

Modeling TGF- β signaling pathway in epithelial-mesenchymal transition

Pasquale Laise, Duccio Fanelli, Pietro Lió, and Annarosa Arcangeli

Citation: *AIP Advances* **2**, 011201 (2012); doi: 10.1063/1.3697962

View online: <http://dx.doi.org/10.1063/1.3697962>

View Table of Contents: <http://aip.scitation.org/toc/adv/2/1>

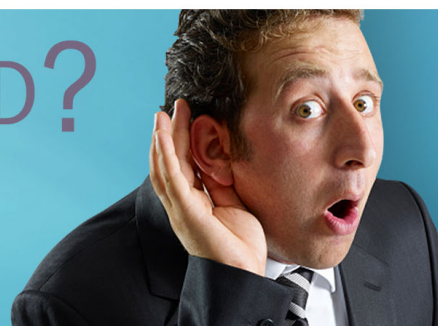
Published by the [American Institute of Physics](#)

HAVE YOU HEARD?

Employers hiring scientists and
engineers trust

PHYSICS TODAY | JOBS

www.physicstoday.org/jobs



Modeling $TGF-\beta$ signaling pathway in epithelial-mesenchymal transition

Pasquale Laise,¹ Duccio Fanelli,^{1,2} Pietro Lió,³ and Annarosa Arcangeli⁴

¹*CSDC Centro Interdipartimentale per lo Studio di Dinamiche Complesse and INFN, University of Florence, Italy*

²*Dipartimento di Energetica “S. Stecco” and INFN, University of Florence, Via S. Marta 3, 50139 Florence, Italy*

³*Computer Laboratory, University of Cambridge, 15 JJ Thompson Avenue, CB03FD Cambridge, United Kingdom*

⁴*Dipartimento di Patologia e Oncologia Sperimentali, University of Florence, Istituto Toscano Tumori, Florence, Italy*

(Received 4 November 2011; accepted 31 January 2012; published online 21 March 2012)

The epithelial-mesenchymal transition (EMT) consists in a morphological change in epithelial cells characterized by the loss of the cell adhesion and the acquisition of mesenchymal phenotype. This process plays a crucial role in the embryonic development and in regulating the tissue homeostasis in the adult, but it proves also fundamental for the development of cancer metastasis. Experimental evidences have shown that the EMT depends on the $TGF-\beta$ signaling pathway, which in turn regulates the transcriptional cellular activity. In this work, a dynamical model of the $TGF-\beta$ pathway is proposed and calibrated versus existing experimental data on lung cancer A549 cells. The analysis combines Bayesian Markov Chain Monte Carlo (MCMC) and standard Ordinary Differential Equations (ODEs) techniques to interpolate the gene expression data via an iterative adjustments of the parameters involved. The kinetic of the Smad proteins phosphorylation, as predicted within the model is found in excellent agreement with available experiments, an observation that confirms the adequacy of the proposed mathematical picture. *Copyright 2012 Author(s). This article is distributed under a Creative Commons Attribution 3.0 Unported License. [http://dx.doi.org/10.1063/1.3697962]*

I. INTRODUCTION

The expression “epithelial-mesenchymal transition (EMT)” indicates the morphological transformation of epithelial cells in mesenchymal cells. This is a cellular process that occurs under physiological conditions, as in embryonic development and/or regulation of the tissue homeostasis in adults. EMT is however also implicated in cancer progression and others pathological conditions.¹ Biochemically, the transformation takes place through the down-regulation of specific epithelial proteins and the up-regulation of mesenchymal proteins. Cells which undergo the EMT transformation present a new phenotype, acquiring a pronounced migratory ability that promotes their diffusion towards adjacent districts. This is a fundamental step which seeds the physiological differentiation during embryonic development. Importantly, it is also fueling the spreading of cancer signaling.² At the cellular level, no significant difference is currently detected between pathological and physiological EMT in that they both appear to be governed by similar signaling pathways.¹ Starting from this setting, it would be extremely important to identify dedicated EMT features, specifically associated to tumour derives, an ambitious aim that could eventually translate in novel protocols to stratify the patients and predict the outcome of the disease. Several experiments unambiguously established that the EMT depends on the $TGF-\beta$ signalling, the coupling being mediated by an intricate network of microscopic molecular reactions.² To gain insight into the $TGF-\beta$ signaling pathway, and elucidate the interlaced connection among molecular constituents, a plethora of models were proposed in the past and benchmarked to direct experiments.³⁻⁶ Among the others, the model proposed by Schmierer



