320

Statistical Construction of Chemical Reaction Mechanisms from Measured Time-Series

Adam Arkin[†,‡] and John Ross[*,†]

Department of Chemistry and Department of Neurobiology, School of Medicine, Stanford University, Stanford, California 94305

Received: September 12, 1994; In Final Form: November 3, 1994[⊕]

We present a new approach to the construction of chemical reaction mechanisms by methods derived, in part, from electronic circuit and systems theoretical techniques. The approach, correlation metric construction (CMC), is based on the calculation and analysis of a time-lagged multivariate correlation function of a set of time-series of chemical concentrations. The time-series are composed of the observed responses of species composing a chemical reaction network to random changes in the concentration of a set of input species. The four-dimensional correlation-time lag function is subsequently transformed into a metric distance function and is analyzed by multidimensional scaling and cluster analysis in order to (1) determine a measure of effective dimensionality of the system; (2) construct a correlation diagram of the reaction mechanism that graphically recapitulates, in large part, the reaction steps in the network by a technique that emphasizes the strengths of coupling among the constituent species; and (3) determine the hierarchy of control in the network and identify possible weakly coupled or uncoupled subsystems. In order to demonstrate the technique, we analyze three different models of common types of chemical reactions. The analysis of these examples, which include enzymatic substrate cycles, mass action kinetics, networks with rate-determining steps, and networks satisfying the steady-state hypothesis, demonstrates that CMC is able to construct informative diagrams which construct, in large part, the underlying chemical reactions and strengths of interactions among the measured species in the network.

I. Introduction

The establishment of chemical and biochemical reaction mechanisms can be a difficult task. In prior work we have proposed a methodology for deducing at least core parts of oscillatory reaction mechanisms from experiments designed for that purpose.[1-4] Key features of the complex reaction mechanisms of both the oscillatory chlorite-iodide[5] and horseradish peroxidase[6] reactions have been deduced by these methods.[6] In parallel work we showed the possibility of implementation of logic functions and computations by means of macroscopic chemical kinetics.[7-12] From there we proceeded to demonstrate that parts of complex biochemical reaction networks implement logic functions.[13] From these two lines of research emerges the possibility of a new approach to the problem of the construction and interpretation of reaction mechanisms by the development and application of analyses from electronic circuit theory, general systems theory,[14-16] and multivariate statistics.[17-19] We demonstrate this approach with the application of just one method, which we call correlation metric construction (CMC), to common types of chemical and biochemical reaction mechanisms.

The goal of a CMC is to derive, from a set of time-series of chemical concentrations (collected near the steady state (or equilibrium) of a chemical reaction network), a skeletal diagram representing the connectivity (the reactions among measured species) of the mechanism and obtain a graphical measure of the regulatory structure of the network. The analysis is designed to find the sites of strong control within networks of chemical reactions and to simplify the analysis of such networks by identifying possible chemical subsystems, each of which might then be analyzed separately. CMC is composed of multiple correlation analysis[13,17,18,20] from which a connection matrix among the measured species is defined; multidimensional scaling,[17,18,21] which produces the graphical representation of the correlation structure of the mechanism; and hierarchical clustering[17,18,21,22] which delineates a hierarchy of regulation and weakly coupled subsystems within the mechanism.

II. Methods and Applications

Consider a chemical system as shown in Figure 1. This mechanism performs the function of a biochemical NAND gate.[13] Mechanisms of this type are common in biochemical networks. For example, the subnetwork of Figure 1 containing S_3-S_5 is based on a simple model of fructose interconversion in glycolysis,[13] and the subnetwork composed of S_6-S_7 is similar to the phosphorylation-dephosphorylation cycles found in cyclic cascades.[23,24] The goal of CMC is to determine both the regulatory structure and the connectivity of the species in the elementary reaction steps of this mechanism, solely from measurements of the response of the concentrations of S_3-S_7 to fluctuations in the input concentrations, I_1 and I_2. We assume initially that it is possible to identify and measure all the chemical species that make up the chemical network. This is a strong assumption and is seldom satisfiable; we return to this point later in this article and following articles. Further, we assume that we may impose concentration variations (noise) independently on a subset of chemical species in the network at each time in the ordered set of times $\mathcal{T} = \{t_1, t_2, t_3, ..., t_n\}$. This subset of chemical species, I, is designated as the inputs to the system. The rest of the species are designated as the set, S. The total number of chemical species is, therefore, $|I| + |S| = I + S = M$. For the system in Figure 1, we choose $I = \{I_1, I_2\}$ and $S = \{S_3, S_4, S_5, S_6, S_7\}$; thus, $M = 7$. The differences between adjacent measurement times, t_n and t_{n-1}, in the set $\mathcal{T}$ are also assumed to be on the order of, or longer than, the slowest relaxation time in the network; that is, the network is

* Author to whom correspondence should be addressed.
† Department of Chemistry.
‡ Department of Neurobiology, School of Medicine.
⊕ Abstract published in *Advance ACS Abstracts*, January 1, 1995.

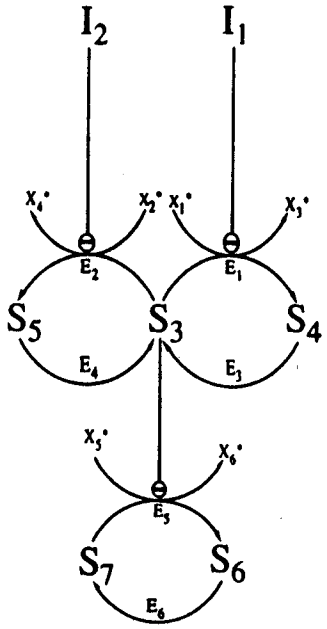

Figure 1. Chemical reaction mechanism representing a biochemical NAND gate: At steady state, the concentration of species S_6 is low if and only if the concentrations of both species I_1 and I_2 are high. All species with asterisks are held constant by buffering. Thus, the system is formally open though there are two conservation constraints. The first constraint conserves the total concentration of $S_3 + S_4 + S_5$, and the second conserves $S_6 + S_7$. All enzyme-catalyzed reactions in this model are governed by simple Michaelis–Menten kinetics. Lines ending in a circle-enclosed minus sign over an enzymatic reaction step indicate that the corresponding enzyme is inhibited (noncompetitively) by the relevant chemical species. We have set the dissociation constants, $K_{D,i}$, of each of the enzymes, E_1–E_6, from their respective substrates equal to 5 concentration units. The inhibition constants, K_{I1} and K_{I2}, for the noncompetitive inhibition of E_1 and E_2 by I_1 and I_2, respectively, are both equal to 1 unit. The V_{max} for both E_1 and E_2 is set to 5 units/s, and that for E_3 and E_4 is 1 unit/s. The V_{max}'s for E_5 and E_6 are 10 and 1 units/s, respectively.

driven near its steady state. If this assumption is not strongly violated, then the following analysis still produces meaningful results. However, if the driving occurs rapidly compared to the relaxation times, the network acts as a low-pass filter and the input signal is degraded. The frequency response of chemical networks is an important area for further development of this method but is left for later investigation.

A CMC of an unknown chemical reaction mechanism proceeds with the following seven steps.:

(1) The set of measurements of the concentrations of the species in $\mathcal{J}$ is obtained as a function of the externally controlled concentrations of the species in $\mathcal{I}$ at each of the time points $\mathcal{T}$. Figure 2 is a plot of the time series for each of the species in $\mathcal{I}$ and $\mathcal{J}$ in Figure 1.[25] One time point is taken every 10 s for 3600 s. The effects of using a much smaller set of observations is discussed in section IV. The first two plots are the time-series for the two externally controlled inputs. The concentrations of I_1 and I_2 at each time point are chosen from a Gaussian distribution centered at 30 concentration units with a standard deviation of 30 units. The choice of Gaussian noise guarantees that in the long time limit the entire state-space of the two inputs is sampled and that there are no autocorrelations or cross correlations between the input species. The bottom five times-series are the responses of S_3–S_7 to the concentration variations of the inputs.

