The rational parameterisation theorem for multisite post-translational modification systems

Matthew Thomson [a], Jeremy Gunawardena [b,*]

[a] Biophysics Program, Harvard University, Cambridge, MA 02138, USA
[b] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

ARTICLE INFO

ABSTRACT

Post-translational modification of proteins plays a central role in cellular regulation but its study has been hampered by the exponential increase in substrate modification forms ("modforms") with increasing numbers of sites. We consider here biochemical networks arising from post-translational modification under mass-action kinetics, allowing for multiple substrates, having different types of modification (phosphorylation, methylation, acetylation, etc.) on multiple sites, acted upon by multiple forward and reverse enzymes (in total number L), using general enzymatic mechanisms. These assumptions are substantially more general than in previous studies. We show that the steady-state modform concentrations constitute an algebraic variety that can be parameterised by rational functions of the L free enzyme concentrations, with coefficients which are rational functions of the rate constants. The parameterisation allows steady states to be calculated by solving L algebraic equations, a dramatic reduction compared to simulating an exponentially large number of differential equations. This complexity collapse enables analysis in contexts that were previously intractable and leads to biological predictions that we review. Our results lay a foundation for the systems biology of post-translational modification and suggest deeper connections between biochemical networks and algebraic geometry.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Post-translational modification (PTM) of proteins is a central regulatory mechanism in eukaryotic cells (Walsh, 2006). Although phosphorylation was the first modification to be discovered (Fischer, 1997; Krebs, 1997), and remains the best studied, proteins are subject to other types of modification by covalent attachment of molecules to the side chains of amino acid residues. The modifiers include other small molecules, as in methylation, acetylation and sulfation, complex sugars, as in glycosylation, and small protein moieties, as in ubiquitin-like modification (Walsh, 2006). It has become increasingly clear that these modifications work together to orchestrate cellular function (Hunter, 2007).

PTMs are dynamically maintained. Forward enzymes, which transfer modifiers from donor molecules to specific residues, are usually competing against reverse enzymes, which hydrolyse modified residues, detaching the modifier and returning it to the pool from which donor molecules are synthesised. In the case of phosphorylation, these enzymes are protein kinases and phosphoprotein phosphatases, respectively. These so-called "futile cycles" of modification and demodification can use energy to keep the concentrations of modified substrates far from equilibrium, testifying to the importance of such processes as regulatory mechanisms.

A given protein may be modified on multiple sites. For instance, the transcription factor and tumour suppressor p53 has 18 serine/threonine sites that can be phosphorylated and 10 lysine sites that can accommodate acetylation, methylation and attachment of ubiquitin, SUMO and NEDD (Kruse and Gu, 2008). O-linked modifications like phosphorylation, on residues like serine, threonine or tyrosine, are digital—at most one phosphate group is attached to each residue—but N-linked modifications like methylation, on residues like lysine or arginine, can be more complex (Walsh, 2006). Ubiquitin, in particular, can form linear and branched poly-ubiquitin chains connected by iso-peptide linkages through lysine residues. In general, an individual molecule with n sites of modification may be in one of several global states of modification ("modforms"), whose numbers increase exponentially with n. Cartoon diagrams often pick one of these modforms, usually the maximally modified one, to depict the state of a protein in the cell. However, there is always a population of such molecules present and individual molecules in the population may be in different states. The state of the population is best described as a frequency distribution over these single molecule states. We call this the modform distribution. Not all modforms will necessarily be present at any one time but this

begs the question of which modforms are present and of how the relevant forward and reverse enzymes cooperate to shape the modform distribution. The methods of this paper were developed in part to address such questions.

The combinatorial explosion in the number of potential modforms is a challenge for both experiment and theory. Mass spectrometry techniques have only recently begun to provide data on modform distributions (Pesavento et al., 2008; Phanstiel et al., 2008), prompted, in part, by the growing realisation that different modforms may have distinct biological effects (Park et al., 2006; Pufall et al., 2005; Wu et al., 2004). The theoretical challenge lies in the complexity of any mathematical model, which must accommodate some number, L, of forward and reverse enzymes and some total number, N, of modforms targeted by those enzymes. Additionally, the biochemistry of modification and demodification usually requires intermediate enzyme–substrate complexes associated to each enzyme and its substrates. If there are P such intermediate complexes, then the model will have $L+N+P$ state variables, where N and P grow exponentially with n, while L is relatively much smaller:

$$L \ll N, \quad P \propto a^n \quad \text{for some } a \geq 2.$$

Because the dynamical equations are nonlinear, these models cannot be analytically solved for the temporal trajectories of the variables. Simulation provides the widely used alternative to analytical solution. The trajectories of the system can be calculated by numerical integration once the site-specific rate constants have been given values. These values have usually not been measured and it may be necessary to search through the space of rate constants to determine whether a particular behaviour is robust to the choice of rate constant values (Gunawardena, 2009). Calculations of this kind rapidly become infeasible with increasing n, which has limited simulation studies to systems with few sites. These difficulties have made it hard to see what, if any, general principles lie behind the widespread use of multisite modification systems in cellular regulation.

In this paper we show that, if attention is restricted to the steady states of a multisite PTM system, then it is not necessary to numerically integrate $L+N+P$ differential equations but only to solve L algebraic equations (Theorem 3). Furthermore, this reduction can be carried out without having to specify the rate constant values in advance. For instance, for the case of two enzymes, the steady states can be found by an analogue of the nullcline analysis that is widely used for two-dimensional systems of differential equations (Strogatz, 2001): the steady states correspond to the intersections of two curves in the plane, no matter what the number of sites. This exponential reduction in complexity leads to biological insights in contexts that were previously intractable, as reviewed in Section 4, and provides a new theoretical foundation for studying multisite PTM systems.

The present paper generalises and conceptually simplifies a method that emerged in previous work for systems with two enzymes and one substrate (Manrai and Gunawardena, 2008; Thomson and Gunawardena, 2009). It allows for multiple enzymes, multiple types of modification, multiple substrates and complex biochemistry of modification and demodification. These assumptions are substantially more general than any used previously in the literature (Ferrell and Bhatt, 1997; Goldbeter and Koshland, 1981; Gunawardena, 2005; Kim and Ferrell, 2007; Lisman, 1985; Markevich et al., 2004; Salazar and Höfer, 2007; Shacter-Noiman et al., 1983). The main restrictions in terms of applicability of our methods are the requirements that enzymes cannot also be substrates and that the recharging mechanism for each modification should keep the concentration of donor molecules constant on the time scale of steady state formation. The latter requirement is widely believed to hold for phosphor-

ylation, in which the donor molecule is ATP, and it has been taken for granted in all mathematical models of phosphorylation, but its validity for other forms of modification appears not to have been widely studied.

The two most significant examples that currently fall outside the scope of our analysis are kinase cascades and ubiquitin-like modifications. Both violate the first requirement by relying on chains of enzymatic modification. We discuss how our results might be extended to such cases in Section 4. It is an interesting question whether ubiquitin-like modifications satisfy the second requirement. Peptide modifiers are synthesised by mRNA translation rather than by the cell's central metabolism, as is the case for small molecule modifiers, and little is known about how effectively this translation process is buffered against varying demand.

Our method of proof is one of hierarchical elimination of variables from the steady-state equations, which, because of mass-action, are polynomials in the variables. The intermediate enzyme–substrate complexes are first eliminated in favour of the substrates and the enzymes (Proposition 1). Using the results of this, the substrates are then eliminated in favour of the enzymes (Proposition 2). Each elimination step is framed as the solution to a system of linear equations (Lemma 2) but is undertaken over an extension field of the real numbers, which carries the nonlinearity that is present in the underlying steady-state equations. This allows us to solve a nonlinear system of equations using essentially linear methods. The extension field also permits the rate constants to be treated symbolically during the elimination. The Matrix-Tree theorem (Theorem 1) plays a key role in identifying the nonlinear coefficients that arise in the elimination steps. This important graph-theoretic result goes back to the 19th century but was first stated in the form we use it in Tutte (1948). It does not seem to have been previously noticed that the Matrix-Tree theorem immediately implies the famous King–Altman method, developed in King and Altman (1956), for calculating the rate function of an enzyme (Cornish-Bowden, 1995). We find it intriguing that the method we use to analyse multisite PTM systems is so closely related to a key result of classical biochemistry. Section 2 reviews the graph theoretic preliminaries, including the Matrix-Tree theorem, before the main results are presented in Section 3.

Our results are not derivable from any existing mathematical theories of biochemical reaction networks, such as Chemical Reaction Network Theory (CRNT) (Feinberg, 1979; Gunawardena, 2003), more recently developed injectivity methods (Craciun et al., 2006; Soulé, 2003) or Monotone Systems Theory (Angeli et al., 2004). We believe that they are best interpreted in algebraic geometric terms, as discussed at more length in an earlier paper (Manrai and Gunawardena, 2008). Under the mass-action kinetics used here, a network of biochemical reactions gives rise to a polynomial dynamical system. Hence, for given rate constant values, the steady states form a real algebraic variety (Cox et al., 1997). Despite this, the use of algebraic geometric methods to study biochemical networks has been surprisingly limited, the most interesting exception being the use of toric varieties to reinterpret the Deficiency Zero theorem of CRNT (Craciun et al., 2009; Gatermann and Huber, 2002). For multisite PTM systems, we show that the variety of steady-state modform concentrations can be parameterised by L rational functions (Theorem 4). A rational parameterisation provides an explicit description of points on a variety, in contrast to their implicit definition as solutions of polynomial equations. Rationally parameterisable varieties are rare and of considerable interest in their own right. The rationality of multisite PTM systems suggests that algebraic geometry may provide powerful tools for analysing biochemical reaction networks and overcoming molecular complexity. We may

hope, thereby, to better see the biological wood for the molecular trees.