and colleagues³ appears particularly interesting for the quality of the experimental input processed and, at the same time, the richness of the biochemical details accommodated for in the hypothesized reactions scheme. Smad intracellular proteins are for instance explicitly included: Smad transduce extracellular signals from *TGF- β* ligands to the nucleus where they activate downstream the transcription of specific target genes. The mechanism of Smad nucleocytoplasmic shuttling was studied in Schmierer *et al.*³ by monitoring the nuclear accumulation of phosphorylated Smad proteins. To this end, a fitting strategy was implemented to quantitatively compare the model predictions to time course kinetics experiments, performed on HaCaT cells in response to *TGF- β* . The experiments covered however a relatively short time window, 150 minutes upon administration of *TGF- β* . Noteworthy, while constituting a reasonable skeleton for validating the model, the data employed in Schmierer *et al.*³ did not allow one to appreciate, and fully resolve, the *TGF- β* induced dynamics. This latter in fact requires 72 hours for the morphological transformation of the cells to eventually take place.⁷ In this work, extending beyond the former analysis, we intend to characterize the dynamics of *TGF- β* in the EMT, embracing the time frame of interest. To constrain our model, we will make use of gene expression data relative to A549 cells, in response to 5 ng/mL of *TGF- β* , over a window of observation of 72 hours. These data were produced using single-color microarray technology, which enabled one to simultaneously measure the mRNA concentration of thousands of genes,⁸ including those involved in the *TGF- β* pathway, object of our analysis. As we shall discuss in the following, the model that we will put forward is rather simplified and inspired to a reductionist approach. A limited number of biochemical proteins is solely considered that however organize in a self-consistent dynamical picture to describe the *TGF- β* pathway. Clearly, a much larger gallery of reactions should be accounted for when aiming at resolving all the details implicated in the examined processes. Our philosophy is however different. We are in fact interested in elucidating a minimalistic subset of microscopic reactions that are supposedly responsible for the observed macroscopic experimental behaviours. It is worth emphasizing that the hypothesized chemical reactions, although simple, follows current understanding of the processes involved, see e.g. Ting Liu *et al.*⁹ Despite its inherent simplicity, our model proves capable of successfully explaining the alteration in the transcription process as induced by *TGF- β* in EMT. The quality of the fitting and the predictive ability of the model provide in turn an a posteriori validation to the proposed formulation. More specifically, as concerns the model setting, we shall focus our attention on the following molecular species: the *TGF- β* , the *TGF- β* receptor, the Smad proteins, the Phosphatases and finally the gene *hmg2*. *TGF- β* exerts its activity by forming a ligand-receptor complex with the *TGF- β* receptor, which then switches to its active form and consequently transmits the signal into the cytoplasm by phosphorylation of the Smad proteins. The Smad proteins are transcription factors. They act as effective carriers which transport the *TGF- β* signal from the cell surface to the nucleus. Inside the nucleus, the Smad proteins can be dephosphorylated by phosphatases or express their transcriptional activity by stimulating the mRNA production of *hmg2* gene, which belongs to the non-histone chromosomal high mobility group (HMG) protein family. Within our formulation, schematically summarized in figure 1, the modulation in gene expression of *hmg2* is interpreted as a direct transcriptional effect of *TGF- β* .² By invoking a straightforward application of the law of mass action, a closed set of ordinary differential equations for the concentration of the chemical species involved is recovered and numerically benchmarked to available experiments. As we will clarify in the forthcoming sections, the model parameters have been adjusted so to reproduce *hmg2* gene expression data. This fitting procedure is customarily adopted in pharmacogenomics^{10,11} and yields to reliable estimates for the a priori unknown kinetic rates, providing in turn a viable strategy to reconstruct the underlying genetic networks.¹² Surprisingly, the model constrained to interpolate the time evolution of *hmg2* gene expression data, predicts a non trivial behavior for the Smad protein phosphorylation dynamics which perfectly matches the experimental measurements of Keshamouni and colleagues.¹³ These latter data, not employed in the calibration phase, and therefore treated as an independent experimental input, were generated using western immunoblotting technique, and quantify the protein expression of phosphorylated Smad in lung cancer A549 cells induced by 5ng/mL of *TGF- β* for 72 hours. The observation of a quantitative correspondence between (constrained) predictions and (independent) experiments, constitutes the main conclusion of our work and testifies on the validity of the proposed formulation. From a methodological point