(2) The time-lagged-correlation matrix, $\mathbf{R}(\tau) = (r_{ij}(\tau))$, is calculated for all $\tau < n$ with the equations[20]

$$S_{ij}(\tau) = \langle (x_i(t) - \bar{x}_i)(x_j(t+\tau) - \bar{x}_j) \rangle \qquad (1)$$

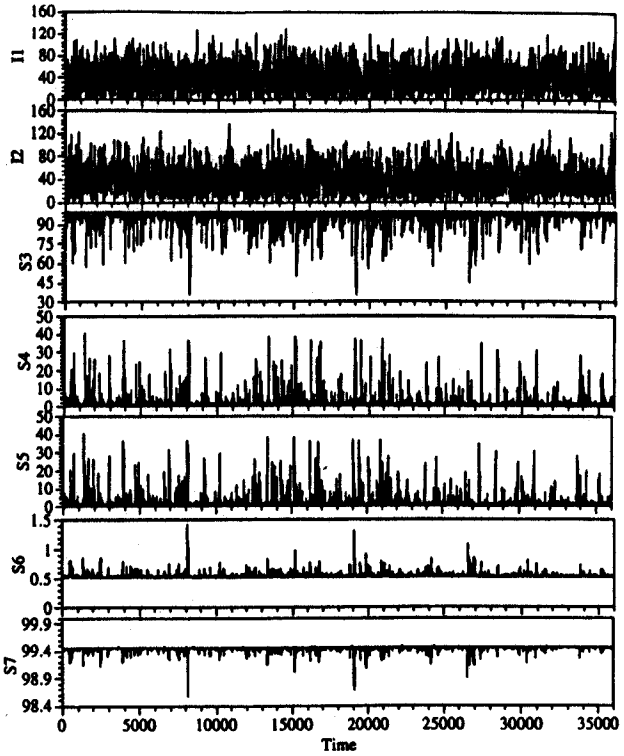

Figure 2. Plot of the calculated concentration time-series for all the species composing the mechanism in Figure 1. Only the first two time courses (those for I_1 and I_2) are set by the experimenter. The concentrations of I_1 and I_2 are chosen independently from a Gaussian distribution with a mean and standard deviation of 30.0 concentration units. Since the lower limit of concentration is zero, the actual distribution of input concentrations has a tail toward high concentrations. See step 1 in the text for a full explanation.

$$r_{ij}(\tau) = \frac{S_{ij}(\tau)}{\sqrt{S_{ii}(\tau)\, S_{jj}(\tau)}} \qquad (2)$$

where the angle brackets denote a time average, $x_i(t)$ is the tth time point of the time-series generated from species i, and $\bar{x}_i$ is the time average of the ith time-series. The indices i and j range over all species in the set $I \cup J$. $\mathbf{R}(\tau)$ is dependent, in a complicated way, on the elementary reactions and their rate coefficients. Figure 3 is a set of plots of some three-dimensional cross sections from the four-dimensional $\mathbf{R}(\tau)$ surface calculated from the data in Figure 2. Each cross section represents the correlations, at all calculated time lags (here every ±10 s up to a time lag of ±190 s), of the times-series corresponding to a given species with those of each of the other species, and itself. If the system were truly at steady state at each time point, then the correlation surface for combinational networks (those in which there are no feedback loops) would be flat except on the $\tau = 0$ plane. Figure 3 shows only four cross sections corresponding to the choice of three independent species and one of the inputs. Since there are two conservation constraints (see below) in this mechanism, only three of the S_n's are independent; hence, these plots are representative of the entire $\mathbf{R}(\tau)$ surface (except for the $r_{I_2,J}(\tau)$ section). Examination of the $r_{S_6,J}(\tau)$ section of $\mathbf{R}(\tau)$ (Figure 3D) shows that, as expected, the concentration of S_6 perfectly anticorrelates with that of S_7 (by conservation) and negatively correlates with that of S_3.

Figure 4 shows two projections of the $r_{S_3,J}(\tau)$ cross section (Figure 3B). Figure 4A is a projection on the species-correlation plane. This projection makes clear the relative correlations of each of the species with S_3. Figure 4B is a projection on the time lag-correlation plane. Here, a rough sequence of events is clear when it is noted that if one species correlates with a

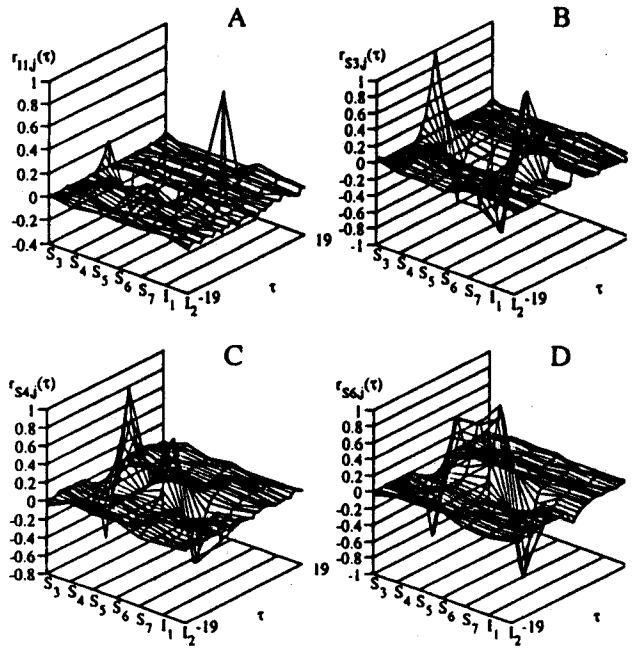

Figure 3. Plots of cross sections through the four-dimensional time lag-correlation surface calculated from eqs 1 and 2. The cross sections are A = $r_{I_1 j}(\tau)$, B = $r_{S_3 j}(\tau)$, C = $r_{S_4 j}(\tau)$, and D = $r_{S_6 j}(\tau)$. See step 2 in the text for explanation.

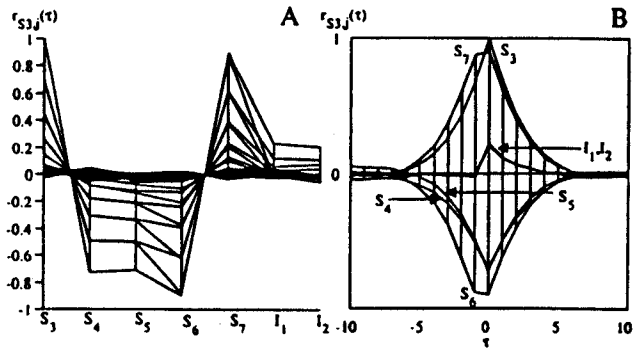

Figure 4. Projections of Figure 3B down the species (A) and time lag (B) axes. Figure 4A emphasizes the correlation of S_3 with all the other species. It correlates most highly with S_6 and S_7, next most with S_4 and S_5, and finally with I_1 and I_2. Figure 4B (truncated to emphasize the region between $\tau = \pm 10$ units) shows the large amount of time information which is largely ignored by the simple analysis presented here. If a species correlates at a positive lag with respect to S_3, then that species leads S_3 in time; a negative lag implies the correlating species follows S_3. Since Figure 4B shows that the correlation with the inputs tails toward positive lags, action on the inputs precedes action at S_3. Action at S_4 and S_5 occurs simultaneously with S_3, and S_6 and S_7 follow action at S_3 (as indicated by their left tails). In this particular network this sequence of lags relates directly to the causal sequence. The inputs affect the concentration of S_3–S_5, which then affects the concentration of S_6 and S_7. The amount of this information available for a given experiment is dependent on the magnitude of the rate coefficients with respect to the time between measurements.