2. Preliminaries

2.1. Symbols and polynomials

In this paper we will analyse systems of ordinary differential equations arising from networks of biochemical reactions under mass-action kinetics. The rate constants and dynamical variables are usually treated as real variables in $\mathbb{R}$. In our analysis, we will make use of certain directed graphs whose edges have labels like aX, where a is a rate constant and X is a dynamical variable in steady state. These labels must satisfy a positivity condition, expressed in Lemma 1 below. The rate constants can reasonably be taken to be positive but whether or not a dynamical variable is positive in steady state is more delicate. Even if a system is started with all its variables positive, it may not be persistent within the positive orthant and may reach a steady state on the boundary in which some variables are zero. To avoid having to rule out such situations and thereby limit our analysis, we take a more algebraic approach. We treat the rate constants and the dynamical variables in steady state as uninterpreted symbols in an appropriate extension field of $\mathbb{R}$. We show that the calculations can be carried out over this extension field and, having done them, we then give real values to the symbols and draw conclusions over $\mathbb{R}$. While this avoids the problem and brings added benefits, it incurs some technical cost. The reader will not lose much by assuming that all calculations take place over $\mathbb{R}$ and ignoring the problems that arise with loss of positivity in the dynamical variables. The symbolic calculations only appear in Sections 3.3 and 3.4; rate constants and dynamical variables are treated as real variables elsewhere.

For more information about the algebraic methods reviewed here see, for instance, Herstein (1975). For any finite set $Q = \{q_1, \ldots, q_n\}$, $\mathbb{R}[Q]$ will denote the ring of real polynomials in the elements of Q, considered as algebraically independent symbols. Recall that a polynomial $p \in \mathbb{R}[Q]$ is a linear combination of monomials, $p = \sum_\alpha c_\alpha q^\alpha$, where $c_\alpha \in \mathbb{R}$ and each monomial q^α is a product of symbols,

$$q^\alpha = q_1^{\alpha_1} \cdots q_n^{\alpha_n} \quad \text{with } \alpha_i \geq 0.$$

$\mathbb{R}(Q)$ will denote the field of rational functions in the elements of Q: the smallest field in which the symbols in Q can be added, subtracted, multiplied and divided as if they were (non-zero) numbers. Equivalently, $\mathbb{R}(Q)$ is the field of fractions of $\mathbb{R}[Q]$: each element $f \in \mathbb{R}(Q)$ can be expressed as the ratio of two polynomials, $f = p_1/p_2$ where $p_1, p_2 \in \mathbb{R}[Q]$. $\mathbb{R}[Q]$ sits inside $\mathbb{R}(Q)$ in the obvious way: $p = p/1$.

For various finite sets Q, we will use elements of $\mathbb{R}[Q]$ as labels in graphs, and use $\mathbb{R}(Q)$ as a field over which we solve linear equations. For instance, Q may consist of rate constants. To relate calculations over $\mathbb{R}[Q]$ or $\mathbb{R}(Q)$ back to biology, the symbols ultimately have to be given real values. This is not a problem for the polynomial ring $\mathbb{R}[Q]$. Any assignment, $\iota : Q \to \mathbb{R}$, extends to a homomorphism of rings $\iota : \mathbb{R}[Q] \to \mathbb{R}$. Hence, any symbolic algebraic expression in $\mathbb{R}[Q]$ gives rise to a corresponding expression in $\mathbb{R}$. However, the field of rational functions, $\mathbb{R}(Q)$, contains elements like $1/(\iota(q_1) - q_1)$, which become undefined in $\mathbb{R}$ no matter what assignment of real values are made to elements of Q. To infer an expression over $\mathbb{R}$, it is essential to show that nothing blows up in this way.

We make use of S-positivity to do this. A polynomial $p \in \mathbb{R}[Q]$ is said to be S-positive ("sum positive") if it is a non-zero sum of positive monomials. That is, p is S-positive if $p = \sum_\alpha c_\alpha q^\alpha \neq 0$ and

if $c_\alpha > 0$ whenever $c_\alpha \neq 0$. A rational function in $\mathbb{R}(Q)$ is said to be S-positive if it can be expressed as the ratio of two S-positive polynomials. If the elements of Q are given positive real values, which is biochemically realistic for rate constants, then any S-positive rational function in $\mathbb{R}(Q)$ will be well defined over $\mathbb{R}$. Note that, if $Q = \emptyset$, so that $\mathbb{R}[Q] = \mathbb{R}$, then $x \in \mathbb{R}[Q]$ is S-positive if, and only if, it is positive in the usual sense.

If $p = \sum_\alpha c_\alpha x^\alpha \in \mathbb{R}[Q]$ then $p = 0$ means that $c_\alpha = 0$ for all α. If $\iota : Q \to \mathbb{R}$, then $p = 0$ in $\mathbb{R}[Q]$ implies that $\iota(p) = 0 \in \mathbb{R}$. The converse, of course, is false: if $\iota(p) = 0 \in \mathbb{R}$ it does not imply that $p = 0 \in \mathbb{R}[Q]$. However, if $p \neq 0$ then the variety in $\mathbb{R}^n$ corresponding to the set of solutions of $p = 0$ has dimension strictly less than n. Hence, if $\iota(p) = 0$ for sufficiently many assignments ι, then $p = 0$. Let $\mathbb{R}^+$ denote the positive reals.

Remark 1. If $\iota(p) = 0$ for all $\iota : Q \to \mathbb{R}^+$, then $p = 0 \in \mathbb{R}[Q]$.

2.2. Graphs and Laplacians

A labelled, directed graph is a triple, (V, E, ℓ), where V is a finite set of nodes, $V = \{v_1, \ldots, v_n\}$, E is a finite set of directed edges, $E \subseteq V \times V$, and $\ell : E \to \mathbb{K} - \{0\}$ is a function that associates to each edge a non-zero label in some field $\mathbb{K}$. Because they take values in a field, labels can be treated algebraically as if they were numbers. Usually, $\mathbb{K} = \mathbb{R}(Q)$ for some set of symbols Q and the labels are elements of $\mathbb{R}[Q]$. In place of a labelling function, the notation $v_i \xrightarrow{a} v_j$ will denote an edge from v_i to v_j with label a. We will sometimes abbreviate this to $v_i \to v_j$.

If G is a labelled, directed graph, $G^\star$ will denote the corresponding unlabelled, undirected graph, in which the direction of the edges is forgotten and multiple edges between the same vertices are merged. We use $v_i \leftrightarrow v_j$ to denote an edge in $G^\star$. Note that this could imply $v_i \to v_j$ or $v_j \to v_i$ or both in G. We say that G is connected if $G^\star$ is connected: if there is a chain of undirected edges linking any two vertices. Any labelled, directed graph is a disjoint union of connected components. G is strongly connected if there is a directed path between any two distinct vertices.

There is a bijective correspondence between matrices and labelled, directed graphs. If A is a $n \times n$ matrix over $\mathbb{K}$, let $\mathcal{G}(A)$ be the associated labelled, directed graph with nodes $\{1, \ldots, n\}$ and labelled edges $j \xrightarrow{A_{ij}} i$, if, and only if, $A_{ij} \neq 0$. Note that entry A_{ij} goes from j to i. If G is a labelled, directed graph on $\{1, \ldots, n\}$, let $\mathcal{M}(G)$ be the $n \times n$ matrix for which $\mathcal{M}(G)_{ij} = a$ if, and only if, $j \xrightarrow{a} i$. Evidently, $\mathcal{M}(\mathcal{G}(A)) = A$ and $\mathcal{G}(\mathcal{M}(G)) = G$.

If G is a labelled directed graph then its Laplacian matrix, $\mathcal{L}(G)$, is given by $\mathcal{L}(G) = \mathcal{M}(G) - \mathrm{diag}(1.\mathcal{M}(G))$. Here, 1 denotes the all 1's row vector of the appropriate dimension, $1 = (1, \ldots, 1)$, and $\mathrm{diag}(v)$, where v is a row vector, denotes the diagonal matrix with v on the main diagonal

$$\mathrm{diag}(v)_{ij} = \begin{cases} v_i & \text{if } i = j, \\ 0 & \text{otherwise}. \end{cases}$$

The Laplacian encodes much information about graph structure; see, for instance, Chung (1997), but note that the conventions used here are different. Note also that, by construction, $1.\mathcal{L}(G) = 0$.

Lemma 1. *Let G be a labelled, directed graph on n vertices with no self-loops in which each non-zero label is a S-positive element of $\mathbb{R}[Q]$. If G is strongly connected then the rank of $\mathcal{L}(G)$ over $\mathbb{R}(Q)$ is $n - 1$.*

Proof. Let $u = (u_1, \ldots, u_n) \in \mathbb{R}(Q)^n$. Since $1.\mathcal{L}(G) = 0$, it is sufficient to show that if $u.\mathcal{L}(G) = 0$ then $u_1 = \cdots = u_n \in \mathbb{R}(Q)$. Express each $u_i \in \mathbb{R}(Q)$ as a fraction $u_i = f_i/g_i$, where $f_i, g_i \in \mathbb{R}[Q]$ and we may suppose that $g_i \neq 0$. We can now clear the denominators. Let

$\lambda = g_1 g_2 \cdots g_n \neq 0$ and $v_i = \lambda u_i$. Then, $v_i \in \mathbb{R}[Q]$ and $v.\mathcal{L}(G) = 0$. We can now operate in the polynomial ring $\mathbb{R}[Q]$ in preference to the field of fractions $\mathbb{R}(Q)$. The entries in the Laplacian have the following form. If $i \neq j$, $\mathcal{L}(G)_{ij}$ is either zero or is a S-positive element of $\mathbb{R}[Q]$, while the diagonal entries are given by

$$\mathcal{L}(G)_{ii} = -\sum_{k \neq i} \mathcal{L}(G)_{ki}.$$

Hence, for the i-th column of $v.\mathcal{L}(G) = 0$,

$$\sum_{k \neq i} (v_k - v_i)\mathcal{L}(G)_{ki} = 0. \tag{1}$$

So far, this has all been symbolic, in $\mathbb{R}[Q]$. Now suppose that the symbols in Q are given positive real values and, suppressing the corresponding assignment $\iota : Q \to \mathbb{R}^+$ for readability, let us consider the corresponding system of column equations to (1) in $\mathbb{R}$. Since the v_i are now real numbers, there is a smallest, v_p, for which $v_i \geq v_p$ for all i. Let $U \subseteq \{1, \ldots, n\}$ be the set of those indices i for which $v_i = v_p$. $U \neq \emptyset$ since $p \in U$. If $m \in U$, then, in the m-th column equation,

$$\sum_{k \neq m} (v_k - v_m)\mathcal{L}(G)_{km} = 0,$$

each non-zero term is the product of a non-negative quantity, $v_k - v_m$, since v_m is smallest, and a strictly positive quantity $\mathcal{L}(G)_{km}$, since each label in G is S-positive. It follows that $v_k = v_m = v_p$ whenever there is an edge $m \to k$, so that U is closed under outgoing edges. Since G is strongly connected, $U = \{1, \ldots, n\}$. Hence, $v_i = v_p \in \mathbb{R}$ for all i. Since this holds for any assignment of positive values to symbols in Q we deduce from Remark 1 that $v_1 = \cdots = v_n$ symbolically in $\mathbb{R}[Q]$. Since $\lambda \neq 0$ we see that $u_1 = \cdots = u_n \in \mathbb{R}(Q)$, as required. $\square$