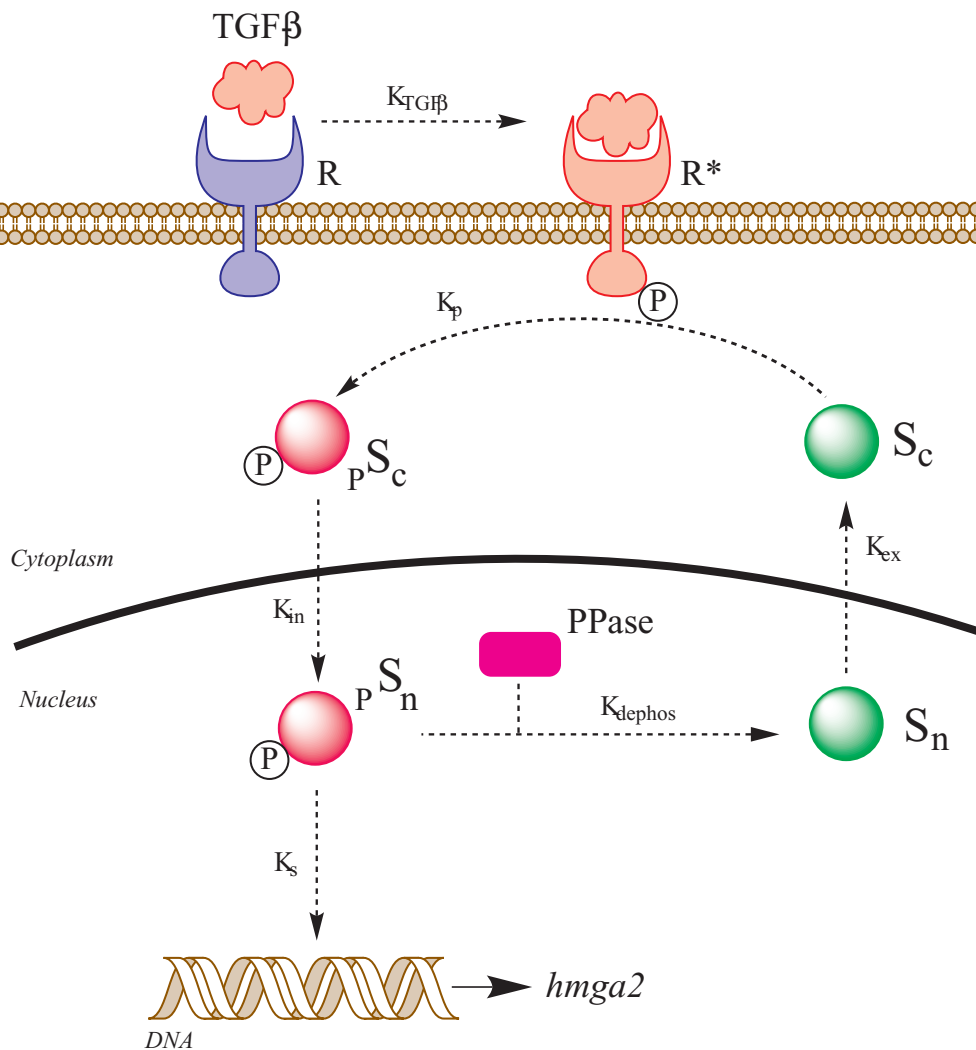


FIG. 1. Cartoon of *TGF-β* pathway dynamics. For a detailed account of the process involved which inspired our formulation refer to Ting Liu *et al.*⁹ The *TGF-β* molecule interacts with the receptor (R), which then turns into its active form (R^*). The active receptor (R^*) phosphorylates the cytoplasmatic Smad proteins (S_c), which transform into pS_c . The phosphorylated Smad shift to the nucleus. In the nucleus pS_n can stimulate the transcription of the target gene *hmga2*. Otherwise it can be dephosphorylated by the phosphatases (PPase). The dephosphorylated nuclear Smad proteins (S_n) are exported back to the cytoplasm.

of view, we complement the direct fitting strategy with a dedicated statistical based analysis to assess the model robustness. Particularly relevant, in this context, is the use of Bayesian MCMC (Markov Chain Monte Carlo) techniques, which allows to estimate the posterior data-dependent distribution of the uncertain parameters, accounting for the unavoidable noise which affects the experimental data.