second species at a positive lag, then the variation of the first species occurs (on average) before that of the second. The correlation of a species with itself at nonzero lag is expected due to relaxation to the steady state and should be symmetric around zero lag. The sequence of events with respect to S_3 is, therefore, as follows: I_1 and I_2 change, followed by the simultaneous changes in S_3–S_5, after which S_6 and S_7 change. In this case, this time line is a sequence of causality. However, in branched networks or networks with feedback, causality may not always be determined in this way. Further, if the time between measurements had been 70 s rather than 10 s, the network would have been at steady state and no time delays would be observed.[26]

TABLE 1: Listing of the "Significant" Connections Calculated among the Chemical Species Composing the Reaction Mechanism in Figure 1 (see Step 3 in the Text)

	S_4	S_5	S_6	S_7
I_1	−0.31			
I_2		−0.32		
S_3	−0.72	−0.71	−0.90	0.90
S_6				−1.00

The nullity (the number of zero eigenvalues) and null-space (the space spanned by the corresponding eigenvectors) of the correlation matrix determine the number of conservation constraints and rapidly established quasi-equilibria in the network and the species involved, respectively (since species involved in such relations are completely dependent on one another). For example, the eigenvalues of the (zero lag) correlation matrix derived from the time-series in Figure 3 are $\lambda_r = \{3.89, 1.3, 0.97, 0.67, 0.17, 1.7 \times 10^{-7}, 7.0 \times 10^{-9}\}$. Thus, since the last two eigenvalues are so small compared to the first five, the nullity of the matrix is 2, as expected from the two conservation conditions.

(3) A connection algorithm generates an approximate dependency list among the different species. Ideally, we would like the dependency list to reflect directly the reactions between species. However, dependency in this context is not necessarily related to direct causation but rather to a high degree of association among sets of variables.[26] There are a number of common analyses, such as multiple regression, canonical correlation,[18] linear structural relations,[27,28] and Box–Jenkins time-series analysis,[29] which attempt to determine the dependencies of a set of observable dependent variables on known independent variables. All of these techniques are essentially regression analyses in which an estimate of the mean of one set of variables is made based on knowledge of another set of variables. In order to perform the analyses, a model of the system generally has to be specified and fit. Since one goal of CMC is to construct an approximate model of the interspecies interactions, we wish to avoid the explicit specification of a model (chemical mechanism). Correlation-based analysis measures a symmetric degree of association between two variables. Thus, correlations yield a convenient metric of distance between two species (see below). Since the balance of our method (steps 4–7) relies on such a metric, we develop a simple correlation-based agglomerative dependency algorithm with the following iterated steps: (a) Designate each species by its own group. (b) Find the two groups, *i* and *j*, each containing a disjoint subset of species for which the magnitude of correlation (at any lag) between a species in *i* and one in *j* is the maximum over all pairs of groups. (c) Make these two groups into one group and list the connection (found in step b) between the two species (one from each original group) with the maximum correlation. If two or more species from one group correlate with species in the other with (nearly) the same maximum magnitude, then list connections between these as well. All listed connections are designated as "significant" connections. (d) Go back to b, now with one less group, and repeat the steps until there is only a single group left. This procedure creates a singly linked system graph in which every species is connected to at least one other species.

Application of the connection algorithm to the full correlation surface of the NAND mechanism yields the values of the significant connections listed in Table 1. The algorithm neglects the direct effect of each input species on the concentration of S_3 since we are using only a single-link dependency; the

magnitude of the correlation between the inputs and S_3 is weak since S_3 is maximally affected only when the inputs are both low or both high at the same time (a relatively rare event). It may be possible to eliminate such statistical misses and redundancies in the connection list by including chemical knowledge, incorporating the available correlation time lag information, and application of the transinformation criterion of probabilistic reconstruction analysis,[14-16,30,31] which uses conditional probabilities derived from the full (calculated) joint probability density function generated from the time-series. We report on this measure of connection in later work. Lastly, the difference in correlations of S_4 and S_5 to S_3 is merely a result of the sampling statistics.

The connections between species, since they are based on correlations, represent a noncausal structural model of the system. Correlations may be decomposed into four parts: (1) direct effects in which one variable is a direct antecedent to a second (e.g. one species directly converted to another by a chemical reaction); (2) indirect effects where one variable influences another by way of a third (e.g. one species affects the production of a second which is then converted directly to a third); (3) spurious effects which occur when two variables have a common antecedent (e.g. when one species is converted into two other species by separate reactions); and (4) unanalyzed effects which arise from correlations between the externally controlled variables.[32] In order to construct a mechanism (a causal model), only the first contribution to the correlation between two variables should be considered. We do not currently separate the measured correlation into these explicit components. In the absence of special knowledge of the chemical mechanism, methods to do so, such as path analysis and LISREL,[27,28] require extensive and complex calculation as well as a number of restrictive assumptions. Our simple definition of connection, though possibly not as informative as a full causal analysis, yields a good first guess for a relational structure among the species (ideally defined by the first component of correlation above).

A partial description of the function of the network may be read from the sign of the connections in the list. For example, since I_1 and I_2 are negatively correlated with S_4 and S_5, both of which are negatively correlated with S_3, we may hypothesize that $[S_3]$ is high only when $[I_1]$ or $[I_2]$ or both are high, and $[S_3]$ is low otherwise. So the subsystem of Figure 1, composed of S_3-S_5 and with the output defined as $[S_3]$, may be functionally analogous to a logical OR (or an AND) between I_1 and I_2. In this particular case, the function is an AND.

(4) The time-lagged correlation matrix, $\mathbf{R}(\tau)$, is converted into a Euclidean distance matrix with the canonical transform[18]

$$d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{1/2} = \sqrt{2}(1.0 - c_{ij})^{1/2} \quad (3)$$

$$c_{ij} = \max|r_{ij}(\tau)|_\tau \quad (4)$$

where the second equality in eq 3 follows from the properties of the correlation matrix. The formula in eq 4 defines c_{ij} to be the absolute value of the maximum correlation between the time-series for species i and that of species j, regardless of the value of τ.[33] We define $\mathbf{D} = (d_{ij})$ to be the distance matrix. Since $\mathbf{D}$ is Euclidean, its elements automatically satisfy the three standard tests for a metric space: identity, symmetry, and the triangle inequality. In the case of perfect correlation between two variables, the triangle inequality is violated, a situation that may be remedied (without dire consequence) by adding a small value, $\epsilon = 1 \times 10^{-10}$, to the distance between them. The particular metric defined by eq 3 is a measure of independence between two variables. If the correlation between two variables is small, then the distance between them is large.