Some condition on the labels is necessary for the conclusion of Lemma 1. The following labelled, directed graph on $\{1, 2, 3\}$, with labels in $\mathbb{R}$, is strongly connected

$$1 \xrightarrow{1} 2, \quad 2 \xrightarrow{2} 1, \quad 1 \xrightarrow{1} 3, \quad 3 \xrightarrow{2} 1, \quad 2 \xrightarrow{-1} 3, \quad 3 \xrightarrow{-1} 2,$$

but its Laplacian has rank 1, not 2:

$$\begin{pmatrix} -2 & 2 & 2 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}.$$

Remark 2. If M is any $n \times n$ matrix, we can construct a labelled directed graph with no self-loops by ignoring the labels on the main diagonal: $G = \mathcal{G}(M - \text{diag}(M_{ii}))$. If, in addition, $1.M = 0$, then M is the Laplacian of G:

$$\mathcal{L}(G) = [M - \text{diag}(M_{ii})] - \text{diag}(1.[M - \text{diag}(M_{ii})]) = M - \text{diag}(M_{ii}) + \text{diag}(M_{ii}) = M.$$

2.3. The Matrix-Tree theorem

In what follows we will need to solve linear systems of the form $M.z = 0$, where M is a $n \times n$ matrix over some field $\mathbb{K}$, z is a column vector of n unknowns and M has rank $n - 1$. Recall that the adjugate matrix of M, $\text{adj}(M)$, is defined by

$$\text{adj}(M)_{ij} = (-1)^{i+j}M_{(ji)}, \tag{2}$$

where $M_{(ji)}$ is the maximal minor given by the determinant of the $(n-1) \times (n-1)$ matrix obtained from M by removing the jth row and ith column. Note the reversal of indices in (2). The adjugate satisfies the Laplace relations

$$\text{adj}(M).M = M.\text{adj}(M) = \det(M)I.$$

If $\text{rk}(M) = n - 1$, any column vector of $\text{adj}(M)$ is a basis for the column null space. Taking the first column, let

$$z_i = \text{adj}(M)_{i1} = (-1)^{i+1}M_{(1i)}, \tag{3}$$

so that $M.z = 0$.

The maximal minors have a particularly striking form when M is a Laplacian matrix. Let G be a labelled, directed graph with no self-loops. T is said to be a spanning tree of G if T is a directed subgraph which reaches each node of G such that $T^\star$ is connected and acyclic. T inherits labels from G. T is said to be rooted at $v \in G$ if v is the unique sink in T. That is, v is the only node of T with no edges leaving it, $v \nrightarrow w$. This implies that any non-root node has exactly one edge leaving it, for otherwise there would either be an additional sink or an undirected cycle. Let $\Theta_v(G)$ denote the set of spanning trees of G rooted at v.

Theorem 1 (*The Matrix-Tree theorem, Tutte, 1948, Section 3.6*). *Let G be a directed graph on $\{1, \ldots, n\}$ with labels in the field $\mathbb{K}$ and no self-loops. The maximal minors of the Laplacian are given by*

$$\mathcal{L}(G)_{(ij)} = (-1)^{n+i+j-1} \sum_{T \in \Theta_j(G)} \left(\prod_{k \xrightarrow{a} l \in T} a \right).$$

It is remarkable that the maximal minor is, up to a sign, just a sum of products of labels, since the determinant itself is a sum with alternating signs. Results like Theorem 1 go back to the 19th century work of Kirchhoff on electrical networks and of Sylvester and others on elimination theory; see Moon (1970, Chapter 5) for references. Combining (3) with Theorem 1, we get the following.

Lemma 2. *Let M be a $n \times n$ matrix over a field $\mathbb{K}$ such that $1.M = 0$ and $\text{rk}(M) = n - 1$. Let G be the labelled, directed graph constructed as in Remark 2 for which $M = \mathcal{L}(G)$. The one-dimensional column null space of M is generated by the vector $\rho = (\rho_1, \ldots, \rho_n)^t$, where,*

$$\rho_i = (-1)^{n+1} \sum_{T \in \Theta_i(G)} \left(\prod_{k \xrightarrow{a} l \in T} a \right). \tag{4}$$

There is a simple condition for x to also be in the null space of M.

Remark 3. Suppose that $x, \rho \in \mathbb{K}^n$ with $\rho_k \neq 0$ for some $1 \leq k \leq n$. Then, $x = \lambda \rho$ if, and only if, $x_i = (\rho_i/\rho_k)x_k$ for $1 \leq i \leq n$.

The quantities ρ_i/ρ_k will play an important role below; see (8), (20) and (24). In these calculations, the labels will lie in $\mathbb{R}[Q]$ and will either be symbols, like q_1, or S-positive polynomials, like $q_1 q_2 + q_3 q_4$. Under these conditions, we see from (4) that each ρ_i is either 0, if there are no spanning trees rooted at vertex i, of, if there are such spanning trees, ρ_i is $(-1)^{n+1}$ times a S-positive polynomial. Accordingly, ρ_i/ρ_k is a S-positive rational function.

2.4. The King–Altman method

If the biochemical mechanism of an enzyme is known, its rate function is often calculated in the quasi-steady state approximation (Cornish-Bowden, 1995). King and Altman worked out a graphical method for doing this that is widely used (Cornish-Bowden, 1995; King and Altman, 1956). It seems not to have been previously noticed that this is an immediate application of the Matrix-Tree theorem. Since we will see the Matrix-Tree theorem at work in more detail later, we illustrate the application to rate functions with the simple example of reversible Michaelis–Menten kinetics. Here, enzyme E and substrate S reversibly form an enzyme–substrate complex Y, which reversibly yields enzyme

a

$$E + S_u \rightleftarrows Y_j \rightleftarrows E + S_v$$

b

$$E + S_u \rightleftarrows Y_j \longrightarrow Y_k \rightleftarrows E + S_v$$
$$\updownarrow$$
$$Y_m$$

e

$$E + S_p$$
$$\updownarrow$$
$$Y_m \quad\quad E + S_w$$
$$\updownarrow \quad\quad \updownarrow$$
$$E + S_u \rightleftarrows Y_j \rightleftarrows Y_k \rightleftarrows E + S_v$$

d

$$E + S_v \longleftarrow Y_k \rightleftarrows E + S_u \rightleftarrows Y_j \longrightarrow E + S_v$$
$$\updownarrow$$
$$Y_m$$
$$\downarrow$$
$$E + S_w$$

c

$$\quad\quad\quad\quad\quad\quad E + S_p$$
$$\quad\quad\quad\quad\quad \nearrow$$
$$E + S_u \rightleftarrows Y_j \longrightarrow E + S_v$$
$$\quad\quad\quad\quad\quad \searrow$$
$$\quad\quad\quad\quad\quad\quad E + S_w$$

f

$$E + S_u \rightleftarrows Y_j \longrightarrow Y_k$$
$$\downarrow \quad\quad \downarrow$$
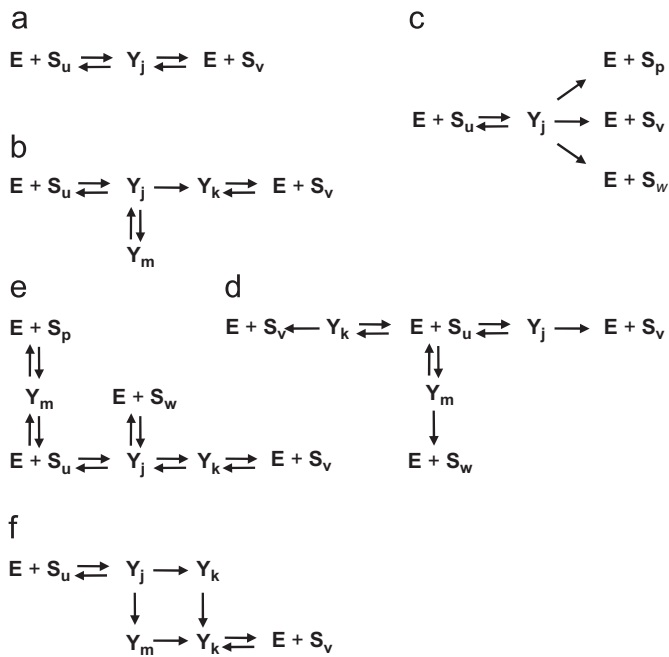$$Y_m \longrightarrow Y_k \rightleftarrows E + S_v$$

Fig. 1. Examples of sub-networks. (a) Michaelis–Menten style enzyme, with a single enzyme–substrate complex, Y_j, and reversible product formation, as in (5). (b) Example with two enzyme–substrate complexes, Y_j, Y_k, leading irreversibly to S_v, along with a dead-end complex Y_m. (c) Example used in Thomson and Gunawardena (2009), with a single enzyme–substrate complex Y_j and multiple products S_p, S_v, S_w. (d) Alternative network to (c) with a different enzyme–substrate complex for each product. (e) Example with partially overlapping routes for different products. (f) Cyclic network, which may not correspond to known biochemistry but is mathematically allowed.

and product P (compare Fig. 1a):

$$E+S \underset{a}{\overset{b}{\rightleftharpoons}} Y \underset{c}{\overset{d}{\rightleftharpoons}} E+P. \tag{5}$$

