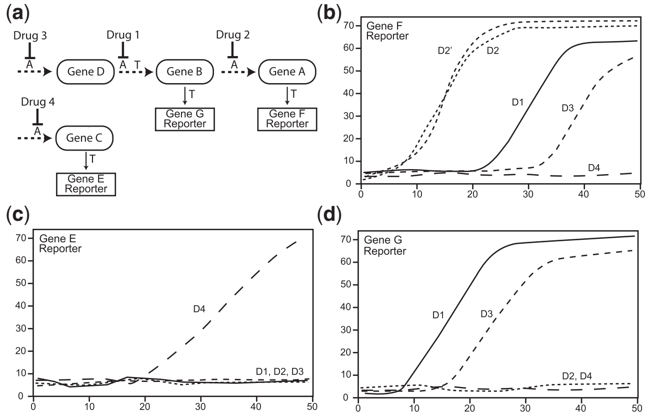


Fig. 2. A variety of possible population-change trajectories resulting from drug responses by the pathways diagrammed at the upper left side of the figure. Activation (A) and transcription (T) steps for which components are not shown in this graph occur in the dashed connections between the transcription factors. *X*-axis represents time in hours and *y*-axis represents percentage of population change

1.3 Rationale for drug response comparisons

The three characteristics of a population-change trajectory mentioned in the previous section correspond to three different phases. First is the dormancy period, where the population change is relatively constant and small. In this period, the cell lines are usually making the necessary conditions ready for the target GFP to show activity. For example, D1 in Figure 2b has a dormancy period of ~20 h. Next comes the responding period, where the right conditions have been prepared and reporters start to be affected by the drug intervention. In this period, we usually see an increase of the population change level. The rate of change depends on a number of factors, including drug dosage. Finally comes the stabilizing period, where the drug has reached its potential and population change is slowed down, eventually reaching a certain level. The final proportion of changed cells depends on the overall efficacy of the drug. Note that the three steps described here do not necessarily happen for every GFP reporter. For example, if the test drug is ineffective, then the treated cell line will remain in the dormancy period without any significant population change.

The most informative region is carried in the responding period described above, which is directly affected by a drug's sensitivity and efficacy. From a previous study, we know that fold change difference can be confidently detected (P-value = 0.05) when the population change is >7.25% (Hua *et al.*, 2012). Hence, we define the *core response* on a population-change trajectory to be the region >7.25%.

We believe that it is meaningful to study a drug's MOA only if it is able to induce a sufficient population change at some time point during the whole experiment.

The following descriptions summarize the key points that determine the mechanistic similarity between two drug responses:

1. The most informative region on a population-change trajectory is contained in the core region. The onset times for the core region to happen may vary from drug to drug; however, after the onset time, two population-change trajectories should proceed side by side if they are induced by two drugs with very similar MOAs. The longer time they continue, the better mechanistic similarity between the two. Here in, we define the *core containing alignment* of two population-change trajectories to be the longest contiguous alignment that contains at least one pair of points in the core region.

2. Noisy measurements or a slight stretch/compression in time can break an originally longer contiguous alignment into several smaller pieces. Hence, we should allow small gaps around the core containing alignment region described in Point 1. In other words, the similarity comparison for two drug responses should start from the core containing alignment and iteratively search its adjacent regions earlier or later in time, with a small gap allowed (e.g. 2 h). In the end, the different sections should be aggregated together to reflect overall similarity level.

3. Due to the 2-dimensional (2D) nature of the drug response data (population change and fold change), similarity requires the responses to be close on both dimensions. However, in reality, we often observe that when population change is small, the variation of fold change is quite large. Therefore, we require 2D similarity when population change is >7.25%; otherwise, we only require similarity on the population-change dimension.

To capture the similarities between reporter responses, we need an algorithm that is robust to noise, time delays and is able to find all the contiguous parts of signals centered about the core mechanism. Directly comparing points on the two signals is unrealistic, as it cannot handle delays and speed variations on the time axis, as shown in Figure 3a. Another popular approach, dynamic time warping (DTW) (Aach and Church, 2001; Sakoe and Chiba, 1978), is based on the concept that the similarity between two time series should be computed by locally deforming the time axis in order to minimize the cumulative difference between the aligned points. There are several disadvantages for the DTW-type algorithms. First, for global DTW, all the points on one signal must be mapped to points on the other signal. Thus, outliers cannot be skipped and they can severely distort the alignment. Even though efforts have been made to relax the global alignment constraint, e.g. the open-end DTW algorithm (Tormene *et al.*, 2009), where the head or the tail sections can be left unaligned, it is still difficult to avoid the middle portion outliers in the signal. Second, DTW algorithms tend to have many-to-one mappings for the alignment. Therefore, when the two signals are different in amplitude, it is often the case that a large portion of one signal will be mapped to a single point on the other signal to minimize the overall cumulative distance between the two. A global DTW alignment is shown in Figure 3b. Many superfluous and spurious matches are seen at the ending sections, making the alignment very counterintuitive.

