A coherent framework for multiresolution analysis of biological networks with “memory”: Ras pathway, cell cycle, and immune system

Paolo Emilio Barbano*†‡, Marina Spivak‡§, Jiawu Feng†, Marco Antoniotti§, and Bud Mishra†¶

*Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012; †Department of Cell Biology, New York University, 550 First Avenue, New York, NY 10016; ‡Department of Mathematics, Yale University, 10 Hillhouse Avenue, New Haven, CT 06520; and §Department of Genetics, Yale University School of Medicine, New Haven, CT 06520

Communicated by Ronald R. Coifman, Yale University, New Haven, CT, March 9, 2005 (received for review September 24, 2003)

Various biological processes exhibit characteristics that vary dramatically in response to different input conditions or changes in the history of the process itself. One of the examples studied here, the Ras–PKC–mitogen-activated protein kinase (MAPK) bistable pathway, follows two distinct dynamics (modes) depending on duration and strength of EGF stimulus. Similar examples are found in the behavior of the cell cycle and the immune system. A classification methodology, based on time-frequency analysis, was developed and tested on these systems to understand global behavior of biological processes. Contrary to most traditionally used statistical and spectral methods, our approach captures complex functional relations between parts of the systems in a simple way. The resulting algorithms are capable of analyzing and classifying sets of time-series data obtained from in vivo or in vitro experiments, or in silico simulation of biological processes. The method was found to be considerably stable under stochastic noise perturbation and, therefore, suitable for the analysis of real experimental data.

Ras–PKC–mitogen-activated protein kinase (MAPK) pathway | systems biology | time-frequency analysis

Systems biology has focused on the task of understanding large-scale complex biochemical processes by using mathematical models, systems theoretic analysis, and data-mining algorithms. The interest in the subject is motivated by the availability of large amounts of genomic and proteomic information that points to the existence of complex networks of interactions in the cell. The structure of these interactions can be represented by means of the connectivity in a graph of the biological “parts” of the network. Major emphasis has been placed on discovering “functional motifs” by data-mining these graphs, thus implicitly assuming that the function of a pathway is suggested by its topology. However, many elegant biological examples have shown recently that the topology of the pathway only guides the function, because it may support substantially different responses to varying stimuli as well as conditions in the cell.

Many such biological systems, when exposed to transient stimuli varying in strength or other characteristics, can display substantially different responses, persisting long after the signal is withdrawn or, in extreme cases, being irreversible. The system is intuitively interpreted as committing itself to different modes of operation and displaying different behavior in these modes. With the recent interest in global analysis of biological processes, these systems have begun to be intensively studied and subjected to high-throughput experiments. Nevertheless, these studies seem to have remained focused only on the final outcome of the evolution of the system: namely, what the states of the system are and which state the system reaches at different times (1, 2).

In this article, we propose a different kind of analysis; we suggest an approach capable of discerning the dynamic characteristics of the time-course behavior of the response of the system to different stimuli. By dynamic characteristics, we mean that the components of the system, as they respond to stimuli, may follow different trajectories before they reach a steady state. If the topology of the system supports differential behavior, the evolution of the components of the system during these different responses can be distinguished by time-frequency analysis techniques. Ultimately, we argue that, given a time-course data under different experimental conditions, it is possible to use time-frequency techniques to gain intuition about topology of the system.

Biological Time Series and Discriminant Space (D-Space)
The algorithm proposed here is designed to analyze data obtained from time-course experiments (e.g., transcriptomic or proteomic data) or in silico simulation of biological processes. The input to the algorithm consists of trajectories for the dynamic evolution of the abundance of various molecules in a biological system generated at different experimental conditions. The goal of the analysis is to determine whether variations in the experimental conditions (e.g., initial conditions or duration of stimuli) cause the system to evolve globally in a substantially different manner. We can then identify different modes of operation in the system and establish a correspondence between the typical experimental conditions and these modes of dynamic behavior. For example, our technique is able to detect the differences in the evolution toward the two stable states of Ras–PKC–mitogen-activated protein kinase (MAPK) bistable pathway activated by EGF stimuli of various strengths. However, the differences in dynamic behavior that we can detect are not at all confined to multistable systems.