The rate constants, a, b, c, d are all taken to be positive. The quasi-steady state approximation assumes that Y reaches steady state while substrate is being converted to product. Under mass-action kinetics, the differential equations for E and Y are

$$\frac{dE}{dt} = -(aS+dP).E+(b+c)Y,$$

$$\frac{dY}{dt} = (aS+dP).E - (b+c)Y. \tag{6}$$

Note that E is at steady-state, if, and only if, Y is at steady state (compare Lemma 3). Hence, (5) is in quasi-steady state, if, and only if, (6) is in steady state. At steady-state, the right-hand side of (6) is a system of linear equations in E and Y. We can ignore the uninteresting case when $S = P = 0$ and assume that the coefficients of (6) are positive, thereby avoiding symbolic calculations. Let M be the corresponding matrix over $\mathbb{R}$ for the basis $\{E, Y\}$. We see from (6) that $1.M = 0$. Since $\text{rk}(M) = 1$ by inspection, we can use Lemma 2 to find solutions of $M.z = 0$. The labelled, directed graph formed from M according to Remark 2, is

$$E \underset{b+c}{\overset{aS+dP}{\rightleftharpoons}} Y. \tag{7}$$

This has a single spanning tree rooted at E, $Y \overset{b+c}{\longrightarrow} E$, and a single spanning tree rooted at Y, $E \overset{aS+dP}{\longrightarrow} Y$. By Lemma 2, a basis for the column null space of M is

$$\begin{pmatrix} -(b+c) \\ -(aS+dP) \end{pmatrix}.$$

By Remark 3, $M.z = 0$ if, and only if,

$$z_Y = \left[\left(\frac{a}{b+c} \right) S + \left(\frac{d}{b+c} \right) P \right] z_E. \tag{8}$$

We see that, at steady state, the enzyme–substrate complex is the free enzyme times a linear combination of substrate and product. The coefficients of the linear form are reciprocals of the forward and reverse Michaelis–Menten constants, $K_f = (b+c)/a$ and $K_r = (b+c)/d$ (Cornish-Bowden, 1995) (compare Proposition 1). The rate function can now be determined using the conservation law for the enzyme, $z_E + z_Y = E_{tot}$, giving, as in Cornish-Bowden (1995),

$$\frac{dP}{dt} = \frac{\dfrac{cE_{tot}}{K_f} S - \dfrac{dE_{tot}}{K_r} P}{1 + \dfrac{S}{K_f} + \dfrac{P}{K_r}}.$$

This simple calculation may provide some orientation for the more involved treatment that now follows.

3. Results

3.1. The system equations

We begin by setting up a general model for a PTM system. There are three kinds of chemical species in the system: enzymes, substrates and intermediate enzyme–substrate complexes. Let Enz, denote a non-empty, finite set of enzymes, Sub, a non-empty, finite set of substrate modforms and Int, a non-empty, finite set of enzyme–substrate complexes. A non-empty, finite set of sub-networks, Net, is defined in terms of these:

$$\text{Enz} = \{E_1, \ldots, E_L\},$$

$$\text{Sub} = \{S_1, \ldots, S_N\},$$

$$\text{Int} = \{Y_1, \ldots, Y_P\},$$

$$\text{Net} = \{T_1, \ldots, T_M\}.$$

Each sub-network, $T \in \text{Net}$, consists of an associated enzyme, $e(T) \in \text{Enz}$, a non-empty subset of modforms, $\sigma(T) \subseteq \text{Sub}$, a non-empty subset of enzyme–substrate complexes, $\gamma(T) \subseteq \text{Int}$, and a reaction network, $N(T)$, defined below. The sub-networks encode the biochemical details of how enzymes convert modforms.

This formulation allows for multiple forward and reverse enzymes, which may catalyse different types of modification and demodification. Multiple substrates are also permitted; the distinction between them will emerge in the calculation, as part of Condition 2. The combinatorics of multisite modification are not directly represented: the modforms of all substrates are simply listed $1, \ldots, N$. As discussed in the Introduction, N and P may be exponentially larger than L.

The enzyme–substrate subsets must be disjoint between distinct sub-networks: if $i \neq j$ then $\gamma(T_i) \cap \gamma(T_j) = \emptyset$. However, distinct sub-networks may share both modforms and enzymes. We assume, without loss of generality, that each substrate is in some $\sigma(T)$ and each enzyme–substrate complex in some $\gamma(T)$

$$\bigcup_{i=1}^{M} \sigma(T_i) = \text{Sub}, \quad\quad \bigcup_{i=1}^{M} \gamma(T_i) = \text{Int}.$$

Given $Y \in \text{Int}$, $t(Y) \in \text{Net}$ denotes the (unique) sub-network containing Y.

For $E = e(T)$, any $S_u \in \sigma(T)$ and any $Y_v, Y_i, Y_j \in \gamma(T)$, the sub-network $N(T)$ is made up of any reactions of the following

three kinds:

$$E + S_u \xrightarrow{a_{u,v}^T} Y_v,$$

$$E + S_u \xleftarrow{b_{u,v}^T} Y_v,$$

$$Y_i \xrightarrow{c_{i,j}^T} Y_j. \tag{9}$$
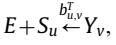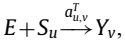
We further assume Condition 1 in Section 3.3 and Condition 2 in Section 3.4. These are strong connectivity conditions on certain graphs that allow Lemma 1 to be used. They will be stated after introducing additional concepts below.

The reactions (9) imply that enzyme is conserved—it is either free or bound in some enzyme–substrate complex—while substrate can flow between different modforms. While not much is known about the biochemical details of how enzymes modify multisite substrates, the sub-network assumptions allow considerable flexibility. They can accommodate, for instance, overlapping site preferences, arbitrary orders of modification and demodification, distributivity or processivity (Ferrell and Bhatt, 1997) and intricate hierarchical dependencies between enzymes (Ferrarese et al., 2007; Roach, 1991). Some simple examples are shown in Fig. 1. The main assumption behind (9) is that the donor molecules that provide the modifier are kept at constant concentration, on the time scale of the PTM dynamics, by mechanisms that are not explicitly modelled. As mentioned in the Introduction, this assumption has always been made for phosphorylation but needs further investigation for other modifications. It means that the donor molecules do not have to be treated as dynamical variables; their affects can be absorbed into the rate constants. This permits both forward and reverse reactions to be bimolecular with only enzyme and substrate. (In fact, we are implicitly making a similar assumption for the reverse reactions by ignoring the water molecules needed for hydrolysis.) Enzymes would normally be expected to give rise to tree networks, as in Fig. 1a–e, but cyclic networks like Fig. 1f are mathematically allowed.

The data above give rise to a polynomial dynamical system defined by mass-action kinetics on the set of chemical species, Sub ∪ Enz ∪ Int. For $1 \leq u \leq N$; for $1 \leq v \leq P$, $T = t(Y_v)$, $E = e(t(Y_v))$; and for $1 \leq w \leq L$,

$$\frac{dS_u}{dt} = \sum_{S_u \in \sigma(T)} \left(\sum_{e(T)+S_u \leftrightarrow Y_j \in N^\cdot(T)} (b_{u,j}^T Y_j - a_{u,j}^T e(T) S_u) \right), \tag{10}$$

$$\frac{dY_v}{dt} = \sum_{Y_v \leftrightarrow Y_j \in N^\cdot(T)} (c_{j,v}^T Y_j - c_{v,j}^T Y_v) + \sum_{E+S_j \leftrightarrow Y_v \in N^\cdot(T)} (a_{j,v}^T S_j E - b_{j,v}^T Y_v), \tag{11}$$

$$\frac{dE_w}{dt} = \sum_{e(T)=E_w} \left(\sum_{E_w + S_i \leftrightarrow Y_j \in N^\cdot(T)} \left(b_{i,j}^T Y_j - a_{i,j}^T E_w S_i \right) \right). \tag{12}$$

Terms like $c_{j,v}^T Y_j - c_{v,j}^T Y_v$ in Eq. (11) are indexed over edges in the undirected graph $N^\cdot(T)$. If one or other corresponding directed edge is not present in $N(T)$, the associated label should be treated as if it were zero. In other words,

$$\sum_{Y_v \leftrightarrow Y_j \in N^\cdot(T)} (c_{j,v}^T Y_j - c_{v,j}^T Y_v) = \sum_{Y_j \to Y_v \in N(T)} c_{j,v}^T Y_j - \sum_{Y_v \to Y_j \in N(T)} c_{v,j}^T Y_v.$$

Indexing over $N^\cdot(T)$ is purely a notational convenience, which allows for a more compact syntax.

3.2. Conservation laws

In this section, the rate constants and dynamical variables in (10)–(12) are treated as real variables. The structure of these equations can be better seen in terms of f_T, the net flux of enzyme out of sub-network T. By definition,

$$f_T = \sum_{e(T)+S_i \leftrightarrow Y_j \in N^\cdot(T)} b_{i,j}^T Y_j - a_{i,j}^T e(T) S_i,$$

so that Eq. (12) can be rewritten as

$$\frac{dE_w}{dt} = \sum_{e(T)=E_w} f_T. \tag{13}$$

Moreover, if (11) is added up for all $Y_v \in T$, we get a sum of two terms. The first term counts each binomial $c_{j,v}^T Y_j - c_{v,j}^T Y_v$ twice with opposite sign and hence vanishes. The second term is just $-f_T$. Hence,

$$f_T = -\sum_{Y_v \in \gamma(T)} \frac{dY_v}{dt}. \tag{14}$$

Finally, the enzyme flux from T is equal to the flux of all substrate modforms from T. Summed over all sub-networks, this gives the total substrate flux. More formally, if Eq. (10) is added up for all substrates, then, rearranging the order of summation, and noting that each $E+S \leftrightarrow Y$ is unique since the Y are not shared, we see that

$$\sum_{u \in \text{Sub}} \frac{dS_u}{dt} = \sum_{T \in \text{Net}} f_T. \tag{15}$$

From (13) and (14) we see that

$$\frac{d}{dt} \left(E_w + \sum_{e(T)=E_w} \left(\sum_{Y_v \in \gamma(T)} Y_v \right) \right) = 0.$$

The term being differentiated is evidently the total amount of enzyme E_w in the system, which we conclude to be the same at all times. Similarly, from (14) and (15) we see that

$$\frac{d}{dt} \left(\sum_{u \in \text{Sub}} S_u + \sum_{T \in \text{Net}} \left(\sum_{Y_v \in \gamma(T)} Y_v \right) \right) = 0.$$

The term being differentiated is the total amount of substrate in the system, which must also be the same at all times. Let S_{tot} be the total amount of substrate and $E_{w,tot}$ the total amount of enzyme E_w. We see that, at all times, the following $L+1$ conservation laws hold:

$$E_w + \sum_{e(T)=E_w} \left(\sum_{Y_v \in \gamma(T)} Y_v \right) = E_{w,tot}, \tag{16}$$

$$\sum_{u \in \text{Sub}} S_u + \sum_{T \in \text{Net}} \left(\sum_{Y_v \in \gamma(T)} Y_v \right) = S_{tot}. \tag{17}$$

Of course, these conservation laws are evident from the form of the allowed reactions in (9) but the derivation above checks the correctness of the system equations and illuminates the role of f_T. Reinforcing that, the following immediate consequence of (14) is a key result.