An alternative solution is shown in Figure 3c, where the two signals are aligned based on the concept of *longest common substring* (LCSS), which belongs to the class of edit distance problems (Bergroth *et al.*, 2000; Hirschberg, 1977). LCSS finds the longest string that is a substring of two strings. In this article, we extend this approach to real-valued signals in a recursive fashion and call it recursive RLCSS (RLCSS). The benefit of RLCSS is that the aligned signals are contiguous and only one-to-one mapping is allowed, which satisfies our assumptions of biological similarity. Furthermore, small gaps (≤2 h) are allowed between different sections to account for noisy measurements.

2 SYSTEMS AND METHODS

2.1 LCSS on time series

The original LCSS model refers to a 1D sequence with discrete values, i.e. strings. For example, the sequences ABAB and BABB have their LCSS to be BAB. Our data are 2D and real-valued. The first dimension is population change and the second is fold change. These reflect two important aspects of the same cell population over time. Therefore, it is natural to consider both dimensions simultaneously when defining similarities.

Formally, let $A=((a_{x,1},\ a_{y,1}),\ \dots,\ (a_{x,m},\ a_{y,m}))$ and $B=((b_{x,1},\ b_{y,1}),\ \dots,\ (b_{x,n},\ b_{y,n}))$ be two drug responses, where x is the dimension for population change, y is the dimension for fold change, m is the length of A and n is the length of B. Let $A[1,\dots,i]=((a_{x,1},a_{y,1}),\dots,(a_{x,i},a_{y,i}))$.

Given an integer δ, a real value $k\in[0,1]$ and a pair of non-negative real values $\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2)$, we define the length of the LCSS between the 2D time series A and B to be the largest element in matrix $R_{\delta,\boldsymbol{\varepsilon}}$, where the element $R^{i,j}_{\delta,\boldsymbol{\varepsilon}}$ is defined by

$$R^{i,j}_{\delta,\boldsymbol{\varepsilon}}=\begin{cases} 0 & \text{if } i=0 \text{ or } j=0\\ 1+R^{i-1,j-1}_{\delta,\boldsymbol{\varepsilon}} & \text{if } |a_{x,i}-b_{x,j}|\le\varepsilon_1,\\ & |a_{y,i}-b_{y,j}|\le\varepsilon_2,\\ & |i-j|\le\delta,\ |a_{x,i}|\ge k,\ |b_{x,j}|\ge k,\\ & \text{or if } |a_{x,i}-b_{x,j}|\le\varepsilon_1,\ |i-j|\le\delta,\\ & \text{and } |a_{x,i}|\le k \text{ or } |b_{x,j}|\le k,\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

where the constant δ controls the flexibility of matching in time, the constant vector $\boldsymbol{\varepsilon}$ controls the matching threshold and k determines the population change threshold which fold change constraint will be applied (i.e. the population threshold for the core mechanism). Throughout the article, we will set $k=0.0725$, because fold change difference can be confidently detected (P-value = 0.05) when population change has reached 7.25% (Hua *et al.*, 2012). Intrinsic to equation (1) is that $R^{i,j}_{\delta,\boldsymbol{\varepsilon}}$ depends only on the previous diagonal element and the current element-wise distance. Hence, $R^{m,n}_{\delta,\boldsymbol{\varepsilon}}$ can be efficiently found by filling the table starting from $R^{0,0}_{\delta,\boldsymbol{\varepsilon}}$. After the table has been filled, the actual common substring can be found by going back diagonal-wise from the largest entry in the table until a 0 entry is reached (the trace-back path shown in Table 1). Intuitively, equation (1) says that two drug responses are similar either: (1) if population change is large enough, data along each dimension must be similar or (2) if population change is not large enough, only population change has to be similar, because the measurement on fold change is no longer reliable. The requirement is made to be consistent with Condition 3 described in Section 1.3. For illustrative purposes, a numerical example is shown in Table 1.

By definition, the substring must be contiguous. This differs from the concept of *longest common subsequence*, where the subsequence is not necessarily contiguous (for example, the longest common subsequence between ABACD and BABD is BAD or ABD). Furthermore, different