The simple mathematical observation that we make is that it is possible to choose a small number of vectors in an orthonormal basis so that all the trajectories of the system under consideration are effectively described only by the coefficients with respect to those vectors. In mathematical terms, we study the characteristics of the set of trajectories of a complex biological system by projecting them onto a suitable, low-dimensional vector space. Because any trajectory can be projected onto this “coefficient space” (more formally, the D-Space), it is then possible to project a large number of randomly sampled trajectories into points in the D-Space and identify the different modes of evolution of the system by inspecting the clusters that these projected points form in D-Space. We then identify the modes of the biological system by studying the geometric properties of these projected points. A more formal explanation of such techniques is given in Supporting Appendix, which is published on the PNAS web site (see also refs. 3–5).

Methods
A biological process is described by the set of trajectories of concentration, e.g., time series functions: \( F = \{ f(t), 0 \leq t \leq T \} \).

Abbreviations: D-Space, discriminant space; MAPK, mitogen-activated protein kinase; Cycl B, cyclin B; CDK, cyclin-dependent kinase; APC, antigen-presenting cell; SK, start kinase; CKI, casein kinase I.

†P.E.B. and M.S. contributed equally to this work.
‡To whom correspondence should be addressed. E-mail: mishra@nyu.edu.
© 2005 by The National Academy of Sciences of the USA
in the Hilbert space $L^2([0, T])$. Here, we approximate them through a projection of $\mathcal{F}$ onto $\mathbb{R}^M$, a Euclidean space. Let $\mathcal{B}$ be a finite orthonormal basis $\mathcal{B} = \{g_i(t), 0 \leq j \leq \infty\}$. $\mathcal{F} = \{f_j(t), g_j(t) = \sum_{i=1}^M c_i g_i(t)\}$ where $c_i$ are coefficients. With a suitably chosen subset of $\mathcal{B}$, the Euclidean distance between two functions determines similarity between their behaviors. The most typical behaviors (or modes) of the system are determined by functions whose projections lie all within a Euclidean sphere. Formally, given a family of trajectories $\mathcal{T}_1, \ldots, \mathcal{T}_n$ associated with different states of a biological system, we determine a basis $\mathcal{B}$, a finite subset $\mathcal{B} \subset \mathcal{B}$ and points $(x_1, \ldots, x_n) \in \mathbb{R}^M$ with the following properties. (i) For any two functions $f_i = \sum_{j=1}^M c_{ij} g_j$ and $f_j = \sum_{k=1}^M c_{jk} g_k$ in $\mathcal{B}$, it holds that: $f_i, f_j \in \mathcal{B}(x_i, \delta)$, where $B(x_i, \delta)$ denotes the ball of radius $\delta$ centered at $x_i$. (ii) For all distinct $i$ and $j$, $\|x_i - x_j\| \geq 3\delta$. In Supporting Appendix, we describe an adaptation of a multiresolution algorithm (3) for our basis selection and projection. We use the LOCAL DISCRIMINANT BASIS (LDB) algorithm (3), an efficient method to select the “Best Basis” in a wavelet-packet scheme (5). It works well for biological processes, even when projected onto the plane ($M = 2$).

**Biological Examples**

The three systems that we chose to examine (Ras–PKC–MAPK, cell cycle, and immune system) display switch-like behavior, either in response to graded transient stimuli or as a consequence of the evolution process of the system. We use the term “switch-like” in a loose sense to describe both reversible and irreversible behavior. In the reversible case, a steady state is reached and persists long after the stimulus is withdrawn before the system relaxes to its original state. When we say that a system is “irreversible,” we mean that, after a sufficient stimulus, the system commits to a new mode of operation from which it never returns to the previous state.

To demonstrate that the different characteristics of the three networks trajectories are induced by multimodal responses of a particular topology, we selected three data sets to analyze, all of which were obtained by simulation. Also, we tested our method on time-course yeast microarray data from Spellman et al. (6).