Lemma 3. *In any steady state, for any sub-network T, $f_T = 0$.*

It is instructive to see this another way, which does not depend on the details of how (14) is proved. In any steady state, the enzyme flux out of T cannot be negative. In other words, there cannot be positive flux of enzyme into T. If there were, this flux cannot escape through any $Y_v \in T$, since the enzyme–substrate

complexes are not shared between sub-networks. Hence, it would accumulate somewhere, violating the steady state assumption. Accordingly, $f_T \geq 0$. But then from (13), dE_w/dt is a sum of non-negative terms, so that if $dE_w/dt = 0$, then $f_T = 0$ for each $e(T) = E_w$.

3.3. Generalised Michaelis–Menten constants

Lemma 3 says that, at steady state, not merely is $dE_w/dt = 0$, but each individual $f_T = 0$. This decouples the system at steady state and allows us to treat each sub-network in isolation.

In this section and the next we will work symbolically. Let Con denote the set of all rate constants for all sub-networks $T \in$ Net and all reactions (9) in $N(T)$,

$$\text{Con} = \{a_{u,v}^T, b_{u,v}^T, c_{i,j}^T\}.$$

Let T be any sub-network and $E = e(T)$. In any steady state of the system, it follows from (11) and Lemma 3 that for all $Y_v \in \gamma(T)$ the following equations are satisfied:

$$\sum_{Y_v \leftrightarrow Y_j \in N^\bullet(T)} (c_{j,v}^T Y_j - c_{v,j}^T Y_v) + \sum_{E+S_j \leftrightarrow Y_v \in N^\bullet(T)} (a_{j,v}^T S_j E - b_{j,v}^T Y_v) = 0, \quad (18)$$

$$\sum_{E+S_i \leftrightarrow Y_j \in N^\bullet(T)} (b_{i,j}^T Y_j - a_{i,j}^T S_i E) = 0. \quad (19)$$

These form a system of linear equations in the variables E and $Y_v \in \gamma(T)$, with coefficients in $\mathbb{R}[\text{Con} \cup \text{Sub}]$. Assume, without loss of generality, that $\gamma(T) = \{Y_1, \ldots, Y_{p-1}\}$ where $p - 1 \leq P$. Let us use the notation $Y_p = E$ temporarily, for the purposes of this argument. Let M_T denote the $p \times p$ matrix over the field $\mathbb{R}(\text{Con} \cup \text{Sub})$ corresponding to (18) and (19). Evidently, $M_T.Y^t = 0$ where Y is the row vector, $Y = (Y_1, \ldots, Y_p)$. Furthermore, it follows from (14) that $1.M_T = 0$.

Let G_T be the labelled, directed graph formed from M_T as in Remark 2, so that $M_T = \mathcal{L}(G_T)$. Fig. 2b shows this graph for the sub-network in Fig. 1e. G_T is identical to $N(T)$ for all edges between Y_i for $1 \leq i < p$. However, each edge $E+S_u \xrightarrow{a_{u,v}^T} Y_v$ in $N(T)$ corresponds to the edge $Y_p \to Y_v$ in G_T with $a_{u,v}^T S_u$ added to its label. Similarly, each edge $Y_v \xrightarrow{b_{u,v}^T} E+S_u$ in $N(T)$ corresponds to the edge $Y_v \to Y_p$ in G_T with $b_{u,v}^T$ added to its label. The labels in G_T are all S-positive elements of $\mathbb{R}[\text{Con} \cup \text{Sub}]$. We can now state the first additional condition.

Condition 1. For any sub-network T, G_T is strongly connected.

All the examples in Fig. 1 satisfy this condition, which seems biochemically reasonable.

By Lemma 1, M_T has rank $p - 1$. Accordingly, a basis vector for the column null space is given by (4). The labels on the edges of G_T are rate constants, $c_{i,j}^T, b_{u,v}^T \in$ Con, except for the edges outgoing from Y_p, whose labels are homogeneous linear combinations of modforms (Fig. 2b). It follows that each spanning tree rooted at Y_p has a label product in $\mathbb{R}[\text{Con}]$, while any spanning tree rooted at Y_i, for $i \neq p$, has a label product that is homogeneous linear in the modforms with coefficients in $\mathbb{R}[\text{Con}]$. Hence, by (4), ρ_p is a S-positive element of $\mathbb{R}[\text{Con}]$, while ρ_i, for $i \neq p$, is a homogeneous linear combination of modforms whose non-zero coefficients are S-positive elements of $\mathbb{R}[\text{Con}]$. Let $\mu_{i,u}^T \in \mathbb{R}(\text{Con})$ be such that, for $1 \leq i < p$,

$$\frac{\rho_i}{\rho_p} = \sum_{S_u \in \sigma(T)} \mu_{i,u}^T S_u. \quad (20)$$

The $\mu_{i,u}^T$ are, by definition, generalised Michaelis–Menten constants (compare the discussion in Section 2.4). Note that, as defined here, these are reciprocals of the usual Michaelis–Menten constants (Cornish-Bowden, 1995). We prefer this convention because it allows these constants to be 0 when necessary, in preference to having to be ∞. By construction, the Michaelis–Menten constants are either 0 or are S-positive elements of $\mathbb{R}(\text{Con})$. In particular, they are well defined for any positive values of the rate constants in Con. The Michaelis–Menten constants for the example in Fig. 1e are shown in Table 1.

Using Remark 3, we have proved the following generalisation of (8).

Proposition 1. *For any sub-network $T \in$ Net, the sub-network Eqs. (18) and (19) are satisfied, if, and only if,*

$$Y_i = \left(\sum_{S_u \in \sigma(T)} \mu_{i,u}^T S_u \right) e(T), \quad (21)$$
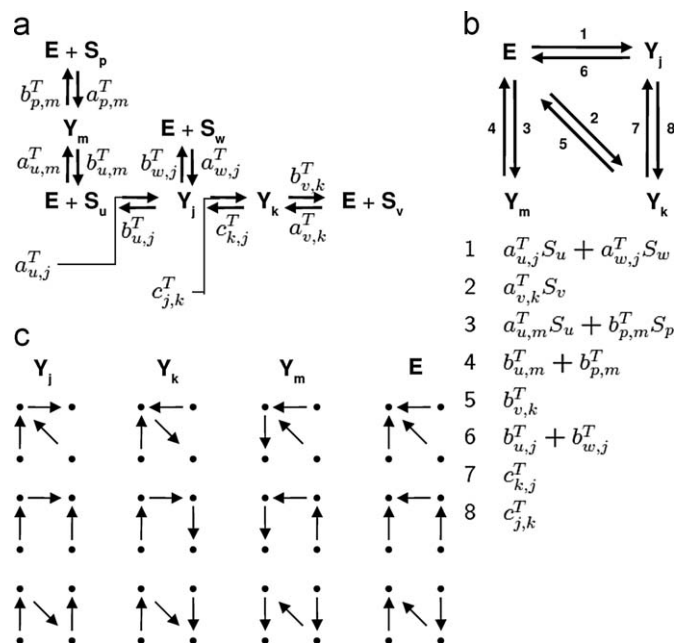
for $1 \leq i < p$.

Fig. 2. Steady state calculation for the sub-network in Fig. 1e. (a) The network in Fig. 1e with labels on the reactions according to (9). (b) The modified graph, G_T, on the vertices Y_j, Y_k, Y_m and E with the new labels in $\mathbb{R}[\text{Con} \cup \text{Sub}]$, as listed in the table below. The edges outgoing from E, numbered 1, 2 and 3, have labels which are linear and homogeneous in the modforms, while all other edges have labels which are rate constants. (c) The spanning trees rooted at each vertex of G_T, from which the maximal minors in (4) are calculated.

Table 1
Generalised Michaelis–Menten constants for the example in Fig. 1e.

	S_u	S_v	S_w	S_p
Y_j	$\dfrac{a_{u,j}^T(b_{v,k}^T + c_{k,j}^T)}{D}$	$\dfrac{a_{v,k}^T c_{k,j}^T}{D}$	$\dfrac{a_{w,j}^T(b_{v,k}^T + c_{k,j}^T)}{D}$	0
Y_k	$\dfrac{a_{u,j}^T c_{j,k}^T}{D}$	$\dfrac{a_{v,k}^T(b_{u,j}^T + b_{w,j}^T + c_{j,k}^T)}{D}$	$\dfrac{a_{w,j}^T c_{j,k}^T}{D}$	0
Y_m	$\dfrac{a_{u,m}^T}{b_{p,m}^T + b_{u,m}^T}$	0	0	$\dfrac{a_{p,m}^T}{b_{p,m}^T + b_{u,m}^T}$
D	$b_{v,k}^T(b_{w,j}^T + c_{j,k}^T) + b_{w,j}^T c_{k,j}^T + b_{u,j}^T(b_{v,k}^T + c_{k,j}^T)$			

The entries in the first three rows are the $\mu_{x,y}^T \in \mathbb{R}(\text{Con})$ defined in Proposition 1 for $x = j, k, m$ and $y = u, v, w, p$. The last row gives the denominator term D appearing in the entries for Y_j and Y_k. Note the S-positivity of all entries.