(5) The classical multidimensional scaling (MDS) method[18,21] is applied to the distance matrix calculated in step 4 in order to find both the dimensionality, Δ, of the system and a consistent configuration of points representing each of the species. This is accomplished by finding the eigenvalues, λ_i, and eigenvectors, $\mathbf{z}_i$, of the centered inner product matrix, $\mathbf{B}$, defined by

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}(d_{ij}^2)\mathbf{H} \quad (5)$$

$$\mathbf{H} = \mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}' \quad (6)$$

where $\mathbf{H}$ is the centering matrix,[18] $\mathbf{I}$ is the $M \times M$ identity matrix, and $\mathbf{1}\mathbf{1}'$ is the $M \times M$ unit matrix. The operation of the symmetric, idempotent matrix $\mathbf{H}$ on the vector $\mathbf{x}$ has the effect of subtracting off the mean of the entries of $\mathbf{x}$ from each of its elements, i.e., $\mathbf{Hx} = \mathbf{x} - \bar{x}\mathbf{1}$ where $\bar{x} = n^{-1}\sum x_i$. The number of significant eigenvalues of $\mathbf{B}$ (defined in Table 2) is the dimensionality, Δ, of the system, and the vectors of the first Δ coordinates of the M eigenvectors compose the principal coordinates of points representing the chemical species in the correlation diagram. (Formally, each point represents the particular time-series generated for a given chemical species.) The eigenvectors are normalized such that $\mathbf{z}_i\cdot\mathbf{z}_i = \lambda_i$. The distance between each pair of points is inversely related to the correlation between the corresponding species. If the M series are independent, then all the d_{ij} are equal to $\sqrt{2}$ and points representing each species must fall on the vertices of a regular $M - 1$ dimensional hypertetrahedron. At the other extreme, when all species perfectly correlate (or anticorrelate), then the d_{ij} are all equal to 0 and there is a single degenerate point in the system. By construction, however, the inputs are completely uncorrelated so the minimum dimension derived from a CMC is $|I| - 1$. Most often, we are interested in the first two principle coordinates of the MDS solution since the configurations may then be plotted on a plane and are thus easy to visualize.

The numerical results from the MDS analysis of the distance matrix derived from eqs 3 and 4 are shown in Table 2. The columns of Table 2A are the eigenvectors of $\mathbf{B}$, and the rows are the coordinates of each point (time-series). Each eigenvector, $\mathbf{z}_k$, corresponds to the projection of the kth coordinate of each of the points on an orthogonal basis vector. Each of the eigenvalues (Table 2B) are an indicator of the degree to which the vectors from the origin of the configuration to each point are projected along the corresponding basis vector. In this example, over 99% (for $a_{2,3}$) of the distance matrix is "explained" by the first three eigenvalues. Thus, we find the dimensionality $\Delta \approx 3$, which represents a reduction of three dimensions from the theoretical maximum of six. Two of these dimensions are lost due to the two conservation constraints. The third dimension is lost due to constraints imposed by the degree of interaction among the subsystems. The two-dimensional projection of the coordinates of the points listed in Table 2A is shown in Figure 5A. Each pair of points with a nonzero entry in Table 1 is connected by a line. This represents the correlation-metric diagram which depends on the reaction mechanism and rate coefficients (as well as the properties of the perturbations, such as the frequency spectra, average values, and standard deviations). This diagram reproduces many features of the standard mechanistic diagram (Figure 1) but differs in that it has additional information, such as the extent of coupling (tightly or loosely) among subsystems in the network.

TABLE 2: Eigenvectors and Eigenvalues of the Matrix B (Eq 5) Determined by the Classical MDS Analysis of a Distance Matrix Calculated from the Correlation Surface in Figure 3 (See Steps 4 and 5 in the Text)[a]

(A) Eigenvectors, z_k

point/z_k	1	2	3	4	5	6	7
1 (I_1)	6.68E−01[b]	−5.84E−01	4.05E−01	5.51E−02	1.77E−02	−8.20E−09	−3.52E−10
2 (I_2)	7.00E−01	5.26E−01	−4.30E−01	4.93E−02	1.42E−02	−8.20E−09	−3.52E−10
3 (S_7)	−4.20E−01	7.29E−03	−8.16E−03	2.05E−01	1.90E−03	−6.82E−09	−7.67E−09
4 (S_6)	−4.20E−01	7.29E−03	−8.16E−03	2.05E−01	1.90E−03	−9.58E−09	6.97E−09
5 (S_4)	−1.44E−01	−5.51E−01	−4.02E−01	−1.60E−01	−7.55E−02	−8.20E−09	−3.52E−10
6 (S_5)	−7.15E−02	5.60E−01	4.30E−01	−1.38E−01	−7.65E−02	−8.20E−09	−3.52E−10
7 (S_3)	−3.14E−01	3.49E−02	1.27E−02	−2.16E−01	1.16E−01	−8.20E−09	−3.52E−10

(B) Eigenvalues, λ_k

k	λ_k	$a_{1,k}$	$a_{2,k}$
1	1.413496E+00	3.978311E−01	4.937104E−01
2	1.237497E+00	7.461268E−01	8.721279E−01
3	6.958617E−01	9.419783E−01	9.917824E−01
4	1.805556E−01	9.927960E−01	9.998381E−01
5	2.559576E−02	1.000000E+00	1.000000E+00
6	4.747935E−16	1.000000E+00	1.000000E+00
7	1.080736E−16	1.000000E+00	1.000000E+00

[a] The coefficients $a_{1,k}$ and $a_{2,k}$ are agreement measures of the degree to which the distance matrix is "explained" by the k-dimensional MDS solution.[18] The measures are calculated from: $a_{1,k} = (\sum_{i=1}^{k}\lambda_i)/\sum_{i=1}^{n}|\lambda_i|$ and $a_{2,k} = (\sum_{i=1}^{k}\lambda_i^2)/\sum_{i=1}^{n}\lambda_i^2$, where the eigenvalues are sorted in decreasing order. Thus, according to $a_{2,k}$, 49.4% of the distance matrix is explained by one dimension ($a_{2,1} = 0.494$) and 100% of the matrix is explained by a configuration in five dimensions ($a_{2,5} = 1.0$). Columns 1 and 2 of the eigenvector matrix define the x and y positions of points in Figure 4A, respectively. [b] Read as 6.68×10^{-1}.

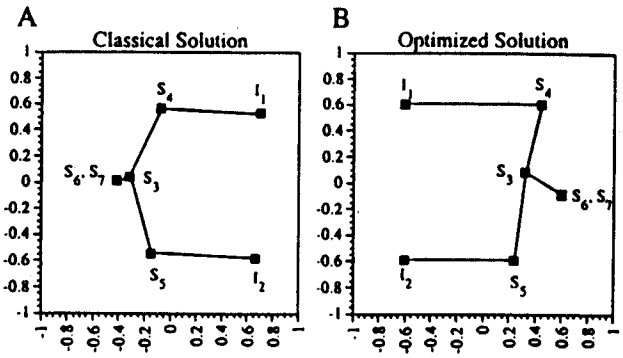

Figure 5. Classical and optimized MDS solutions calculated from the canonically transformed correlation surface from Figure 3. Figure 5A is the two-dimensional projection of the three-dimensional ($a_{2,3}$ = 99.2%) (or four-dimensional; $a_{2,4}$ = 99.98%) object found by the matrix method described by step 5 in the text. The coordinates for the points in Figure 5A are the same as the first two columns of the matrix of eigenvectors in Table 2A. The second diagram (Figure 5B) is derived from the MDS optimization method discussed in step 6. The lines between points in both diagrams are obtained from the results of step 3 (Table 1). The respective stresses of the two diagrams (A and B) are 3.50 and 0.71, as calculated from eq 7. Thus, Figure 5B is more representative of the measured distance matrix. Both diagrams correspond to rotated and slightly distorted approximations to the reaction mechanism in Figure 1.

(6) Figure 5A is a projection of the high-dimensional MDS object onto two dimensions. The distances between the points in the 2-D representation are therefore less than or equal to actual measured distances. A more representative 2-D diagram may often be obtained with an optimization-based MDS method that allows the distances between the optimized points to be both greater, equal to, and less than the actual distances.[21] Here, an optimization algorithm minimizes a stress function. One possible definition of such a function is

$$\text{Stress}(\delta) = \left(\sum_{i,j}(\tilde{d}_{ij}(\delta) - d_{ij})^2\right)^{1/2} \qquad (7)$$

where $\tilde{d}_{ij}(\delta)$ are the calculated geometrical distances between pairs of initially randomly placed δ-dimensional points in a test configuration. The positions of these test points are moved by the optimization algorithm such that the distances between them are as close as possible, for a given δ, to the experimentally measured distances between the species, d_{ij}. Since we have already determined the full dimensionality of the object in step 5, we set $\delta = 2$ and minimize the stress using a simulated annealing algorithm,[34] which is a numerical global optimization techniques.