The first system that we studied is the MAPK enzymatic cascade activated by EGF through two interconnected pathways: the PLC–PKC and the Ras–Raf–MAPK pathways (Fig. 1A). Depending on the duration and concentration of the EGF stimulation, the system follows different evolution dynamics after the EGF stimulus is withdrawn. The simulations of this system were carried out by Bhalla and Iyengar (1).

The second example uses Tyson and Novak’s (7) simulated model of the yeast cell cycle and examines their hypothesis that the model involves only two stable states: G1 and S–G2–M, separated by two transitions (Start and Finish) (Fig. 1B). After each transition, the decision to commit to one or the other state is irreversible. We also studied this system by using the experimental mRNA time-course data. We applied time-frequency analysis to the yeast microarray data presented in ref. 6. This analysis shows that, among the three different cell-cycle-synchronization methods (namely, the α-factor arrest, cdc-15 temperature-sensitive mutant, and elutriation), the first two methods produce more regular oscillatory patterns than the last method.

The last example is based on a cellular-automata model of the immune system and a detailed examination of its adaptive dynamics in response to primary and secondary infections (Fig. 1C). In this case, the dynamics is again irreversible, because the primary response to an antigen brings that system into a different mode, which determines the subsequent secondary response to the same antigen.

In the three cases that we describe, knowledge of the topology of the systems provides insufficient information about their behavior, because the topology describes only qualitative relations among the components of the system and the direction of the evolution.
Furthermore, careful analysis at this qualitative level is even more difficult because experiments may often involve noisy and incomplete experimental samples, or generate data sets that are too large and/or scattered to allow their study by mere visual inspection. Last, it is often not known at all whether the system is multimodal and, in this case, it is not evident which component of the system needs to be examined in order to observe the differential behavior. For example, the main result of Bhalla and Iyengar (1), whose simulation we also use for our analysis, is that the levels of active forms of PKC and MAPK can reach different steady states under EGF stimuli of different strength (see Fig. 2). However, it is evident that the examination of other components of the system may not reveal this characteristic behavior. We emphasize that because our analysis uses the trajectories of all the members of the network (active and inactive forms), we are able to detect that the system as a whole operates multimodally under different conditions.

In the experimental set up, this technique may prove to be useful because the time course of only a limited number of the components can be measured, and very possibly, the member of the network with the most characteristic behavior is omitted. Moreover, the two first examples in our article focus mostly on the systems with multiple steady states given by complex topology; however, it is possible to detect completely different modes of evolution of the biological systems, unrelated to multiple steady states and not confined to specific topology features, but reflect certain extraneous properties. For example, such a deviation could be due to differences in the system’s response to variations in experimental design and other similar effects, as our microarray example and immune system simulation example demonstrate.

**Fig. 2.** The trajectories of MAPK and PKC change if EGF stimulus is provided. The MAPK (A) and PKC (B) time course for different values of EGF stimulus are shown. A 6,000-s EGF stimulus is provided at five levels (1, 2, 3, 5, and 7 nM). The two different modes for MAPK and PKC are shown.

More specifically, it was observed that the behavior of the components in the feedback loop depends on the amplitude and duration of the stimulation by EGF; the time evolution of the concentration of activated PKC and MAPK will follow different patterns. If the stimulation by EGF is below a threshold, the system will become deactivated soon after the signal is withdrawn; otherwise, if the stimulation is above this threshold, the active MAPK and PKC will persist even after the signal is withdrawn (1) (Fig. 2D). Bhalla and Iyengar suggest that these properties, allowing persistent enzymatic activation, raise the possibility of storing information by switching between the steady states and, in this way, displaying “learning behavior.”