3.4. Linearising the modforms

In steady state, the enzyme–substrate complexes satisfy (21). Substituting these into the expressions for the modforms given by (10) we obtain

$$\sum_{S_u \in \sigma(T)} \left(\sum_{e(T)+S_u \leftrightarrow Y_j \in N^\bullet(T)} \left(\left(\sum_{S_v \in \sigma(T)} b_{u,j}^T \mu_{j,v}^T e(T) S_v \right) - a_{u,j}^T e(T) S_u \right) \right) = 0. \tag{22}$$

These expressions are linear in the S_u with coefficients which are S-positive polynomials in $\mathbb{R}(\text{Con})[\text{Enz}]$. Note the critical need at this point for Y_i in (21) to be linear in the S_u. Let M_S be the corresponding $N \times N$ matrix over $\mathbb{R}(\text{Con} \cup \text{Enz})$ for the basis $S_1, \ldots, S_N$. By Lemma 3, $f_T = 0$ in steady state for all sub-networks $T \in \text{Net}$. Hence, by (15), $1.M_S = 0$. Let G_S be the labelled, directed graph obtained from M_S as in Remark 2, so that $M_S = \mathcal{L}(G_S)$. To understand the structure of G_S, it is convenient to extend the definition of the Michaelis–Menten constants so that $\mu_{i,u}^T = 0$ for all $S_u \notin \sigma(T)$. We can then replace $S_u \in \sigma(T)$ in (20) with $S_u \in \text{Sub}$. With this convention, after collecting the coefficients of S_v in (22), we see that the label on the edge $S_v \to S_u$ may be written

$$\sum_{T \in \text{Net}} \left(\sum_{S_u + e(t) \leftarrow Y_j \in N(T)} b_{u,j}^T \mu_{j,v}^T \right) e(T),$$

whenever that expression is non-zero. Terms like $b_{u,j}^T \mu_{j,v}^T$ are familiar to biochemists as catalytic efficiencies. For each $T \in \text{Net}$ and any distinct pair $S_u, S_v \in \text{Sub}$, define the (generalised) catalytic efficiency, $\kappa_{u,v}^T$, by

$$\kappa_{u,v}^T = \sum_{S_u + e(T) \leftarrow Y_j \in N(T)} b_{u,j}^T \mu_{j,v}^T.$$

Note that $\kappa_{u,v}^T = 0$ if $S_v \notin \sigma(T)$. If not 0, $\kappa_{u,v}^T$ is a S-positive element of $\mathbb{R}(\text{Con})$. With this notation, we can rewrite the label on the edge $S_v \to S_u$ as

$$\sum_{T \in \text{Net}} \kappa_{u,v}^T e(T), \tag{23}$$

whenever that expression is non-zero. Fig. 3 shows G_S for the example in Fig. 1e.

The catalytic efficiencies provide a more concise set of labels for G_S than the original rate constants. Let Cat denote the set of all non-zero generalised catalytic efficiencies:

$$\text{Cat} = \{\kappa_{u,v}^T \neq 0 | S_u, S_v \in \text{Sub}\}.$$

Note that $\mathbb{R}(\text{Cat})$ is a subfield of $\mathbb{R}(\text{Con})$ and that any S-positive element of $\mathbb{R}(\text{Cat})$ is also S-positive as an element of $\mathbb{R}(\text{Con})$. The labels on G_S are S-positive polynomials in $\mathbb{R}[\text{Cat} \cup \text{Enz}]$, which are
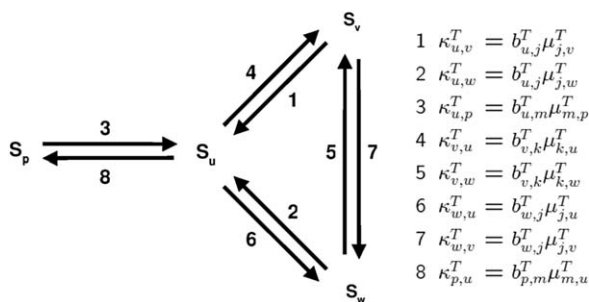
homogeneous and linear in the E_w. We may regard M_S as a $N \times N$ matrix over $\mathbb{R}(\text{Cat} \cup \text{Enz})$.

Since the modforms may include distinct substrates, we cannot assume that G_S is connected. We can now state the second condition for the systems considered here.

Condition 2. The connected components of G_S are strongly connected.

This condition is biochemically reasonable. For a given substrate, each modification is usually balanced by another enzyme carrying out a de-modification. This implies strong connectivity of the corresponding component of G_S. Distinct substrates will give rise to distinct components. Although Fig. 1e has only a single enzyme, the reversibility of the sub-network is sufficient in this case to give a single component which is strongly connected, as in Fig. 3.

Condition 2 implies that M_S is block diagonal, with each block corresponding to one of the connected components. Each block can be treated separately. To avoid further complicating the exposition, we assume from now on that M_S is a single block and that G_S is strongly connected. We point out what needs to be done for the general case but leave it to the reader to write down the details. Since, by Lemma 1, $\text{rk}(M_S) = N - 1$, we can apply Lemma 2 to obtain a basis vector for the column null space of M_S. Since G_S is strongly connected, it has at least one spanning tree rooted at each vertex. It follows from (23) that, in the notation of (4), ρ_i is a S-positive element of $\mathbb{R}[\text{Cat} \cup \text{Enz}]$ which is homogeneous in the E_w and of degree $N - 1$, since there are that many edges in any spanning tree. Hence, ρ_u/ρ_1 is a rational function in $\mathbb{R}(\text{Cat} \cup \text{Enz})$, which we may write as a rational function of the enzymes,

$$\frac{\rho_u}{\rho_1} = r_u(E_1, \ldots, E_L), \tag{24}$$

whose non-zero coefficients are S-positive elements of $\mathbb{R}[\text{Cat}]$. Here, we have, without loss of generality, chosen S_1 as a reference modform. With multiple components, a reference modform will be needed for each component. Since the ρ_u are each homogeneous of degree $N - 1$, it follows that, for $\lambda \in \mathbb{R}$,

$$r_u(\lambda E_1, \ldots, \lambda E_L) = r_u(E_1, \ldots, E_L), \tag{25}$$

so that the r_u are, in fact, inhomogeneous functions of any $L - 1$ of the variables. For instance, if $E_L \neq 0$, then

$$r_u(E_1, \ldots, E_L) = r_u\left(\frac{E_1}{E_L}, \ldots, \frac{E_{L-1}}{E_L}, 1\right).$$

Using Lemma 2 and Remark 3, we deduce the following analogue of Proposition 1.

Proposition 2. *The modform equations* (22) *are satisfied if, and only if,*

$$S_u = r_u(E_1, \ldots, E_L)S_1. \tag{26}$$

Examples of the rational functions r_u for the case of a single kinase, single phosphatase and single substrate with two sites are given in Manrai and Gunawardena (2008).

3.5. The main results

We can now put everything together and return from symbols to real variables. Let us consider any multisite PTM system satisfying the assumptions in Section 3.1 along with Conditions 1 and 2. We assume that the rate constants have positive real values. In this case, the maximal minor ρ_1 appearing in (24) is a real polynomial of degree $N - 1$ in the enzymes and $\rho_1 = 0$ defines a hypersurface in $\mathbb{R}^L$. As long as the vector of enzyme values lies off this hypersurface, the rational functions in (24) are well

1	$\kappa_{u,v}^T = b_{u,j}^T \mu_{j,v}^T$
2	$\kappa_{u,w}^T = b_{u,j}^T \mu_{j,w}^T$
3	$\kappa_{u,p}^T = b_{u,m}^T \mu_{m,p}^T$
4	$\kappa_{v,u}^T = b_{v,k}^T \mu_{k,u}^T$
5	$\kappa_{v,w}^T = b_{v,k}^T \mu_{k,w}^T$
6	$\kappa_{w,u}^T = b_{w,j}^T \mu_{j,u}^T$
7	$\kappa_{w,v}^T = b_{w,j}^T \mu_{j,v}^T$
8	$\kappa_{p,u}^T = b_{p,m}^T \mu_{m,u}^T$

Fig. 3. Modform graph, G_S, for the example in Fig. 1e with the labels shown as generalised catalytic efficiencies, $\kappa_{*,*}^T$, in the accompanying table. The $b_{*,*}^T$ are given in Fig. 2a while the generalised Michaelis–Menten constants, $\mu_{*,*}^T$ are given in Table 1.

defined. If G_S has many components, a similar proviso must be made for each of them, in respect of the maximal minor for the corresponding reference modform. Note that, since ρ_1 is S-positive in $\mathbb{R}[\text{Enz}]$, any vector in the positive orthant will satisfy $\rho_1 \neq 0$. Hence, under biochemically realistic conditions, the r_u are always well defined.