The diagram derived with this algorithm from our example is shown in Figure 5B. To determine whether part A or B of Figure 5 is the better representation of the four-dimensional diagram, we calculate the stress of each diagram with eq 7. The diagram with lower stress is the most representative configuration. According to the stress criterion, the configuration of points in the optimized diagram is more representative of the actual distance matrix. (See the caption of Figure 5.) Note that all rotations, reflections, and translations of the MDS diagrams are also valid MDS solutions. Both diagrams, parts A and B of Figure 5, are rotated and slightly distorted versions of the diagram of the reaction mechanism shown in Figure 1. Thus, from the state measurements (calculations) of the time-series, we derive a construction related to the reaction mechanism for the system, but with the added information of the relative coupling strengths among species.

(7) Finally, a cluster analysis is performed on the distance matrix. This method is used to summarize the grouping of chemical subsystems within the reaction mechanism and to give a hierarchy of interactions among the subsystems. There are many techniques of cluster analysis,[18,21,22] but for simplicity we employ a nonparametric hierarchical clustering technique called the weighted pair group method using arithmetic averages.[22] The algorithm operates on D in three iterated steps: (a) Search the $M \times M$ distance matrix for the minimum distance between pairs of clusters and let this distance be d_{ij}. In the initial step, there are M clusters (u_1, u_2, ..., u_M) each containing a single point (species). (b) Define a new cluster, u_k, containing the two objects (i and j) with the minimum distance between them. Define the branching depth of this cluster as $h_{ij} = d_{ij}/2.0$. Finally, define the distance between the new cluster, k, and all other clusters, $l \neq i$ or j, with the equation $d_{kl} = (d_{il} + d_{jl})/2.0$.

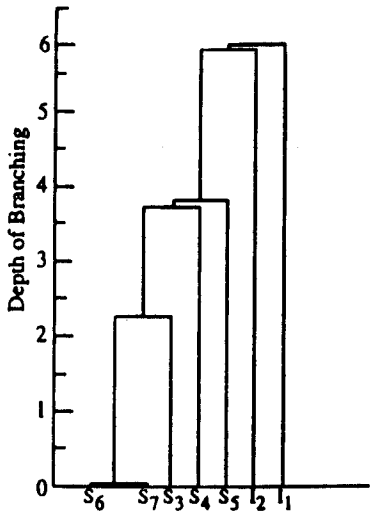

Figure 6. Hierarchically clustered dendrogram calculated from the CMC analysis of the mechanism in Figure 1. The dendrogram is produced as described in step 7 in the text. The hierarchy is a good representation of the flow of control from the inputs down to S_6 and S_7.

(c) Delete clusters i and j from the distance matrix and add the newly computed cluster distances to the matrix. If M is not equal to 1, then decrease M by one and return to step 1, otherwise the algorithm is done. This procedure generates a dendrogram.

The dendrogram, derived from a cluster analysis of the distance matrix calculated in step 4, is shown in Figure 6. The tree may be interpreted as follows. Since the branching depths of species I_1 and I_2 are nearly identical but they do not belong to the same cluster, we may consider the two as separate subsystems which couple with similar strengths to the rest of the network (the species below the inputs in the diagram). A similar assumption may be made about species S_4 and S_5. The diagram is then simple to analyze. Since we control species I_1 and I_2, these in turn most strongly influence the changes in concentrations of both S_4 and S_5. Species S_4 and S_5 then combine to control the concentrations in the subnetwork $\{S_3, S_6, S_7\}$. This three-member subnetwork may be further divided into $\{S_3\}$ and $\{S_6, S_7\}$. From the correlation-metric diagram and the eigenanalysis of the correlation matrix we know that S_4 and S_5 are the determinants of S_3 (since the connections are significant and the three species are involved in a conservation relation), which subsequently controls the concentrations of S_6 and S_7. These last two species also satisfy a conservation constraint. Thus, a hierarchical diagram of control is derived.

III. Further Examples

The following examples are constructed in order to clarify the usage and interpretations of CMC analysis. Figure 7 shows a chemical system composed of two subsystems, one like that of Figure 1 (subsystem 1) and another realizing a NOT I_1 AND NOT I_2 function (subsystem 2). The kinetics of subsystem 1 are chosen to be somewhat faster than the analogous system in Figure 1. Subsystem 2 is composed from substrate cycles similar to those in subsystem 1, but the kinetics of the enzyme reactions are much slower than that of subsystem 1. This sytem was chosen to demonstrate the concept of chemical subsystems as defined by CMC and to demonstrate two interesting interrelated problems which arise during the analysis: (1) ambiguity arising from common causal antecedents and (2) low-pass filtering of the input signal(s). Figures 8 and 9 are the result

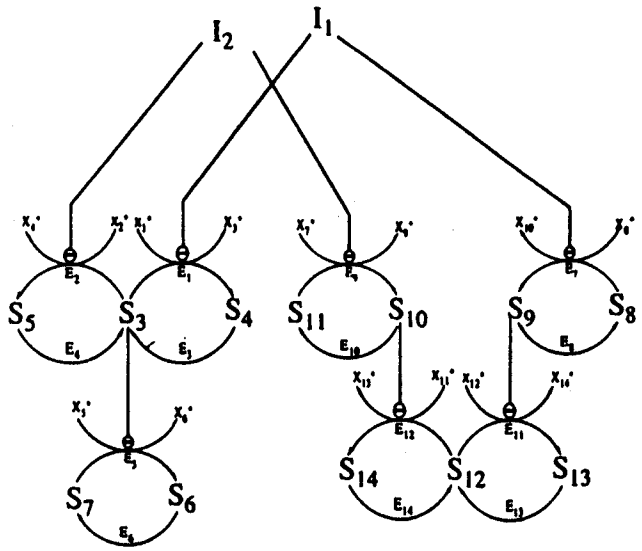

Figure 7. Mechanism composed of two different realizations of a chemical NAND gate. The first NAND-like submechanism is composed of species S_3–S_7 plus the inputs. The second is composed of S_8–S_{14} plus the inputs. Both subsystems have the inputs as common causal antecedents. As in Figure 1, the concentrations of all species bearing an asterisk are considered to be held constant by buffering or external flows.

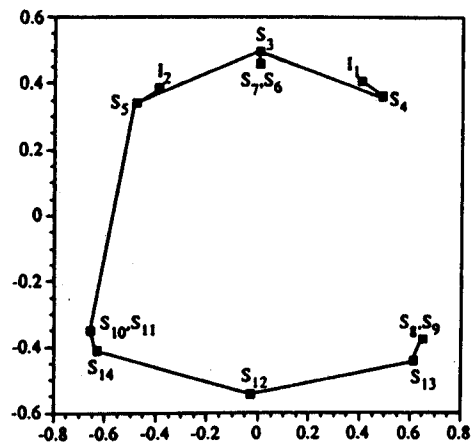

Figure 8. 2-D projection of the classical MDS solution resulting from a CMC analysis of the mechanism in Figure 7. Note that, despite the fact that both NAND gate subsystems of the mechanism are driven by common inputs, the different subsystems group on different half-planes of the diagram. The placement of the inputs of the diagram is due to the choice of rate constants (see text). The reasons for the connection drawn between $\{S_{13}, S_{14}\}$ and S_5 and for the lack of connection between $\{S_{11}, S_{12}\}$ and the other subsystem or inputs are described in the text. For this diagram $a_{2,4} > 99.8\%$ and the stress is 7.47 (versus 3.38 for the 13-dimensional solution).

of the CMC anaysis of this network. The two subsystems separate onto the upper and lower-half-planes of Figure 8 and onto two different branches of the dendrogram in Figure 9, respectively. The MDS configurations of the two subsystems have similar structures, as expected. The input species I_1 and I_2, however, group very close to S_4 and S_5 and on the far side of the MDS diagram away from S_{13} and S_{14}. This occurs because the rates of interconversion of, for example, S_3 and S_4 are faster than the corresponding interconversion of S_8 and S_9; the concentration of species S_4 is able to follow better the fluctuations in the input species than S_8–S_9. If the enzymatic reactions are slow, then not very much material is converted each time the inputs change state. Large changes in concentration are only obtained (assuming Gaussian driving noise) when there are substantial low-frequency components in the input

Figure 9. Hierarchically clustered dendrogram calculated from the CMC analysis of the mechanism in Figure 7. The two different NAND subsystems separate onto the two major branches of the dendrogram. The inputs cluster with the first subsystem (the one similar to the mechanism in Figure 1) since its kinetics are very rapid compared to the second subsystem. The branch containng subsystem 1 is not structurally equivalent to the dendrogram in Figure 6 because, as a result of the kinetics of subsystem 1 being fast compared to those of the mechanism in Figure 1, the inputs couple more tightly to S_4 and S_5.