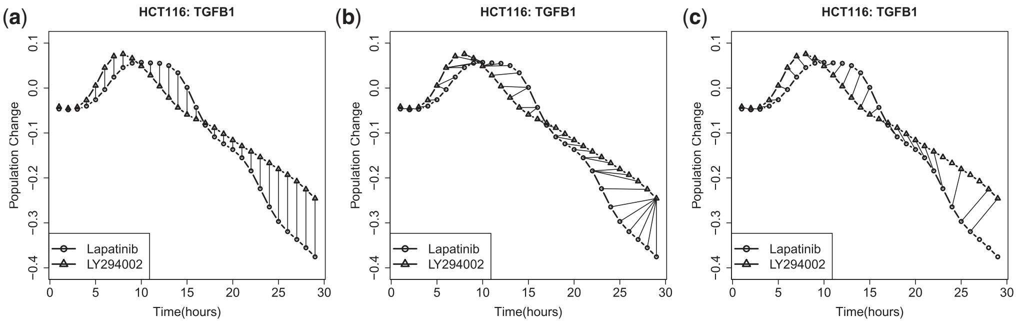


Fig. 3. (**a**) Direct alignment. It completely ignores the variation in time axis. (**b**) Global DTW alignment. Many superfluous and spurious matches are seen at the ending sections. (**c**) RLCSS algorithm. Only one-to-one mapping is allowed and small gaps are allowed to account for noisy measurement

Table 1. The DP table for finding the LCSS of two 1-D sequences [0.2,0.1,0.1,0.2] and [0.1,0.2,0.1,0.1], with the parameters set to be: $\delta=4, k=0, \boldsymbol{\varepsilon}=\varepsilon_1=0$

		0.1	0.2	0.1	0.1
	0	0	0	0	0
0.2	0	0	**1**	0	0
0.1	0	1	0	**2**	1
0.1	0	1	0	1	**3**
0.2	0	0	2	0	0

The trace-back path is highlighted in bold face and it contains time indices: $\{(1,2),(2,3),(3,4)\}$.

choices of δ and $\boldsymbol{\varepsilon}$ will lead to different alignment results. A small δ will restrict the alignment points to be close in time. In fact, $\delta=0$ degenerates to the calculation of direct matching (Fig. 3a). For $\boldsymbol{\varepsilon}$, a very small threshold will lead to almost no alignment between the two signals, whereas a very loose threshold will lead to every pair of points being aligned as similar. Therefore, an appropriate choice of δ and $\boldsymbol{\varepsilon}$ is important for the application of LCSS alignments. A good choice of the two values depends on the application. Thus, we will have a detailed discussion in Section 3.

2.2 RLCSS algorithm

The LCSS algorithm described in the previous section is not yet sufficient for drug response comparisons. First, there is no guarantee that the LCSS will intersect with the region where sufficient population change has been reached and, therefore, cannot be called the core mechanism described in Section 1.3. Second, even if the core mechanism is found, small gaps should be allowed around it to compensate for noisy measurements. For these two reasons, we define an algorithm that will use the LCSS concept recursively to identify the core containing alignment as well as its surrounding pieces.

1. For a pair of 2D sequences A and B, fill the dynamic programming (DP) table as described in equation (1). Find the longest trace-back path (ending with the largest element in the DP table). If its corresponding matched points contain any member that has sufficient population change on both sequences ($\geq 7.25\%$), then record the trace-back path; otherwise, keep searching the second longest trace-back path in the DP table until it satisfies the population change threshold requirement. Denote that track-back path to be T, where (p,q) is the pair of starting time indices and (s,t) is the pair of ending indices. Stop and go to Step 2. If no such trace-back path exists, exit the program and return $T=$NULL.

2. For the head section sequences $A[1,\ldots,p-1]$ and $B[1,\ldots,q-1]$ of A and B, respectively, fill the DP table to find the LCSS path of the truncated head section sequences. If the ending indices of the LCSS trace-back path are within the time gap allowed from (p,q), then add the newly found trace-back path to the beginning of T; otherwise, continue to search for the second LCSS until it meets the time gap constraint. Update (p,q) so that it represents the starting indices of the newly formed T and go to Step 3. If no such trace-back path is found, stop and continue to Step 4.

3. Repeat Step 2 for the remaining head sections of A and B until no trace-back path satisfies the condition described in Step 2. Stop and continue to Step 4.

4. For the tail section sequences $A[s+1,\ldots,m]$ and $B[t+1,\ldots,n]$ of A and B, respectively, fill the DP table to find the LCSS path of the truncated tail section sequences. If the starting indices of the LCSS trace-back path is within the time gap allowed from (s,t), then add the newly found trace-back path to the end of T; otherwise, continue to search for the second LCSS until it meets the time gap constraint. Update (s,t) so that it represents the ending indices of the newly formed T, and go to Step 5. If no such trace-back path is found, stop and exit program.

5. Repeat Step 4 for the remaining tail sections of A and B until no trace-back path satisfies the condition described in Step 4. Stop and exit program.

Note that by aligning the head sections and tail sections separately, it is guaranteed that the time order will not be destroyed. Figure 4 shows a graphic illustration of the RLCSS algorithm.

The similarity $S_{\delta,\boldsymbol{\varepsilon}}(A,B)$, expressed in terms of the RLCSS similarity between the time series A and B, is given by

$$S_{\delta,\boldsymbol{\varepsilon}}(A,B)=\frac{|T|}{min(n,m)}, \tag{2}$$

where T is found by the RLCSS algorithm and $|T|$ is the cardinality of T. Note that $S_{\delta,\boldsymbol{\varepsilon}}(A,B)$ is always between 0 and 1—the larger the value, the greater the similarity.