Theorem 2. *In any steady state of the system for which the enzyme values satisfy $\rho_1 \neq 0$, Eqs. (26) and (21) hold. Conversely, if the enzymes have values in $\mathbb{R}$ for which $\rho_1 \neq 0$, S_1 has values in $\mathbb{R}$, the S_u are defined by (26) and the Y_v are defined by (21), then these quantities form a steady state of the system.*

Proof. We have shown the first part above. Now suppose that $E_i \in \mathbb{R}$, such that $\rho_1(E_1, \ldots, E_L) \neq 0$, and $S_1 \in \mathbb{R}$. Define S_u by (26), which we may do since $\rho_1 \neq 0$, and Y_v by (21), as specified. For any $T \in \text{Net}$, it follows from Proposition 1 that the corresponding sub-network equations (18) and (19) are satisfied. Since (18) is just (11) at steady state, we see that $dY_v/dt = 0$ for all $Y_v \in T$. Furthermore, the expression on the left-hand side of (19) is, by definition, the net flux of enzyme out of T, f_T. Hence, $f_T = 0$. Since this holds for all $T \in \text{Net}$, we see from Eq. (13) that $dE_w/dt = 0$. Since (26) had to be satisfied to define the S_u, it follows from Proposition 2 that (22) is satisfied. Since this is just (10) at steady state after substitution of (21), which has also been satisfied, we see that $dS_u/dt = 0$. It follows that the system is at steady state. $\square$

If the system has specified total amounts of substrate and enzymes, it satisfies the conservation laws (16) and (17). These provide $L+1$ equations for the L enzymes and the substrate. For each sub-network $T \in \text{Net}$, let $\phi_T \in \mathbb{R}[\text{Enz}]$ denote the S-positive polynomial

$$\phi_T(E_1, \ldots, E_L) = \sum_{Y_v \in \gamma(T)} \left(\sum_{S_w \in \sigma(T)} \mu_{v,w}^T \rho_w(E_1, \ldots, E_L) \right).$$

It follows from Proposition 1, that, whenever $\rho_1 \neq 0$,

$$\sum_{Y_v \in \gamma(T)} Y_v = \phi_T e(T) \frac{S_1}{\rho_1}.$$

Let $\Delta \in \mathbb{R}[\text{Enz}]$ denote the S-positive polynomial,

$$\Delta = \sum_{u \in \text{Sub}} \rho_u + \sum_{T \in \text{Net}} \phi_T e(T). \tag{27}$$

Rewriting (17) to solve for S_1 in terms of S_{tot}, we see that, whenever $\Delta \neq 0$,

$$S_1 = \frac{\rho_1 S_{tot}}{\Delta}. \tag{28}$$

We can then rewrite (16) to get

$$E_w \left(1 + \left(\sum_{e(T) = E_w} \phi_T \right) \frac{S_{tot}}{\Delta} \right) = E_{w,tot}. \tag{29}$$

These L equations are well-defined whenever $\Delta \neq 0$. With multiple components, S_{tot} can be apportioned among the components and each reference modform will have a corresponding equation to (28).

Let $\Phi : \mathbb{R}^L \times \mathbb{R} \to \mathbb{R}^L$ be defined by the left-hand side of (29) so that, for $1 \leq w \leq L$,

$$\Phi_w(E_1, \ldots, E_L, S) = E_w \left(1 + \left(\sum_{e(T) = E_w} \phi_T \right) \frac{S}{\Delta} \right).$$

Theorem 3. *The steady states of a multisite PTM system are given by the solutions of a system of L equations for the L free enzyme concentrations. More precisely, if $A, E \in \mathbb{R}^L$, $S \in \mathbb{R}$, $\Delta \rho_1 \neq 0$ and*

$\Phi(E, S) = A$, then there is a steady state of the system for which E gives the free enzyme concentrations, $S = S_{tot}$ and $A_w = E_{w,tot}$. *Conversely, any steady state of the system having these totals, for which $\Delta \rho_1 \neq 0$, arises in this way.*

Proof. We have shown above that any steady state of the system having the specified totals, for which $\Delta \rho_1 \neq 0$, satisfies $\Phi(E, S) = A$. Now suppose that $E \in \mathbb{R}^L$, $S \in \mathbb{R}$ satisfy $\Delta \rho_1 \neq 0$ and that $\Phi(E, S) = A$. Define $S_1^* \in \mathbb{R}$ so that

$$S_1^* = \frac{\rho_1 S}{\Delta}$$

which we may do since $\Delta \neq 0$. Now use Theorem 2 to define a steady state of the system, S_u^*, Y_v^*, which we may do since $\rho_1 \neq 0$. We need only check that the total amount of substrate, S_{tot}^*, and enzymes, $E_{w,tot}^*$, satisfy $S_{tot}^* = S$ and $E_{w,tot}^* = A_w$. The total amount of substrate is

$$S_{tot}^* = \sum_{S_u \in \sigma(T)} S_u^* + \sum_{T \in \text{Net}} \sum_{Y_v \in \gamma(T)} Y_v^*.$$

Since (26) and (21) are satisfied through the use of Theorem 2, this gives

$$S_{tot}^* = \Delta \frac{S_1^*}{\rho_1} = S,$$

as required. Similarly, the total amount of enzyme E_w is

$$E_{w,tot}^* = E_w + \sum_{e(T) = E_w} \left(\sum_{Y_v \in \gamma(T)} Y_v^* \right) = E_w \left(1 + \left(\sum_{e(T) = E_w} \phi_T \right) \frac{S_1^*}{\rho_1} \right)$$

$$= \Phi_w(E, S) = A_w,$$

as required. $\square$

Since $\Delta \rho_1 \in \mathbb{R}[\text{Enz}]$ is S-positive, it never vanishes for positive free enzyme concentrations, which corresponds to the biochemically realistic case.

3.6. Algebraic geometry of the steady state

As suggested in the Introduction, Theorem 2 should be seen as an assertion that an appropriate algebraic variety is rationally parameterisable. This implies that points on the variety can be explicitly constructed as rational functions of some auxiliary parameters, in contrast to the implicit definition of points as solutions of polynomial equations (Cox et al., 1997). For instance, $x^2 + y^2 = 1$ provides an implicit definition of the unit circle, while the expression of x and y as

$$x = \frac{2t}{t^2 + 1}, \quad y = \frac{t^2 - 1}{t^2 + 1} \tag{30}$$

provides an explicit rational parameterisation. In general, a rational parameterisation may be undefined at points where the denominators of the rational functions vanish (which is not the case for (30)) and not all points on the variety may be represented (such as the point $(0, 1)$ in (30)).

Recall from Section 3.1 that there are $L+N+P$ dynamical variables in our system. Let $V \subseteq \mathbb{R}^{L+N+P}$ denote the steady-state variety, corresponding to the simultaneous solutions of the system equations (10)–(12). V is a real algebraic variety. Let $\pi : \mathbb{R}^{L+N+P} \to \mathbb{R}^N$ denote the projection on to the space of modforms. $\pi(V)$ may lack some points required for it to be an algebraic variety but it may be completed to one, if required (Cox et al., 1997).

The set $\pi(V)$ has additional structure, not present in V. Eq. (26) gives the modforms only up to a constant. If $\lambda \in \mathbb{R}$, then it is easy to see from the system equations (10)–(12) that multiplying each modform and each enzyme–substrate complex by λ, changes all the rates by λ. In particular, if the system is at steady state, then it

is still at steady state after such a change. Hence, $\pi(V)$ is a projective set: given $x \in \pi(V)$ the line through the origin and x also lies in $\pi(V)$. We can therefore consider $\pi(V)$ as a subset, $\pi(V)_{\mathbb{p}}$, of the real projective space $\mathbb{RP}^{N-1}$. The following is a corollary of Theorem 2.

Theorem 4. *$\pi(V)_{\mathbb{p}}$ has a rational parameterisation. Specifically, the rational functions r_u in (24) define a surjective mapping $\mathbb{R}^L \to \pi(V)_{\mathbb{p}}$, which is well-defined away from the hypersurface $\rho_1 = 0$.*

In view of (25), the dimension of $\pi(V)_{\mathbb{p}}$, once completed to a variety, is at most $L - 1$. For instance, in the case considered in Manrai and Gunawardena (2008) and Thomson and Gunawardena (2009), with a single kinase and a single phosphatase, $\pi(V)_{\mathbb{p}}$ is a rational curve.

4. Discussion

Our results show that the exponentially large number of state variables, $L+N+P$, of a multisite PTM system is determined at steady state by a relatively small "core" of L variables. We have provided in Propositions 1 and 2 a linear algebraic algorithm for calculating all steady-state variables in terms of the core. Tools like Mathematica can readily carry out linear algebra over symbolic fields like $\mathbb{R}(Q)$ and an example of such a program is available as the Supplementary Information to Manrai and Gunawardena (2008).

Previous steady-state analyses of multisite PTMs have largely focussed on phosphorylation. They have either used approximations, such as Michaelis–Menten or linear kinetics (Goldbeter and Koshland, 1981; Lisman, 1985; Markevich et al., 2004; Salazar and Höfer, 2007), which ignore sequestration effects when enzymes have multiple substrates (Ciliberto et al., 2007), or made simplifying biological assumptions, such as small numbers of sites (Manrai and Gunawardena, 2008) or sequential modification (Gunawardena, 2005), which limits their applicability. The results of the present paper provide a foundation for developing models that are closer to biological reality while also extending the scope of analysis to allow for multiple enzymes, multiple types of modification and multiple substrates. The permitted biochemical mechanisms are also considerably enlarged. Our method of proof reveals the significance of the Matrix-Tree theorem, which seems to play a key role in several forms of algebraic elimination in biochemical networks (Craciun et al., 2009; King and Altman, 1956).

The present paper develops the methodology. Applications of these results are found in previous papers which have focussed on phosphorylation with a single kinase, single phosphatase and single substrate (Manrai and Gunawardena, 2008; Thomson and Gunawardena, 2009). The capability to treat the number, n, of sites as a variable has allowed us to show that, for appropriate rate constants, the number of stable phospho-form distributions can be as many as $\lfloor (n+2)/2 \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer not greater than x (Thomson and Gunawardena, 2009). In particular, the number of stable phospho-form distributions increases with n. While some of these distributions are focussed on a few phospho-forms, others are more diffuse. Since different phospho-forms may have distinct biological effects (Park et al., 2006; Pufall et al., 2005; Wu et al., 2004), the phospho-proteome could be capable of substantial information processing. Indeed, it has been suggested that the remarkable variety of PTMs found on histone proteins in chromatin form a "code" for transcriptional regulation (Jenuwein and Allis, 2001; Turner, 2002). The analysis in Thomson and Gunawardena (2009) provides the first example of a PTM mechanism which is capable of encoding arbitrary amounts of information and gives an estimate of its information capacity.

A second application, for systems with two sites, has shown that the steady-state geometry can distinguish between different reaction networks (Manrai and Gunawardena, 2008). The geometry is detected algebraically through the use of "invariants", or polynomial functions of the steady-state phospho-form concentrations which depend only on the rate constants (Gunawardena, 2007). Invariants take the same value no matter what amounts of enzymes and substrates are present or what steady state is formed. To exploit such results we have been developing, in collaboration with Hanno Steen's group at Children's Hospital in Boston, mass-spectrometric methods for accurately quantifying phospho-form distributions. We are using this to map out, for the first time, the steady-state geometry of a kinase, phosphatase, substrate system: the MAP kinase Erk, which is doubly phosphorylated by the MAP kinase Mek and dephosphorylated by the dual-specificity phosphatase MKP3. Our experimental studies have already shown that these proteins engage in a more complex set of reactions than is commonly described in the literature and we are developing the method of invariants as a tool to work out the missing pieces.