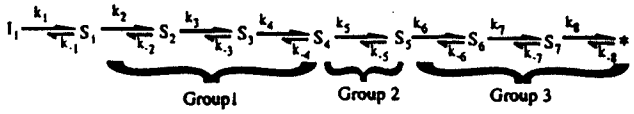

Figure 10. Linear reaction network. All reactions are first-order mass-action kinetics. The back reactions (k_{-1}–k_{-8}) are always 0.1 s^{-1}. k_1 is set to 1.0 s^{-1}. The remainder of the forward rate constants are broken into three groups, as shown. Within a group, all the forward coefficients are assumed to be identical. These rate constants are chosen from three possible rates: 0.7, 70.0, and 7000.0 s^{-1} (slow (S), medium (M), and fast (F)).

signals. Thus, slow reactions act as low-pass filters for their input signals. Though both the subnetworks S_3–S_5 and S_8–S_{11} filter the signals sent by changes in concentration of I_1 and I_2, examination of the Fourier transforms (not shown) of the time-series for each set of species shows that the slower kinetics of S_8–S_{11} leads to a much stronger exponential decay of the components in their frequency spectra than those of S_3–S_5. The correlation (which is related to the product of the Fourier transforms of two time-series) of I_1 with S_8 is, therefore, much less than with S_4. The fact that both S_3–S_5 and S_8–S_{11} filter the signals (albeit to different extents) implies that subsystem 2 is better correlated with S_3–S_5 than with I_1 and I_2. This also explains why I_1 and I_2 appear on the far side of subsystem 1 with respect to subsystem 2. It may be possible to correct for the filtering effects by weighting the Fourier transforms of the series by an appropriate exponential factor. This may allow the better decomposition of the correlation into its direct, indirect, and spurious components. It must be remembered, however, that CMC relies on the fact that each chemical mechanism filters input signals in a characteristic fashion, resulting in an identifiable MDS diagram. This point is emphasized again in the next example.

The formulation of many reaction mechanisms is based on two simplifying possibilities:[35] (1) The reaction mechanism has one rate-determining step. (2) There is no rate-determining step, but the concentrations of intermediates are (nearly) constant (stationary-state hypothesis). Figure 10 shows a simple unbranched chain of chemical reaction steps. Such conversions are common structures in metabolic pathways and, thus, provide an interesting case study for CMC. In order to demonstrate the effects of different patterns of rate coefficients on the outcome of a CMC analysis, we break the linear network into three sets of reactions. Those governed by the first three rate coefficients, k_2–k_4, are the first group; k_5 defines the second

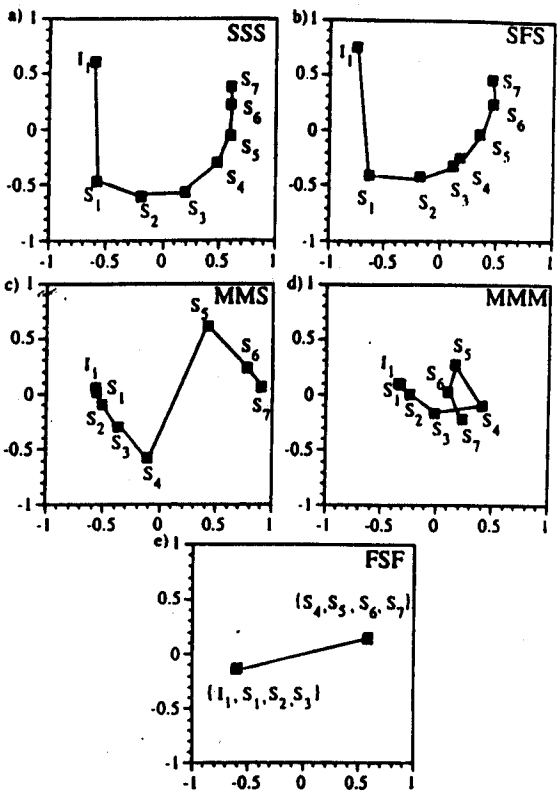

Figure 11. Two-dimensional projection of the MDS solutions of the five kinetic schemes for the linear mechanism in Figure 9. The three-letter triad in the upper right corner of each diagram indicates the pattern of rate constant for the first, second, and third parts of the network, respectively. See the caption of Figure 10 and the text for more information.

group; and the last three rate coefficients define the third. Each group of coefficients may be defined to be slow (0.7 s^{-1}), medium (70 s^{-1}), or fast (7000 s^{-1}) with respect to the switching time of the input time-series in I_1 (switching frequency = 20 s^{-1}). Figure 11 shows MDS diagrams resulting from CMC analysis of the linear networks with five different sets of rate coefficients. In all cases time-series were collected every 0.05 s for 100 s. The average concentration of the input was one unit, and the standard deviation was 0.1 units. Figure 11a results from a scheme in which all three groups of rate constants are chosen to be slow (no rate-determining step). When all the rate constants in a set of consecutive reactions are comparable, the stationary-state hypothesis is often employed to solve analytically the kinetic equations. The approximate autocorrelation time for the time-series, defined as the time lag at which the correlation of a series with itself decays to zero, is about 2.5 s (50 lag times, i.e., 50 observations). The system is not very close to steady state in this case, but a reasonably linear MDS diagram is produced nonetheless. There are three points to be made: (1) The large distance between the input and S_1 occurs because fluctuations in the input species occur too quickly for S_1 to follow. Thus, the filtering effect described above leads to a larger distance than expected. (2) This distance between adjacent species decreases with the number of steps away from the input. (3) The diagram is a curve instead of a line. These last two points are related. The distance between adjacent species decreases because of filtering. As the signal propagates through the network, the high-frequency components get successively filtered out until only the frequencies slower than the characteristic relaxation times of the network survive. As filtering becomes more severe in the network, subsequent species become more strongly correlated (since they are able to follow

one another exactly). This same effect is partly responsible for the curvature of the line. Since later species become more correlated, their correlation with earlier species also becomes more similar. That is, S_6 and S_7 are highly correlated, which, in this case, implies that they are correlated at nearly the same (relatively low) level with S_1. Therefore, the point representing S_7 must be placed at approximately the same distance from S_1 as S_6. Further, since lower correlations are measured less accurately than high correlations (see below), the difference between r_{1,S_6} and r_{1,S_7} becomes less significant (not as accurately measured). This leads to the so-called horseshoe effect often seen in the MDS of serial data.[18]

Figure 11b shows the case in which the middle step is fast compared to the other reaction steps. Since the rate of conversion of S_4 to S_5 is large, we expect the distance between these two species to be small. However, in the diagram it is the distance between S_3 and S_4 which is markedly decreased. Since S_4 is immediately converted to S_5, which itself is slowly degraded, the autocorrelation time of S_5 increases. This implies that its correlation with S_4 (the autocorrelation of which relaxes immediately) is weakened. However, since the back reactions are small, the (relatively) small concentration of S_4 follows S_3 nearly perfectly.