3 RESULTS AND DISCUSSION

In this section, we apply the RLCSS algorithm to drug response data. To test its performance, we need to know *a priori* the MOAs of the testing drugs and see whether the alignment results agree with our prior knowledge. For instance, if we know that Drugs X and Y have very similar effects on some cell line (due to similar MOAs), can the proposed RLCSS algorithm capture their similarities and claim they are similar? Conversely, if Drugs X and Z are very different in their MOAs, is the RLCSS algorithm also able to claim that they are dissimilar?

To test the performance of the proposed RLCSS algorithm, we consider the results of a study in which the detailed MOA of each

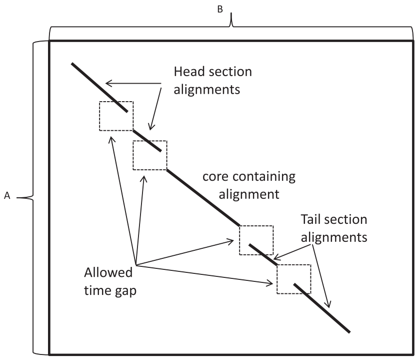


Fig. 4. Illustration of the RLCSS algorithm. The DP table is represented by the big solid black box. The algorithm starts by finding the core containing alignment, and subsequently recursively finds the head section alignments and tail section alignments around it with small time gaps allowed

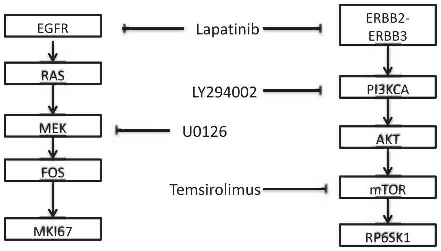


Fig. 5. A simplified pathway diagram showing drug interventions

drug has been performed (Hua *et al.*, 2012). In this study, five drugs (Lapatinib, LY294002, Temsirolimus, U0126 and AG1024) are tested against the colorectal carcinoma cell line HCT116 (ATCC Number: CCL-247). Referring to Figure 5 and ranking drug responsiveness across the drugs for HCT116, one observes responses over most reporters but with decreasing percentages of cells shifted for Lapatinib (EGFR/ERBB2), LY29004 (PI3K) and Temsirolimus (mTOR). This similarity of action with decreasing efficacy falls directly along a survival signaling pathway headed by an activated receptor heterodimer, ERBB2/ERBB3, and then proceeds along the canonical PI3K/AKT/mTOR pathway. The remaining drugs' inhibitory powers would be ranked U0126 (MEK1/2) and AG1024 (IGF1R)—AG1024 not shown in the figure because it acts on an unrelated kinase. All of these second tier drugs deliver very low reductions of transcription of MKI67, the current 'gold standard' (Gerdes, 1990; Scholzen and Gerdes, 2000) in tumor pathology for determining the proliferative state of a tumor. The results show that Lapatinib, LY294002 and Temsirolimus have similar MOAs in the sense that they all target the same survival pathways. On the other hand, U0126 and AG1024 have very dissimilar MOAs compared with the three previously mentioned drugs.

We design several sets of experiments to test whether the proposed RLCSS algorithm reaches the same conclusions as just described. First, to get a sense of the variation presented in the drug response data, we test RLCSS on a set of technical replicates (TRs), with the idea that TRs should exhibit high degrees of similarity among each other. Furthermore, by studying the TRs, we can find a proper range for the two key parameters in the RLCSS algorithm. Second, we test RLCSS on Lapatinib, LY294002 and Temsirolimus, since they are related in their MOAs, and the degree of similarity should be high, but not as high as with the TRs. Last, we test RLCSS on Lapatinib, U0126 and AG1024, because Lapatinib is very different from the last two in their MOAs. We set the parameter $k = 0.0725$ in equation (1) and the time gap to be 2 h for all experiments.

3.1 RLCSS performance on real data

3.1.1 ***TRs to determine*** δ ***and*** $\boldsymbol{\varepsilon}$ For RLCSS to work in practice, a proper set of values must be determined for δ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)$. δ controls the maximum time delay allowed for the matching. In practice, the experiment lasts around 30–50 h and we know that transcription is a relatively slow process. It usually takes about 8–12 h for a reporter to show some activity after the initial treatment. Therefore, in our application, we set δ to be between 8 and 12 h, which should be enough to compensate for the time delays of different drugs. Moreover, we observe that RLCSS is usually insensitive to δ variations in the sense that changing δ in that range does not change the alignment results significantly (Table 2). $\boldsymbol{\varepsilon}$ determines the threshold for similarity. The idea for determining $\boldsymbol{\varepsilon}$ is to set it to be large enough to compensate for the discrepancies in TRs. Herein, we set $\boldsymbol{\varepsilon}$ to be the value so that the worst case TRs similarity is at least 75% (Fig. 6c and d). In practice, we have found that it is enough to account for the biological variations with ε_1 close to 0.09 and ε_2 close to 0.8, as we show in Table 2.