The underlying mathematical model consists of a set of polynomial ordinary differential equations,

\[ [X_i] = \sum_{1 \leq j, k \leq n} a_{ijk} [X_j]^p [X_k]^q \quad 1 \leq i \leq n, \]

whose coefficients are derived from *in vitro* data. The simulations were implemented as in Bhalla and Iyengar (1) by using their simulation software GENESIS. The simulation was as follows: after letting the system equilibrate, we applied EGF stimulus for 6,000 sec and then let the system relax for 4,000 sec. We generated two sets of data corresponding to the response of the system to 2- and 5-nM stimuli by EGF and analyzed the time-course trajectories generated by all the components of the system. However, before our complete analysis, we eliminated trajectories that were too similar in the sets generated by 2- and 5-nM EGF stimuli. This step is necessary because we assume that identical trajectories cannot verify whether multimodal behavior is present. We observed that two sets of time-course trajectories of the components of the network produced by two levels of EGF (2 and 5 nM for 6,000 sec) possess distinct time-frequency characteristics. In other words, by means of LDB algorithm described in Supporting Appendix, we were able to isolate two vectors with respect to which the trajectories of the components of the system at “high” and “low” EGF stimuli could be classified. The clustering in the D-Space, for EGF at low and high levels, is shown in Fig. 3. We hypothesized that the two “modes” of the evolution of the system detected by our analyses are caused by the presence of the feedback loop. To confirm our hypothesis, we broke the loop by modifying our simulation (e.g., made *in silico* “knockout”) and analyzed the trajectories of the same components as in the previous experiment. Fig. 4A shows that the separation is no longer achieved. For further confirmation that the analysis reflects the effects of the feedback loop on the network behavior, we created two sets of trajectories, the first set containing all the components belonging to the loop at both 2- and 5-nM EGF stimuli and the second set containing all the rest of the components of the network at both 2- and 5-nM EGF stimuli. Again, we were able to use LDB classification algorithm on these two new sets of trajectories; as before, distinctive features of time course could be detected. The separation of these two sets on Fig. 4B confirms that common time-frequency characteristics exist between loop components at different stimuli and nonloop components at different stimuli. In summary, this analysis is capable of isolating the topological features that were responsible for the differential behavior of the system as the stimuli varied in strength.

**Cell Cycle and Bistability.** Our next example is based on Tyson and Novak’s model of budding yeast cell-cycle genes (7). It is well established that the core control structure of yeast cell cycle is the protein dimer of cyclin B (CycB) and cyclin-dependent kinase (Cdk). Cdk is responsible for regulating a large variety of cell-cycle events. Its activity depends on forming dimers with CycB, which oscillates through the cell cycle. In Tyson and Novak’s model, the factors that regulate CycB/Cdk are divided into two groups with antagonistic activity: {Cd1h1, CKI, Cdc20} and {SK, CycB/Cdk}.
The first group keeps the cell in phase by keeping the levels of active Cdk/CycB dimers low, because both Cdh1 and Cdc20 degrade CycB and because CKI binds CycB/Cdk to form inactive trimers. The second group counters the degradation of CycB and inactivation of CycB/Cdk complex in two ways. First, SK destroys CKI and, as suggested by Tyson and Novak (7), inactivates Cdh1, and second, CycB/Cdk inactivates Cdh1 by phosphorylation. Sufficiently high levels of Cdk/CycB trigger DNA synthesis and entry into S phase.

Tyson and Novak simulated a model of wild-type cell and two mutant models. In the SK mutant cells, CycB/Cdk cannot be activated, which does not allow the cells to enter S phase. Thus, they continue growing uncontrollably (until they “explode”). However, in the SK/CKI double-mutant model, the cell-cycle behavior is restored. The kinetic mass action-based ordinary differential equation model, formulated with a selected subset of a much larger full set of differential equations, is shown below. The complete set of equations can be found in ref. 7, together with simulations based on Tyson and Novak’s hypothesis that the system possesses two steady states.

\[
[CycB_T] = k_1 - (k_2 + k_3) [Cdh] + k_5 [Cdc20] [CycB_T],
\]

\[
[Cdh1] = \frac{(k_3 + k_4) [Cdc20] (1 - [Cdh1])}{J + 1 - [Cdh1] - (k_4 [SK] + k_{M1} [CycB]) [Cdh1]}. \tag{1}
\]