The paper is organized as follows: in the next Section we introduce the chemical model and discuss its mathematical formulation. In Section IV we compare the theoretical results with the experimental data, building on the fitting procedure. Particular emphasis is put on discussing the predictive ability of the model, as concerns the time evolution of the Smad proteins. In section V, we address the problems of sensitivity and MCMC analysis. Finally, in Section VI we sum up and draw our conclusions.

II. THE MODEL

In order to study the dynamics of the $TGF-\beta$ during the EMT, we propose a simple biochemical model which explicitly incorporates a limited number of key molecular species, assumed to mutually interact. The aim of the analysis is to explain, within a minimalist self-consistent dynamical picture, the modification of the gene expressions, driven by $TGF-\beta$, and as seen in the experiments. In the following we shall provide a critical guide through the model formulation, making reference to the schematic outline depicted in figure 1. The model consists in a collection of chemical equations, which encode for the interactions among individual entities. As we shall clarify later on, the chemical model readily translates in a closed set of ordinary differential equations for the coupled evolution of the concentration amounts.

The $TGF-\beta$ belongs to the family of cytokines. It is produced by the cells in response to an external stimulus. Under normal conditions, it plays a role of paramount importance in several cellular processes as e.g. growth, differentiation and death.^{14,15} The first equation of the model, see (1), accounts for the interaction between $TGF-\beta$ and the inactive membrane receptor (R). Following a successful encounter, the receptor turns into its active form here denoted R^* . In formulae:



the parameter $k_{TGF\beta}$ labeling the associated reaction constant.

In reality, the receptor is a complex constituted by two units, the $TGF-\beta$ -RII and $TGF-\beta$ -RI respectively. Here, we drastically simplify the model setting by considering just one receptor type, which can operate in its active (bound to $TGF-\beta$) or inactive configuration. The activated receptor is able to phosphorylate the Smad proteins (S_c) in the cytoplasm, resulting in the formation of a new species, the phosphorylated Smad protein, here labelled pS_c . This process is mimicked by equation (2):



In the above relation we assume that R^* gets degraded once the Smad proteins are phosphorylated. In principle, one should foresee the possibility for the $TGF-\beta$ to leave its targeted receptor upon completion of the deputed task and so free the binding site. In the proposed model, this latter condition is deliberately neglected, a choice which amounts to assume that the release of the $TGF-\beta$, spoiled of the phosphors, occurs on longer times scale. As a matter of facts, we also emphasize that two are the independent Smad components that get phosphorylated (respectively, pS_2 and pS_3) and that subsequently combine to result in the actual unit pS_c . This process is compressed into a single step, as specified by chemical equation (2).

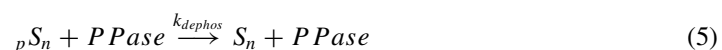
Once phosphorylated, the Smad proteins head towards the nucleus. The translocation of the Smad proteins into the nucleus (pS_n) is necessary to activate the transcriptional activity. This is a delicate process, possibly organized in nested regulatory cycles. We have nevertheless decided to cast it in the minimalistic form:



implying that cytoplasmatic pS_c are mutated into nuclear pS_n with given rate specified by the control parameter k_{in} . The presence of phosphorylated Smad in the nucleus is in turn associated to an increase of the *hmg2* gene expression² as governed by the following equation (4):



The nucleus is also enriched by phosphates ($PPase$) targeting phosphorylated Smad,³ a fact which in our view contributes to the regulation of transcriptional activity of Smad proteins.⁹ The dephosphorylation of the nuclear Smad proteins pS_n , which results in non-phosphorylated nuclear Smad elements S_n , is accounted for by the following reaction:



Non-phosphorylated nuclear Smad S_n can be then exported back to the cytoplasm via the following equation:



The above set of chemical reactions translates straightforwardly into a mathematical model, that we shall benchmark to the existing experimental literature in the forthcoming sections. The model is in particular composed of eight independent variables, respectively $TGF-\beta$, R , R^* , S_c , ${}_pS_c$, ${}_pS_n$, S_n and $hmg2$. The associated concentrations, indicated with the usual notation $[\cdot]$, obey to the following set of ordinary differential equations:

$$\begin{aligned} \frac{d[TGF\beta]}{dt} &= -k_{TGF\beta} [TGF\beta] [R] \\ \frac{d[R]}{dt} &= -k_{TGF\beta} [TGF\beta] [R] \\ \frac{d[R^*]}{dt} &= k_{TGF\beta} [TGF\beta] [R] - k_p [R^*] [S_c] \\ \frac{d[S_c]}{dt} &= -k_p [R^*] [S_c] + k_{ex} [S_n] \\ \frac{d[{}_pS_c]}{dt} &= k_p [R^*] [S_c] - k_{in} [{}_pS_c] \\ \frac{d[{}_pS_n]}{dt} &= k_{in} [{}_pS_c] - k_s [{}_pS_n] - k_{dephos} [{}_pS_n] [PPase] \\ \frac{d[S_n]}{dt} &= k_{dephos} [{}_pS_n] [PPase] - k_{ex} [S_n] \\ \frac{d[hmg2]}{dt} &= k_m [{}_pS_n] \end{aligned} \quad (7)$$

$hmg2$ measures the mRNA quota and as such assumes arbitrary units. All other concentrations are instead expressed in nM. For this reason the auxiliary parameter k_m has been introduced in last of equations (7): $k_m = \alpha k_s$, α being a dimensional quantity that restores the correct dimensional balance in the above equations. Notice that the concentration of $PPase$ remains unchanged all along the dynamical evolution.

Asymptotically, the system admits four distinct families of fixed points. These are listed in the following:

- (i) $R = R^* = S_n = {}_pS_c = {}_pS_n = 0$ and $TGF-\beta$, $hmg2$, S_c arbitrary
- (ii) $R = S_c = S_n = {}_pS_c = {}_pS_n = 0$ and $TGF-\beta$, $hmg2$, R^* arbitrary
- (iii) $TGF-\beta = S_c = S_n = {}_pS_c = {}_pS_n = 0$ and R , $hmg2$, R^* arbitrary
- (iv) $TGF-\beta = R^* = S_n = {}_pS_c = {}_pS_n = 0$ and R , $hmg2$, S_c arbitrary

As we shall make clear in the following, our system converges to solution (iv), given the specific initial conditions that are being imposed so to benchmark theory and experiments. Notice that the quantities R , S_c and $hmg2$ converge to stationary values, that are function of the initial condition and of the kinetic parameters involved.

In the following analysis we shall concentrate on the out of equilibrium dynamics of system (7), this latter being perturbed by the injection of a specific $TGF-\beta$ quota. The time that it is necessary for $TGF-\beta$ to fade off, i.e. for the system to recover its equilibrium steady state configuration, defines our window of observation. The next section is devoted to discussing the initial condition and the parameters choice. Then, in the forthcoming sections, we shall first test the model performance versus the existing experimental literature and subsequently, upon calibration, evaluate its predictive ability.