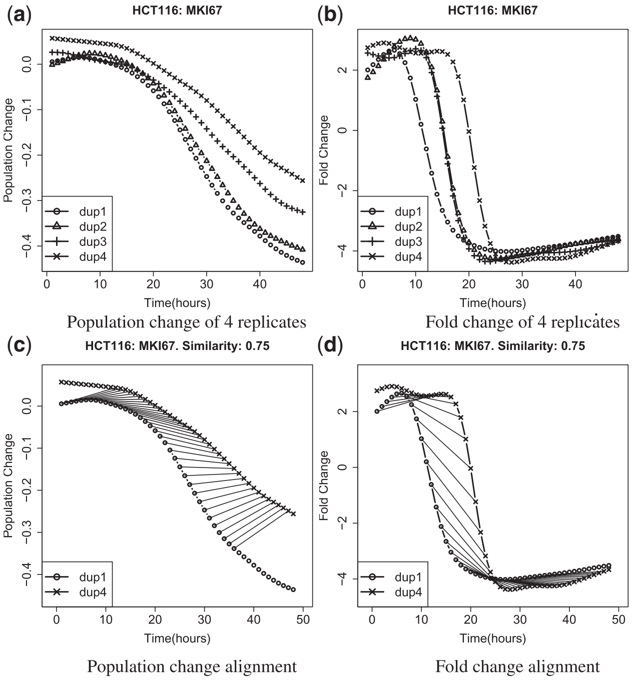
Figure 6a and b shows a set of 4 factor on the reporter MKI67 on the cell line HCT116 after treatment with Lapatinib, and Figure 6c and d shows the RLCSS alignment for the worst case TR similarity. The pairwise alignment results are summarized in Table 2. The alignment results are not affected at all by the different choices of δ ranging from 10 h to 12 h for all the pairwise alignments except dup1 and 4, indicating that the RLCSS algorithm is very robust to δ variations. As seen in Figure 6d, the RLCSS algorithm successfully identifies the time delays between the two replicates.

3.1.2 ***Lapatinib, LY294002 and Temsirolimus*** We design the second set of experiments to show the utility of RLCSS on three drugs with similar MOAs. As described in Section 1.3, it is meaningless to compare drug responses on reporters whose responses have not changed enough during the whole experimental span. Therefore, we select the reporters whose responses show at least 7.25% population change for at least two out of the three drugs. The similarity comparison table of the three is summarized in Table 3. The three drugs show considerable amount of similarity with each other, especially on ERBB3 and MKI67, the two key reporters that reflect the MOA of drugs on cell line HCT116 (Hua *et al.*, 2012). We also observe that Lapatinib is closer to LY294002 than it is to Temsirolimus. The result is also consistent with our prior knowledge that LY294002 is closer to Lapatinib than Temsirolimus in their actual positions of attack (Fig. 5).

3.1.3 ***Lapatinib, U0126 and AG1024*** The third set of experiments is intended to test whether the RLCSS algorithm is able to detect mechanistic differences between drugs. As we know from earlier discussion, Lapatinib is very different from U0126 and AG1024 in their MOAs. The similarity results of the three drugs are summarized in Table 4. Ranking by the closeness to Lapatinib, the order of similarity is U0126 and AG1024. Figure 7a and b shows the five drugs' responses on reporter ERBB3. Note that

Table 2. Pairwise similarity between TRs, with different δ

	Dup1, 2	Dup1, 3	Dup1, 4	Dup2, 3	Dup2, 4	Dup3, 4	δ	ε
Similarity	1	0.917	0.667	0.958	0.75	1	8	(0.09,0.8)
Similarity	1	0.917	0.688	0.958	0.771	1	9	(0.09,0.8)
Similarity	1	0.917	0.708	0.958	0.792	1	10	(0.09,0.8)
Similarity	1	0.917	0.75	0.958	0.792	1	11	(0.09,0.8)
Similarity	1	0.917	0.75	0.958	0.792	1	12	(0.09,0.8)



Population change of 4 replicates

Fold change of 4 replicates

Population change alignment

Fold change alignment

Fig. 6. TRs of Lapatinib treatment on cell line HCT116. $\boldsymbol{\varepsilon}$ is set to be the value so that for the worst case TRs, similarity is at least 75%

Table 3. Pairwise similarity between three drugs with similar MOAs, with $\delta = 11$ and $\boldsymbol{\varepsilon} = (0.09, 0.8)$