We hypothesized that, although in the SK/CKI double mutant the cell cycle is restored, it will behave differently from the wild type in terms of the complexity of the trajectories of the same set of genes (CycB, Cdh1, etc.). We investigated this hypothesis by analyzing the gene trajectories of wild type and double mutant generated from the simulation runs starting with varying the initial levels of CycB and Cdh1. Note that although the double-mutant system restores the oscillatory behavior, it does so in a less coherent manner: the projections of the trajectories of the double mutant onto the D-Space are more scattered in comparison with those of the wild type (Fig. 5). This analysis is another demonstration that differences in the topology of a biological network are reflected in different features of the time-course trajectories of the parts of the system. Note that this relation holds, even when mere visual inspection may fail to immediately reveal these features.

Because our approach can be applied to both model-based and experimental time-series data, we used this technique to investigate the time-course gene-expression data obtained with microarray experiments. We chose to analyze the Stanford yeast public mi-
methods using α-factor arrest and Cdc15 mutant induce cell-cycle synchronization, exhibiting much clearer, consistent, and unambiguous time-frequency features as compared with similar approaches based on elutriation; in other words, the trajectories of gene expression synchronized with the first two methods are more coherent than those obtained with the third one. Because time-frequency features can be read in terms of the oscillatory behavior of the cell cycle, our results suggest that in experiments involving yeast cell-cycle synchronization, α-factor arrest and Cdc15 temperature-sensitive mutant lead to more effective synchronization.

Intuitively, this result seems convincing and easy to explain, because both α-factor and Cdc15 mutant arrests target a specific point in the cell cycle, whereas elutriation divides up the collection of cells into groups of similar size with the assumption that a certain size corresponds to a certain phase in the cell cycle. However, it is clear that the precision of elutriation is limited. To test the robustness of our analysis, we examined its performance when the input data is subjected to 20% (white Gaussian) noise and concluded that the time-frequency analysis is reliable on noisy data (Fig. 6B).

**Immune System and Its Adaptive Behavior.** The last example illustrates the analysis of the dynamics of the immune response simulated by using IMMSIM++ (9–11). IMMSIM++ provides a population-level model of the process of clonal selection during the humoral immune response. It creates a time-course picture of how antigen is captured and processed and how antigen processing affects the population and activity of the cellular components of the immune system: B cells, T cells, nonspecific antigen-presenting cells (APCs), antigens, antibodies, and immune complexes. In contrast to the previous simulations based on differential equations, IMMSIM++ employs hierarchical automata to simulate immune response. Specifically, at every time point, the system is found in a certain state defined in terms of the number of the components participating in the immune response and the simulator applied a set of rules to determine what the next state will be. Note that such an approach based on simulation gives no indication about the topology of the system. Nonetheless, we can still detect important differences in the quality of the primary and secondary seed responses.
immune response evident in the time-course data generated by the simulation.

Intuitively, the purpose of clonal selection is to allow the immune system to “remember” the immunal response to a particular antigen by stimulating the growth of particular clones on memory T and B cells that bind to the antigen with high affinity. The memory stored in T and B cells speeds up the response of the organism to a future infection by the same antigen.

The preceding intuition can be tested rigorously by a multiresolution time-frequency analysis of the response curves to different histories of antigen exposures. We hypothesized that an immune system at different initial conditions (i.e., with the initial number of different immune system cells varying from simulation run to run) will exhibit widely varying primary responses to an injection of an antigen, whereas its secondary responses to the injection of the same antigen are going to be essentially the same. The amount of injected antigen is varied within a selected range of values. This adaptive behavior of the system appears to optimize the effectiveness of the secondary response, regardless of the differences in the primary response.

Our hypothesis was tested as follows. IMMSIM++ simulations generated trajectories of major components of the immune system at different initial counts of APC. It recorded the time-series data for the following components: APC..TOTAL (total amount of APC), APC..EXPOSING (amount of APC to which the immune system was exposed), B..TOTAL (total amount of B cells), B..EXPOSING (amount of B cells exposed to antigen), B..MEMORY (total amount of B memory cells), IC (amount of immune complexes), T..TOTAL (total amount of T cells), T..MEMORY (total amount of T memory cells), ANTIGEN..TOTAL (total amount of antigens), and ANTIBODY..TOTAL (total amount of antibodies).