Figure 11c shows the case in which the reactions in the last part of the network are slow. Here, the distance between S_4 and S_5 is large, as expected. Since this distance is largely a result of a severe low-pass filtering of the signal sent by S_4, the correlation of S_4 with all species in the last part of the network is relatively low. Thus, these last species are almost equidistant to S_4 even though the distances among them are significant. This results in the observed displacement of the S_5-S_7 away from the rest of the network.

Figure 11d is the most pathological of the diagrams. It shows a loop in the structure of the network. The rate constants here are all medium. The autocorrelation time in this network is on the order of the fluctuation time of the input (0.05 s). With these rates, a signal initiated at the inputs travels approximately half-way down the network by the time of the next input concentration. This implies, for example, that at zero time lag S_1 correlates highly with the input and species S_2 and S_3 but not S_5-S_7. S_1 correlates with these latter species at one lag time. Because S_4 is observed just as it is responding to the current signal but still relaxing from response to the last signal, its correlation with S_1 at both lags is smaller than expected. Since the correlation metric does not take into account the time at which the maximum correlation of each species with S_1 occurs, S_5 appears "closer" to S_1 than does S_4. This leads to the observed loop in the diagram. If the metric is modified to take into account only the zero lag correlation, then the expected linear diagram is obtained (data not shown). This example shows how the characteristics of the input noise can strongly effect the outcome of a CMC experiment in some regimes. However, such pathologies can often be diagnosed by direct examination of the correlation surface.

Finally, Figure 11e shows the case in which all reactions are fast except the one in the second group which is slow. This corresponds to the case of a network with a rate-limiting step. Here, the autocorrelation times in the network are much faster than the input fluctuations, so all species in the first and last section of the network are effectively at steady state and therefore perfectly correlated. The slow step in the middle, however, causes filtering of the signal and thereby causes a decrease in correlation between the two sections of the network. This results in the dumbbell shape of the network.

IV. Discussion

With the application and development of concepts of electronic circuit and systems theory, and techniques from multivariate statistics, we obtain a MDS correlation-metric diagram which defines and represents the connectivity and the strength of kinetic interactions among the species of a reaction network from a time-series determination of chemical concentrations. In addition, we obtain the startling result that, for the systems studied, the MDS diagram recapitulates many features of the reaction mechanism. We believe these approaches to be promising and to warrant further study. A number of caveats, improvements, and new directions are apparent.

Correlation metric construction is a statistical analysis, hence, care must be exercised to take into account the significance of all the calculated correlations. If we ignore any error in the concentration measurements themselves, the probability that $|r_{ij}(\tau)|$ is larger than it should be in the null hypothesis (which is that species i and j are uncorrelated) is $\mathrm{erfc}((r_{ij}(\tau)^2 n/2)^{1/2})$,[19] which is a rapidly decreasing function of n, the number of observations. Smaller correlations are, therefore, less significant for a given n than large correlations, a fact which is currently ignored by our connection algorithm. This may lead to the so-called horseshoe effect often seen when MDS is used to seriate data.[18] Experiments might be continued until none of the correlations are changing by more than a given percentage. However, we must be aware that for a given set of parameters defining the random perturbations on the inputs and for a finite number of measurements the calculated correlations are only valid for the portion of the state-space of the system observed. The state-space of a reaction network is defined here by the range of all possible sequences of input concentrations parametrized by the initial values of the concentrations of all the other species composing the network.

Once the correlations are properly measured, the rest of the analysis proceeds simply, except for the calculation of the connection diagram. The connections among points in the MDS solution determine which species participate in each reaction step or interaction (for example, one species might be an effector for the enzymatic production of another species). The connection algorithm described in step 3 is primitive in that (1) causality cannot be hypothesized since even the value of τ at the maximum correlation between two species is ignored (and see discussion above) and (2) the minimum correlation sufficient for connection between two species is chosen merely on the assumption that all species are connected to at least one other species. This algorithm implies a somewhat arbitrary definition of significance of correlation between a single variable and one or more other variables. It may be possible that transinformation,[30] partial correlation analysis, and multiple regression or path analysis[17,18,27,28,30] can be used to discover high-order relations not initially uncovered by the relatively simple correlation analysis. One such example of a high-order relation is the combined effect of I_1 and I_2 on S_3 in Figure 1. Transinformation, multiple regression, and LISREL also provide measures of significance via the χ^2 statistic and provide estimates of the percentage of the variance of one variable explainable by the variance of a subset of the others.[16,18,20,27,28,36] These methods may, therefore, lead to a better definition of the connection among species as well as *provide an indication of the existence of unmeasured variables* in the case when the variation in a measured species cannot be explained by variations in itself, the other measured species, and the inputs. Suppose, for example, we can measure three species in a network, I_1, A, and B. Suppose further that I_1 is a direct precursor of both A and B but that it must react with an unknown (or not directly

measurable) species U in order to produce B. Assuming the concentration of U fluctuates significantly during the time course of the experiment, part of the fluctuation of [B] is due to these changes in [U]. It is, then, impossible to explain fully the variance in B using the two measured variables I_1 and A and prior values of B, and as a result, a significant residual, e_B, will be found when B is regressed on these known variables. A large residual between the measured variance of B and the amount of variance of B explained by the known variables thus indicates the existence of unmeasured species.

The MDS correlation diagram is a new representation of a reaction network. Aside from reproducing in large part the mechanistic diagram for a chemical reaction network, there is information present regarding tightness of kinetic coupling among the species. For example, species S_6 and S_7 from the mechanism in Figure 1 are drawn as far apart as species S_3 and S_4. However, in the diagrams plotted in Figure 5, S_6 and S_7 are plotted on top of one another since they are completely correlated. Thus, the MDS diagram is distorted to reveal loci of tight regulation; for example, on the MDS diagram the point representing S_3 is close to the pair S_6 and S_7 and is the species whose concentration most determines their concentration. The hierarchical clustering results, then, are largely a recapitulation of what is evident by eye from the MDS diagram. There is a related point: in the example derived from Figure 1, the concentrations of the input species are within the region of saturation of their target enzymes (concentrations much greater than K_I) most of the time; however, these concentrations may take on values near or below the inhibition constant, and thus transitions between low and high concentrations of S_3 are observed. If we had chosen the parameters of the Gaussian driving noise such that the input concentrations were always much greater than the inhibition constants, then such a transition would never be observed and the correlation between S_3 and the pair S_6 and S_7 would have been much reduced. Alternatively, if the average values of the input concentrations had been in the transition region of enzymes (i.e., close to the K_I's), the response of the species to variation of the inputs would be highly nonlinear. In this case, linear correlation coefficients would underestimate the relationship between variables. In general, Spearman nonparametric rank correlation[19,20] should be used since it requires only monotonic (as opposed to linear) relationships among variables.

Besides the average concentration and standard deviation, the other important parameter of the driving noise on the inputs is the frequency spectrum of the external perturbations. If the concentrations of the inputs switch at a rate comparable to the slowest relaxation times in the system, then the network will attenuate the high-frequency components in the input sequence. In our analysis we chose the time between changes of input concentrations such that the system was driven near its steady state. Figure 4B shows that the species have completely relaxed in less than 10 time lags (100 s) after perturbation. This implies that if the variation of the input concentrations is much faster than 100 s (say on the order of 1 s), the network will not have time to respond and the correlation strength between system components and the inputs will be lost. This effect is a mixed blessing for the current analysis since it leads to an artificially high correlation between species at loci remote from the entry points of the inputs; the input signal may be filtered similarly at disparate but equally remote sites. The resultant correlations tend to group highly filtered species together far from those species more directly connected to the inputs. (See section III.) This is a desirable feature except that the distances between the highly filtered species will be smaller than expected even when such species do not interact. Further, as shown in the analysis of the mechanism in Figure 7, filtering may result in the spurious assignment of connection between two species. Clearly, the final MDS diagram is partially dependent on the parameters of the noise imposed on the system. It is possible that frequency domain approaches to time-series analysis[29,37] may help in a study of the role of frequency transfer functions[37] in the control of chemical networks.