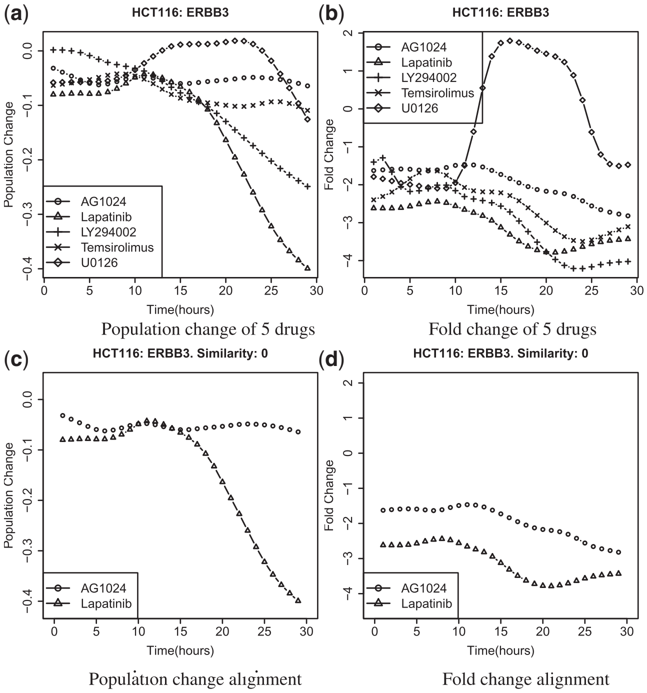
	Lapatinib LY294002	Lapatinib Temsirolimus	LY294002 Temsirolimus
TGFB1	0.862	0	0.138
ERBB3	0.862	0.724	0.793
EGR1	0.611	0.167	1
MKI67	0.621	0.69	0.897
FOS	0.677	0.583	0.25

the RLCSS algorithm can filtered out 'uninterestingly' similarities (Fig. 7c and d).

3.1.4 ***Detect apoptosis*** It is possible that the reporter responses are different in the beginning, but later in time behave similarly due to a common process, e.g. apoptosis. This is because once a cell has determined to go through apoptosis, all the reporters will have a significant drop in their activity level and eventually die out. UNBS1415 is a drug that induces apoptosis on the lung carcinoma cell line A549 (ATCC Number: CCL-185). In Figure 8a and b, we can see that the initial responses are quite different for different reporters; however, later in time, all responses seem to converge to the same behavior after 25–30 h. In Figures 8c and d, we see that

Table 4. Pairwise similarity between three drugs with distinct MOAs, with $\delta = 11$ and $\boldsymbol{\varepsilon} = (0.09, 0.8)$

	Lapatinib U0126	Lapatinib AG1024	U0126 AG1024
TGFB1	0.793	0	0
ERBB3	0	0	0
EGR1	0.306	0	0
MKI67	0	0	0
FOS	0.583	0.611	0.861



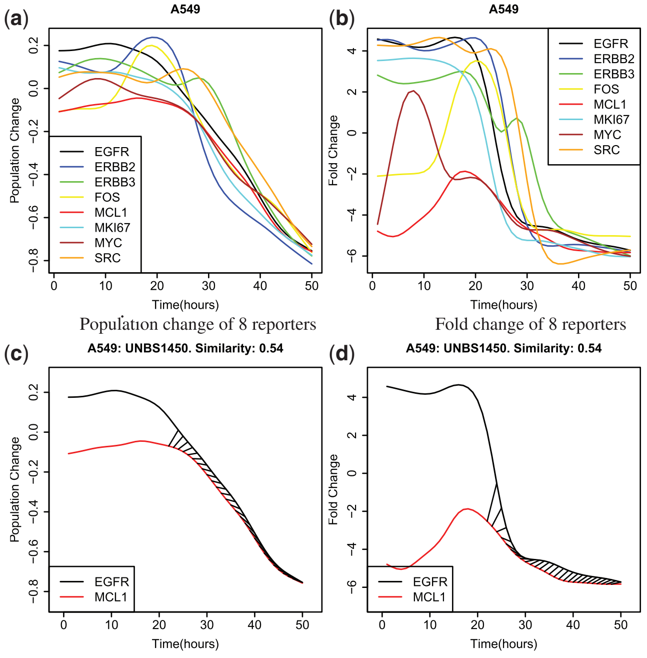
Population change of 5 drugs

Fold change of 5 drugs

Population change alignment

Fold change alignment

Fig. 7. Responses of ERBB3 to five different drugs. The RLCSS algorithm has the advantage to filter out 'uninteresting' similarities, when no core mechanism can be formed



Population change of 8 reporters

Fold change of 8 reporters

Population change alignment

Fold change alignment

Fig. 8. Responses of eight reporters to the drug UNBS1450 on cell line A549. The RLCSS algorithm successfully identified the similarity later in time. The parameters are $\delta=11$ and $\boldsymbol{\varepsilon}=(0.09,0.8)$

RLCSS algorithm successfully identifies the similarity later in time, which is exactly what we expect.