III. THE SELECTED INITIAL CONDITION

As already stressed, the proposed model builds on current biological knowledge,^{3,9} and results in a simplified, though self-consistent, dynamical formulation. The *TGF- β* pathway is in particular reduced to a limited set of chemical reactions that are presumably implicated in the transmission of the signal from the cell surface, as triggered by *TGF- β* , to the cell nucleus. The current analysis is in particular aimed at inspecting the out-of-equilibrium dynamics of the system, as driven by the externally imposed *TGF- β* . The unperturbed steady state is characterized by specific concentrations of cytoplasmatic Smad proteins [S_c] and unbound receptors [R], which have been carefully evaluated in a study by Schmierer and collaborators,³ that we shall here assume as a reference. In particular (see Table S1, in the Supporting Information of Schmierer *et al.*³) we set [S_c] = 121.1 nM and [R] = 1 nM. Moreover, still based on Schmierer *et al.*,³ we require [$PPase$] = 1 nM, this latter being an invariant quantity of the dynamics. All other variables, with the exception of *TGF- β* (see below) are initially set to zero. Moreover, the kinetic rates $k_{TGF\beta}$, k_{in} , k_{dephos} and k_{ex} are also taken from Schmierer *et al.*,³ a working assumption that entails the possibility of dealing with a modest number of residual parameters. Specifically, $k_{TGF\beta} = 0.074 \text{ nM}^{-1} \text{ s}^{-1}$, $k_{in} = 0.0026 \text{ s}^{-1}$, $k_{dephos} = 0.0066 \text{ nM}^{-1} \text{ s}^{-1}$ and $k_{ex} = 0.0056 \text{ s}^{-1}$ (see Fig 2 of Schmierer *et al.*³ and Table S1, in the Supporting Information of Schmierer *et al.*³). Hence, the model rests on just two undetermined kinetic parameters k_p , k_s , and on the, presently unknown, scaling coefficient α . The concentration of *TGF- β* , referred to the initial time when the perturbation is applied, can be assigned based on the particular experimental setting that one aims at reproducing. In the next section, we will consider publicly available gene expression data, that report on the time evolution of *hmg2* concentration, as induced by a perturbation of the steady state regime via *TGF- β* insertion. Initially, [*hmg2*] is assumed to be zero, so to match the experimental conditions.

IV. THEORY VS. EXPERIMENTS: FITTING THE MODEL TO GENE EXPRESSION DATA

The purpose of this section is to benchmark the model to experimental measurements and so produce a direct estimates of the parameters involved. We here recall that only three parameters are to be assigned, namely k_p , k_s and α , the other being set to values previously reported in the literature.

The experimental dataset that is employed in the calibration phase is taken from Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>). The dataset GSE17708 reports in particular on the temporal evolution of the level of expression (corresponding to mRNA concentration) associated to the *hmg2* gene. The data, organized in 9 successive snapshots ranging from 0 to 72 hours, refer to lung cancer A549 cells exposed to *TGF- β* . At time $t = 0$, the cells are treated with 5 ng/mL of *TGF- β* , a concentration that is equivalent to 0.113 nM. This latter value is assumed as the initial condition for the dynamical variable [*TGF- β*].¹⁶ The measures are obtained by Affymetrix one color microarray technology and make use of the platform “U133 Plus 2.0 Array”. The gene expression values of *hmg2* are detected through 9 probes which match the mRNA of the *hmg2* gene. Each probe emits a fluorescence intensity, reflecting the actual mRNA concentration. The temporal behavior is resolved by extracting a limited amount of RNA at successive times, and repeating the microarray analysis. The experimental points are displayed in figure 2 with filled circles: the concentration of mRNA (associated to gene *hmg2*) grows initially and then saturates to an equilibrium value, after about one day from the time of the initial exposure to *TGF- β* .

Starting from this setting one can adjust the free parameters of the model to eventually interpolate the experimental profiles. To this end, several strategies have been implemented (the Levenberg-Marquardt algorithm, the Nelder-Mead algorithm and Port algorithm) and shown to consistently return the same solution, which is therefore interpreted as a global minimum of the scrutinized problem. The fitted profile is depicted in figure 2 (solid line) and accurately interpolates the experimental points, as confirmed by direct visual inspection. The best fit parameters are respectively $k_p = 0.00003 \text{ s}^{-1}$ and $k_s = 0.01 \text{ s}^{-1}$ (as concerns the scaling factor, we get the best fit value $\alpha = 3.446$). The former parameter is definitely small, as compared to the other reaction rates involved. This observation agrees qualitatively with the independent estimate of Schmierer *et al.*,³ where the parameter k_p controls however a slightly modified chemical equation, as compared to the one here