When only the trajectories of the primary and secondary antibody responses were projected on to a 2D D-Space (i.e., the first two major basis functions were considered), they separated into two clusters; the cluster with secondary response is tighter than the cluster with the primary response. These clusters (see Fig. 7, which is published as supporting information on the PNAS web site) lead to the conclusion that the secondary response curves “prune” certain redundant components from their trajectories, thus suggesting that the secondary responses are simpler and more focused. We further analyzed the primary and secondary response curves of all the major components of the system. The two clusters in Fig. 8, which is published as supporting information on the PNAS web site, correspond to the types of cells that persist through the whole simulation (such as B..TOTAL, for example), and the cells which appear only when the immune system is responding to the infection (such as B..EXPOSING). Within both clusters, there is still a distinction between the primary and secondary response curves; primary response curves are significantly more scattered.

These examples have a simple intuitive explanation. For example, if one assumes that the specificity of immune response is determined by the ratio of highly specific B cells and cells with lower affinity for the antigen, then before the primary response, this ratio is low because there are very few clones of B cells, which secrete antibodies with high specificity for the antigen. Thus, during the primary response, B cells with low-affinity receptors also get activated, resulting in a somewhat nonspecific response curve. Nonetheless, because the B cells with higher specificity for the antigen proliferate more efficiently, the proportion of the highly specific B cells in the total B cells population increases. As a result, during the secondary response, the antigen encounters specific B cells with higher probability; fewer B cells with low specificity get activated; and a more specific secondary response, composed of responses of high-specificity proliferated B cells, emerges.

Conclusion

Our focus in this article is on a technique that is capable of detecting significantly distinct features in the dynamics of various biological systems. Need for this technique is spurred by the long-standing interest of the biological community in studying cellular processes, not just in terms of their steady-state outcomes but also their evolution. Unfortunately, the experimental techniques often allow the former rather than the latter approach.

However, it is precisely the differences in the dynamic behavior that might be crucial in understanding many broad biological questions, such as how reversible changes in signaling pathways lead to irreversible changes in cell fate. As Ferrell (2) observed, the answer to this question has been proposed already 40 years ago by Monod and Jacob, who pointed to positive and negative feedback loops in gene regulatory systems for an explanation. According to their theory, modules comprising such feedback loops are “capable of ‘remembering’ a transient stimulus long after the stimulus was removed” (2). In other words, the systems like MAPK cascade simulated by Bhalla and Iyengar (1) may be the key to answering this question.

Although the discussion in Ferrell’s review and elsewhere has concentrated on the bistable systems for an explanation, it is conceivable that a different kind of biological phenomenon may provide a better answer; a specific dynamic behavior of a biological system could disappear with few changes to topologies in strategic points. Yet, these changes may not alter the general characteristics of the biological system but only modify more subtle features of the dynamics. This form of robustness will have a tangible effect on the biological processes that it affects, and it will be selected by a Darwinian evolution. Unfortunately, the traditional tools of mathematical biology (e.g., finding conserved patterns, motifs, or sensitivity analysis) are inadequate here. Detecting various modes in the dynamics is a step towards understanding this multifaceted nature of robustness in biology.

We thank U. Bhalla, C. Calcagnio, F. Celada, D. Ghersia, and M. Wigler for offering many exciting ideas, useful suggestions, and help with various simulation systems (GENESIS, MATLAB, and IMMSIM). We also thank many of our colleagues from the New York University Bioinformatics group, Courant Institute, Cold Spring Harbor Laboratory, Mt. Sinai School of Medicine, Tata Institute of Fundamental Research, Università di Udine, and Università di Genova for directly or indirectly contributing to this effort. This work was supported by grants from the National Science Foundation Qubic program; the Defense Advanced Research Planning Agency; a Howard Hughes Medical Institute biomedical support research grant; the U.S. Department of Energy; the U.S. Air Force; the National Institutes of Health; and the New York State Office of Science, Technology, and Academic Research.

References