3.2 RLCSS performance on synthetic data

We conduct two simulation studies to evaluate the performance of the RLCSS algorithm. As mentioned in Section 1.3, the population-change trajectory usually has three different phases: the dormancy phase the active responding phase and the stabilizing phase. The shape of the curve can be represented by the family of logistic curves, which are often used to model population dynamics in a resource limiting environment. Hence, to model population change, we use the logistic function $p(t)=-1/1+Ke^{-r(t-t_1)}$, where t is time, $p(t)$ is the population change at time t, K and r are positive numbers controlling the shape of the curve and t_1 is an arbitrary time to shift the entire logistic curve along the time axis. In a similar vein, the fold change can also be modeled as logistic curves with $q(t)=-1/1+K'e^{-r'(t-t_1')}$, where we set $K'=K$, $r'>r$ and $\{t_1'|p(t_1')=-0.0725\}$. $r'>r$ guarantees the fold-change dimension to have a sharper transition than the population-change dimension, as is often seen in our experiments (Fig. 6), and by requiring $p(t_1')=-0.0725$, we assume that the fold change starts to decrease quickly when enough population change has accumulated. Note that in ideal situations, fold change should have a sudden drop as soon as one single cell has responded to the drug; however, in practical situations, such change cannot be detected reliably until a considerable amount of population change has happened (Hua *et al.*, 2012). Throughout the simulations, we set $K'=K=100$, $r=0.3$, $r'=1$ and $t=0,1,\ldots,40$. The choice reflects closely real experiment data.

Given two drug responses, $A=(p(t,t_1=0),q(t))$ and $B=(p(t,t_1=r),q(t))$, we first show the effect of the parameter δ on different delays $r=1,2,\ldots,20$. Note that $q(t)$ in A and B are different and are dependent on the respective $p(t)$. As shown in Figure 9, we see a big decrease in similarity between A and B when the delay r is larger than $\delta+1$. This is expected because δ in the RLCSS algorithm controls the flexibility of matching in time, and if the delays are too large compared with δ, the RLCSS will not capture such similarities. In our drug experiments, we have found that $\delta=11$ is sufficiently large to handle all the possible delays. We also note that for a given delay r, the different δ do not change similarity score when $\delta\geq r$. For example, when $r=5$, changing δ from 8 to 11 does not affect the similarity score at all. This indicates that the RLCSS algorithm is robust to δ and the conclusion is consistent with the results shown in Table 2.

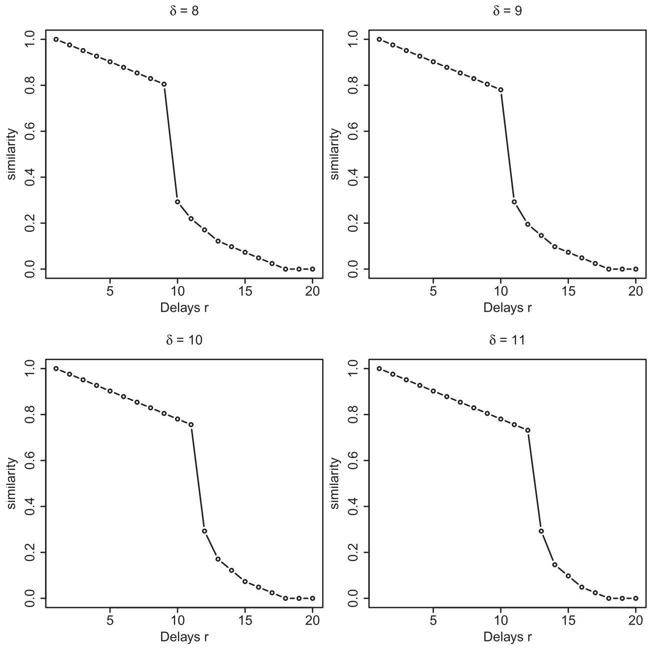


Fig. 9. Effect of δ on different delays r, with $\boldsymbol{\varepsilon}=(0.09,0.8)$

In the second simulation, we show the performance of the RLCSS algorithm with respect to noise. Let $A=(p(t,t_1=0),q(t))+N(0,\sigma)$ and $B=(p(t,t_1=5),q(t))+N(0,\sigma)$, where $N(0,\sigma)$ is Gaussian white noise with 0 mean and σ standard deviation. As shown in Figure 10, the performance of the RLCSS algorithm deteriorates with increased noise level. However, at the noise level $\sigma=0.0725$ encountered in real experiments, we see that 97% of cases have a similarity of 0.75 or more (Fig. 11). This indicates that the false-negative rate is only 3% for the TR experiment.