Our results suggest several directions for future investigation. Which biochemical reaction networks have "cores", or small subsets of variables in terms of which all others can be calculated at steady state? If a core exists, is it unique? How can cores be identified and how can the functional relationship with non-core variables be algorithmically determined? Do these functional relationships give rational parameterisations of the steady-state variety? A particularly interesting generalisation would be to allow substrates to also be enzymes, thereby accommodating kinase cascades. The difficulty here is that those substrates that are also enzymes can, presumably, no longer be eliminated and must hence be in the "core", while, at the same time, as substrates, these variables may have non-trivial algebraic dependencies with other core variables. In the case treated here, the variables in the core are independent: their values can be arbitrarily assigned (Theorem 2). In more general cases there may be additional algebraic constraints on the core variables. We believe that the language and methods of algebraic geometry (Cox et al., 1997; Manrai and Gunawardena, 2008) will be particularly useful for unravelling such issues.

Looking further ahead, it is a tantalising question as to whether the elimination procedures developed here can be extended from the steady state to the dynamics, perhaps, initially, to the local vicinity of the steady state. Since steady-state stability is determined by the eigenvalues of the Jacobian, it does not seem implausible that differential algebraic methods could encompass both the steady state itself as well as its local vicinity. More globally, multisite PTM systems have other attractors, such as limit cycles, of which the cyanobacterial circadian oscillator is a particularly significant example (Nakajima et al., 2005; Rust et al., 2007). Hilbert's sixteenth problem asks about the number of limit cycles of two-dimensional polynomial dynamical systems. It remains unsolved, although some lower bounds are known (Christopher and Lloyd, 1995). Several lines of evidence, including the work presented here, suggest that the polynomial dynamics arising from biochemical reaction networks has very good properties at steady state. Could the same be true for other attractors like limit cycles?

Biologists are continually striving to elicit general principles from experimental data. Mathematical methods have been of less help in this respect than in accounting for the results of individual experiments. The present paper provides the tools to reason about post-translational modification systems without having to fix in advance many of the individual details, such as the number of enzymes or the combinatorics of modification.

Being able to rise above the molecular complexity while retaining biological and biochemical realism provides a complementary capability to that of simulation. If this capability can be extended to a broad range of cellular processes, we will have a powerful tool with which to articulate the principles of cellular information processing.

Acknowledgements

References

Angeli, D., Ferrell, J.E., Sontag, E.D., 2004. Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc. Natl. Acad. Sci. USA 101, 1822–1827.

Christopher, C.J., Lloyd, N.G., 1995. Polynomial systems: a lower bound for the Hilbert numbers. Proc. Roy. Soc. London A 450, 219–224.

Chung, F.R.K., 1997. Spectral graph theory. In: Regional Conference Series in Mathematics, vol. 92. American Mathematical Society, Providence, RI.

Ciliberto, A., Capuani, F., Tyson, J.J., 2007. Modeling networks of coupled enzymatic reactions using the total quasi-steady state approximation. PLoS Comput. Biol. 3, e45.

Cornish-Bowden, A., 1995. Fundamentals of Enzyme Kinetics, second ed Portland Press, London, UK.

Cox, D., Little, J., O'Shea, D., 1997. Ideals, Varieties and Algorithms, second ed Springer, Berlin.

Craciun, G., Dickenstein, A., Shiu, A., Sturmfels, B., 2009. Toric dynamical systems. J. Symbolic Comput. 44, 1551–1565.

Craciun, G., Tang, Y., Feinberg, M., 2006. Understanding bistability in complex enzyme-driven reaction networks. Proc. Natl. Acad. Sci. USA 103, 8697–86102.

Feinberg, M., 1979. Lectures on Chemical Reaction Networks, Lecture Notes, Mathematics Research Center, University of Wisconsin, 1979. ⟨www.che.eng.ohio-state.edu/~feinberg/research/⟩.

Ferrarese, A., Marin, O., Bustos, V.H., Venerando, A., Antonelli, M., Allende, J.E., Pinna, L.A., 2007. Chemical dissection of the APC Repeat 3 multistep phosphorylation by the concerted action of protein kinases CK1 and GSK3. Biochemistry 46, 11902–11910.

Ferrell, J.E., Bhatt, R.R., 1997. Mechanistic studies of the dual phosphorylation of mitogen-activated protein kinase. J. Biol. Chem. 272, 19008–19016.

Fischer, E.H., 1997. Protein phosphorylation and cellular regulation, II. In: Ringertz, N. (Ed.), Nobel Lectures, Physiology or Medicine 1991–1995. World Scientific, Singapore.

Gatermann, K., Huber, B., 2002. A family of sparse polynomial systems arising in chemical reaction systems. J. Symbolic Comput. 33, 273–305.

Goldbeter, A., Koshland, D.E., 1981. An amplified sensitivity arising from covalent modification in biological systems. Proc. Natl. Acad. Sci. USA 78 (11), 6840–6844.

Gunawardena, J., 2003. Chemical Reaction Network Theory for *in-silico* biologists, Lecture Notes, Harvard University, 2003. ⟨vcp.med.harvard.edu/papers/crnt.pdf⟩.

Gunawardena, J., 2005. Multisite protein phosphorylation makes a good threshold but can be a poor switch. Proc. Natl. Acad. Sci. USA 102, 14617–14622.

Gunawardena, J., 2007. Distributivity and processivity in multisite phosphorylation can be distinguished through steady-state invariants. Biophys. J. 93, 3828–3834.

Gunawardena, J., 2009. Models in systems biology: the parameter problem and the meanings of robustness. In: Lodhi, H., Muggleton, S. (Eds.), Elements of Computational Systems Biology. Wiley Book Series on Bioinformatics. Wiley, New York.

Herstein, I., 1975. Topics in Algebra. Wiley, New York.

Hunter, T., 2007. The age of crosstalk: phosphorylation, ubiquitination and beyond. Mol. Cell 28, 730–738.

Jenuwein, T., Allis, C.D., 2001. Translating the histone code. Science 293, 1074–1080.

Kim, S.Y., Ferrell, J.E., 2007. Substrate competition as a source of ultrasensitivity in the inactivation of Wee1. Cell 128, 1133–1145.

King, E.L., Altman, C., 1956. A schematic method of deriving the rate laws for enzyme-catalyzed reactions. J. Phys. Chem. 60, 1375–1378.

Krebs, E.G., 1997. Protein phosphorylation and cellular regulation, I. In: Ringertz, N. (Ed.), Nobel Lectures, Physiology or Medicine 1991–1995. World Scientific, Singapore.

Kruse, J.-P., Gu, W., 2008. SnapShot: p53 posttranslational modifications. Cell 133, 930.

Lisman, J.E., 1985. A mechanism for memory storage insensitive to molecular turnover: a bistable autophosphorylating kinase. Proc. Natl. Acad. Sci. USA 82, 3055–3057.

Manrai, A., Gunawardena, J., 2008. The geometry of multisite phosphorylation. Biophys. J. 95, 5533–5543.

Markevich, N.I., Hoek, J.B., Kholodenko, B.N., 2004. Signalling switches and bistability arising from multisite phosphorylation in protein kinase cascades. J. Cell Biol. 164, 353–359.

Moon, J.W., 1970. Counting labelled trees. In: Canadian Mathematical Monographs, vol. 1. Canadian Mathematical Congress.

Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., Kondo, T., 2005. Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro. Science 308, 414–415.

Park, K.-S., Mohapatra, D.P., Misonou, H., Trimmer, J.S., 2006. Graded regulation of the Kv2.1 potassium channel by variable phosphorylation. Science 313, 976–979.

Pesavento, J.J., Bullock, C.R., LeDuc, R.D., Mizzen, C.A., Kelleher, N.L., 2008. Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry. J. Biol. Chem. 283, 14927–14937.

Phanstiel, D., Brumbaugh, J., Berggren, W.T., Conrad, K., Feng, X., Levenstein, M.E., McAllister, G.C., Thomson, J.A., Coon, J.J., 2008. Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. Proc. Natl. Acad. Sci. USA 105, 4093–4098.

Pufall, M.A., Lee, G.M., Nelson, M.L., Kang, H.-S., Velyvis, A., Kay, L.E., McIntosh, L.P., Graves, B.J., 2005. Variable control of Ets-1 DNA binding by multiple phosphates in an unstructured region. Science 309, 142–145.

Roach, P.J., 1991. Multisite and hierarchal protein phosphorylation. J. Biol. Chem. 266, 14139–14142.

Rust, M.J., Markson, J.S., Lane, W.S., Fisher, D.S., O'Shea, E., 2007. Ordered phosphorylation governs oscillation of a three-protein circadian clock. Science 318, 809–812.

Salazar, C., Höfer, T., 2007. Versatile regulation of multisite protein phosphorylation by the order of phosphate processing and protein–protein interactions. FEBS J. 274, 1046–1060.

Shacter-Noiman, E., Chock, P.B., Stadtman, E.R., 1983. Protein phosphorylation as a regulatory device. Philos. Trans. R. Soc. London B 302, 157–166.

Soulé, C., 2003. Graphic requirements for multistationarity. ComPlexUs 1 123–133

Strogatz, S.H., 2001. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering. Perseus Books.

Thomson, M., Gunawardena, J., 2009. Unlimited multistability in multisite phosphorylation systems. Nature 460, 274–277.

Turner, B., 2002. Cellular memory and the histone code. Cell 111, 285–291.

Tutte, W.T., 1948. The dissection of equilateral triangles into equilateral triangles. Proc. Cambridge Philos. Soc. 44, 463–482.

Walsh, C.T., 2006. Posttranslational Modification of Proteins. Roberts and Company, Englewood, Colorado.

Wu, R.C., Qin, J., Yi, P., Wong, J., Tsai, S.Y., Tsai, M.J., O'Malley, B.W., 2004. Selective phosphorylations of the SRC-3/AIB1 coactivator integrate genomic responses to multiple cellular signaling pathways. Mol. Cell 15, 937–949.