We have assumed that all species involved in the mechanism may be identified and measured. For systems with many species this has not been possible. When there are missing species, CMC may still be performed on the measurable subset of species. The effects of the other species are subsumed into the correlations among the known species, and a consistent diagram can be constructed. The MDS diagram, then, may not be an obvious representation of the underlying mechanism. In fact, due to signal loss in the network, certain connections between known species may be lost. On the other hand, unknown species that participate in the dynamics of the network, but are not ultimately determined by the input concentrations (i.e., unkown species produced and degraded by uncontrolled sources in the network), can be helpful in strengthening the correlations among known species by providing another source of information (chemical material) to be processed by the network. For oscillatory reaction mechanisms it is possible to construct the Jacobian of the reaction network by applying a delayed feedback in each of the species in the mechanism and then measuring the response of only one species.[3] Further, reference 3 shows how to deduce a reaction of mechanism from the Jacobian of the network.

Finally, the networks analyzed in this paper are combinational networks, i.e., networks with no (explicit) feedback loops and, therefore, no memory or autonomous dynamics. Nonzero correlations away from $\tau = 0$ are, therefore, caused only by slow relaxation of the chemical species to their steady states (slow reaction steps). In sequential systems, in which feedback exists, nonzero time-lagged correlations may be indicative of species involved in a feedback relation. For systems that contain feedback in such a way as to generate multistability and oscillations it may be impossible, in the absence of any prior knowledge, to predict in advance how many states are available to the network and how they are triggered. However, a series of experiments has been suggested for such systems from which the essentials of the core mechanism containing feedback may be deduced.[1-3,38] The methods discussed here may be useful complementary approaches to determining reaction mechanisms of coupled kinetic systems.

Acknowledgment. This work has been supported in part by the National Science Foundation and the Air Force Office of Scientific Research. A.P.A. was also supported, in part, by the National Institute of Mental Health, Grant No. MH45324.

References and Notes

(1) Eiswirth, M.; Freund, A.; Ross, J. *J. Phys. Chem.* **1991**, *95*, 1294.
(2) Eiswirth, M.; Freund, A.; Ross, J. *Adv. Chem. Phys.* **1991**, *LXXX*, 127.
(3) Chevalier, T.; Schreiber, I.; Ross, J. *J. Phys. Chem.* **1993**, *97*, 6776.
(4) Stemwedel, J. D.; Schreiber, I.; Ross, J. *Adv. Chem. Phys.*, accepted.
(5) Stemwedel, J. D.; Ross, J. *J. Phys. Chem.*, submitted.
(6) Hung, Y.-F.; Ross, J. *J. Phys. Chem.*, submitted.
(7) Hjelmfelt, A.; Weinberger, E. D.; Ross, J. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 10983.
(8) Hjelmfelt, A.; Ross, J. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 388.
(9) Hjelmfelt, A.; Weinberger, E. D.; Ross, J. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 383.
(10) Hjelmfelt, A.; Ross, J. *J. Phys. Chem.* **1993**, *97*, 7988.
(11) Hjelmfelt, A.; Schneider, F. W.; Ross, J. *Science* **1993**, *260*, 335.
(12) Hjelmfelt, A.; Ross, J. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *97*, 63.
(13) Arkin, A. P.; Ross, J. *Biophys. J.* **1994**, *67*, 560

(14) Klir, G. J. *Int. J. Gen. Sys.* **1981**, *7*, 1.
(15) Conant, R. C. *Int. J. Gen. Sys.* **1988**, *14*, 125.
(16) Conant, R. C. *Int. J. Gen. Sys.* **1988**, *14*, 97.
(17) Marriot, F. H. C. *The Interpretation of Multiple Observations*; Academic Press: New York, 1974.
(18) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: San Francisco, 1979.
(19) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C*; Cambridge University Press: New York, 1988.
(20) Ott, L. *An Introduction to Statistical Methods and Data Analysis*; Duxbury Press: Boston, 1984.
(21) Shepard, R. N. *Science* **1980**, *210*, 390.
(22) Swofford, D. L.; Olsen, G. J. In *Molecular Systematics*; Hillis, D. M., Moritz, C., Eds.; Sinauer Associates, Inc.: Sunderland, MA, 1990; p 411.
(23) Schacter, E.; Chock, P. B.; Stadtman, E. R. *J. Biol. Chem.* **1984**, *259*, 12252.
(24) Schacter, E.; Chock, P. B.; Stadtman, E. R. *J. Biol. Chem.* **1984**, *259*, 12260.
(25) All algorithms and simulations are written in the C programming language interfaced, when necessary, to the Numerical Algorithms Group FORTRAN library (The Numerical Algorithms Group Limited, Oxford, UK. Release Mark 15.) Programs were compiled and executed on DEC 3100 workstations.
(26) Causality, in this case the assignment of which reactants lead to which products, cannot be determined simply by the strength of a statistical relation. For causality to be established three criteria must be met: there must be (1) a temporal ordering of variables; (2) concomitant variation; and (3) control over other factors which might affect the observed relationships between variables.[32] In this analysis we do not always meet the first criterion, and the third criterion can only be fulfilled by data outside the present analysis (for example knowledge of the basic chemistry of the species might preclude or indicate basic causal relations). This point is especially relevant to the connection algorithm discussed in step 3. Again, for reasons of simplicity in this first article we defer the study of causality.
(27) Li, C. C. *Path Analysis—A Primer*; The Boxwood Press: Pacific Grove, CA, 1986.
(28) Hayduk, L. A. *Structural Equation Modelling with LISREL: Essentials and Advances*; John Hopkins University Press: Baltimore, 1987.
(29) Box, G. E. P.; Jenkins, G. M. *Time Series Analysis, Forecasting and Control*; Holden-Day: San Francisco, 1976.
(30) Broekstra, G. *Int. J. Gen. Sys.* **1981**, *7*, 33.
(31) Klir, G. J. *Architecture of Systems Problem Solving*; Plenum Press: New York, 1985.
(32) Dillon, W. R.; Goldstein, M. *Multivariate Analysis: Methods and Applications*; John Wiley and Sons: New York, 1984.
(33) The neglect of the value of τ at the maximum correlation between two time-series represents a loss of useful information. In kinetic systems in which the correlation is peaked at some nonzero lag the implication is that the system is large enough for there to be a delay time between the site of signal generation (the inputs) and reception. This information may be incorporated into the distance metric in order to better represent distances among the species. In addition the sign of the lag may be used to estimate a sequence of events.
(34) Corana, A.; Marchesi, M.; Martini, C.; Ridella, S. *ACM Transactions on Mathematical Software* **1987**, *13*, 262.
(35) Berry, R. S.; Rice, S. A.; Ross, J. *Physical Chemistry Part 3: Physical and Chemical Kinetics*; John Wiley & Sons: New York, 1980.
(36) Conant, R. *Int. J. Gen. Sys.* **1981**, *7*, 81.
(37) Otnes, R. K.; Enochson, L. *Applied Time Series Analysis: Basic Techniques*; John Wiley and Sons: New York, 1978.
(38) Strasser, P.; Stemwedel, J. D.; Ross, J. *J. Phys. Chem.* **1993**, *97*, 2851.

JP942438P