3.3 Concluding remarks

The proposed RLCSS algorithm aligns reporter response dynamics in order to facilitate identification of mechanistic similarities in drug responses. As opposed to 1D discrete alignment algorithms, it operates on 2D real-valued data corresponding to population change and fold change. It achieves three desirable aims: robustness to noise, robustness to time delays and the ability to find contiguous parts of signals centered abut the core mechanism. Its performance has been tested on TRs, drugs with similar MOAs, drugs with

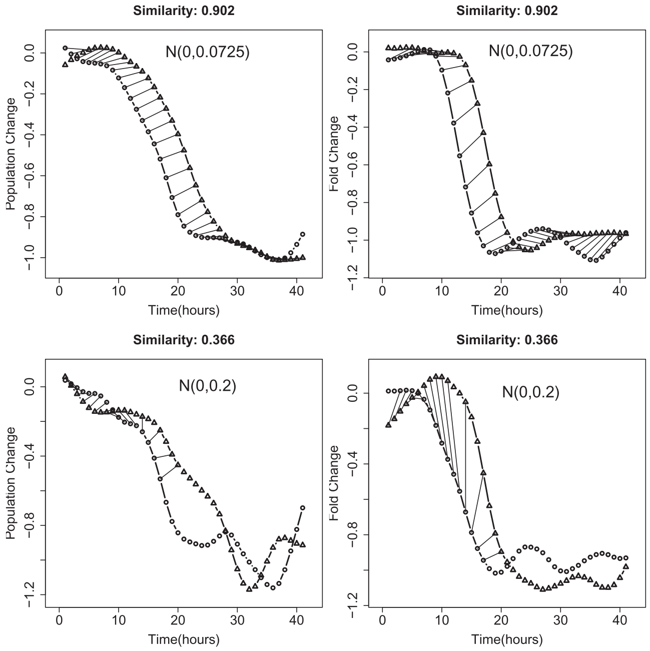


Fig. 10. Performance of RLCSS with respect to different noise level, with $\delta=11$ and $\boldsymbol{\varepsilon}=(0.09,0.8)$

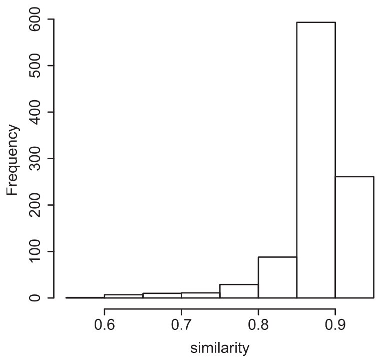


Fig. 11. Similarity histogram for noise level $N(0,0.0725)$, with $\delta=11$ and $\boldsymbol{\varepsilon}=(0.09,0.8)$

different MOAs, an apoptotic drug and synthetic data. In all cases, it performed successfully.

Funding: This publication is based in part on work supported by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

Conflict of Interest: none declared.

REFERENCES

Aach,J. and Church,G.M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

Bergroth,L. *et al.* (2000) A survey of longest common subsequence algorithms. In *Seventh International Symposium on String Processing and Information Retrieval*, IEEE Computer Society, Washington, DC, USA, pp. 39–48.

Chalfie,M. *et al.* (1994) Green fluorescent protein as a marker for gene expression. *Science*, **263**, 802–805.

Chen,S.J. *et al.* (1991) Rearrangements in the second intron of the RARA gene are present in a large majority of patients with acute promyelocytic leukemia and are used as molecular marker for retinoic acid-induced leukemic cell differentiation. *Blood*, **78**, 2696–2701.

Dougherty,E.R. and Lotufo,R.A. (2003) *Hands-on Morphological Image Processing*. SPIE Optical Engineering Press, Bellingham, Wash.

Gerdes,J. (1990) Ki-67 and other proliferation markers useful for immunohistological diagnostic and prognostic evaluations in human malignancies. *Semin. Cancer Biol.*, **1**, 199–206.

Hirschberg,D.S. (1977) Algorithms for the longest common subsequence problem. *J. ACM*, **24**, 664–675.

Hu,G and Agarwal,P. (2009) Human disease-drug network based on genomic expression profiles. *PLaS One*. **4**, e6536.

Hua,J. *et al.* (2012) Tracking transcriptional activities with high-content epifluorescent imaging. *J. Biomed. Opt.*, **17**, 046008 (May 01, 2012).

Hunt-Newbury,R. *et al.* (2007) High-throughput *in vivo* analysis of gene expression in Caenorhabditis elegans. *PLoS Biol.*, **5**, e237.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. USA*, **107**, 14621–14626.

Lamb,J. *et al.* (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Sakoe,H. and Chiba,S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.*, **26**, 43–49.

Scherf,U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.

Scholzen,T. and Gerdes,J. (2000) The Ki-67 protein: from the known and the unknown. *J. Cell. Physiol.*, **182**, 311–322.

Sergina,N.V. *et al.* (2007) Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature*, **445**, 437–441.

Sirota,M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, **3**, 96ra77.

Tormene,P. *et al.* (2009) Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artif. Intell. Med.*, **45**, 11–34.

Yusuf,R.Z. *et al.* (2003) Paclitaxel resistance: molecular mechanisms and pharmacologic manipulation. *Curr. Cancer. Drug. Targets*, **3**, 1–19.