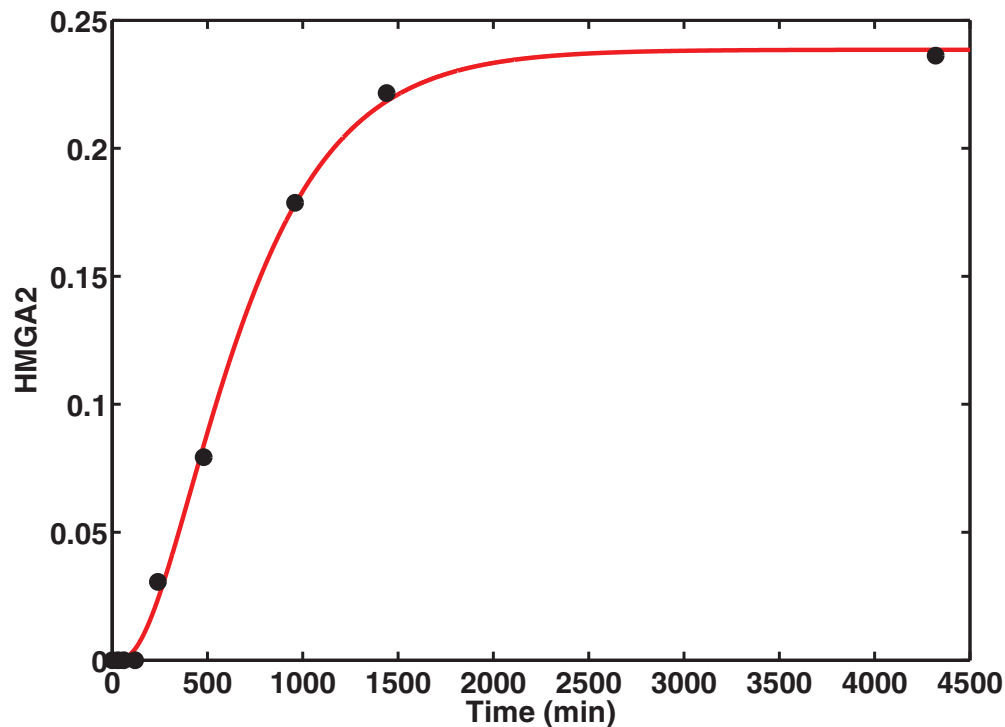


FIG. 2. The experimental level of expression relative to the mRNA quota associated to the *hmga2* gene, is plotted (symbols, filled circles) versus time. The experiments represent microarray investigations, repeated at different times of observation. The solid line stands for the theory solution obtained by tuning the parameters k_p , k_s and α so to optimize the matching with the experimental data points. The best fit parameters are found to be $k_p = 0.00003s^{-1}$, $k_s = 0.01s^{-1}$ and $\alpha = 3.446$.

hypothesized. At variance, to the best of our knowledge, this is the first time that the k_s parameter is being quantified.

Before turning to elaborate on the robustness of the derived estimates, a task that we will later on accomplish by means of dedicated statistical analysis tools, let us further characterize the predictions of the model. In doing so, we are mainly concerned with testing the predictive ability of the model versus other experimental data, independent from the ones exploited in the calibration procedure. As we will see in the next section, the theory predicts the correct behavior of the phosphorylation kinetic of Smad proteins, without no additional fitting parameters.

A. The model predicts the phosphorylation kinetic of the Smad proteins

As previously anticipated, we are here interested in further evaluating the model, and critically test its predictive performances versus existing experimental observations. It should be emphasized that the model is now fully constrained, no additional parameters modulation being de facto allowed.

Consider for instance the kinetic of Smad proteins phosphorylation kinetics. This is an extremely important indicator in *TGF- β* signaling: it senses the nuclear import/export processes, and indirectly controls the transcriptional activity/stability.¹⁷ In figure 3, panel A, we report the time evolution dynamics of $[_pS_c]$ and $[_pS_n]$, respectively solid and dashed line. Both curves present a clear peak after about 6 hours from *TGF- β* injection. Then, the concentrations of phosphorylated (cytoplasmatic and nuclear) Smad decay and eventually fade off in about 48 hours. In figure 3, panel B, the cumulated (cytoplasmatic plus nuclear) concentration of Smad proteins is reported using a different graphical representation that is meant to closely mimic a typical western blotting output. A color code is assigned and the concentration measured at different times plotted accordingly. In particular, we acquire, and hence display, the signal at times 10 minutes and 1, 8, 16, 24, 48, 72 hours. This choice

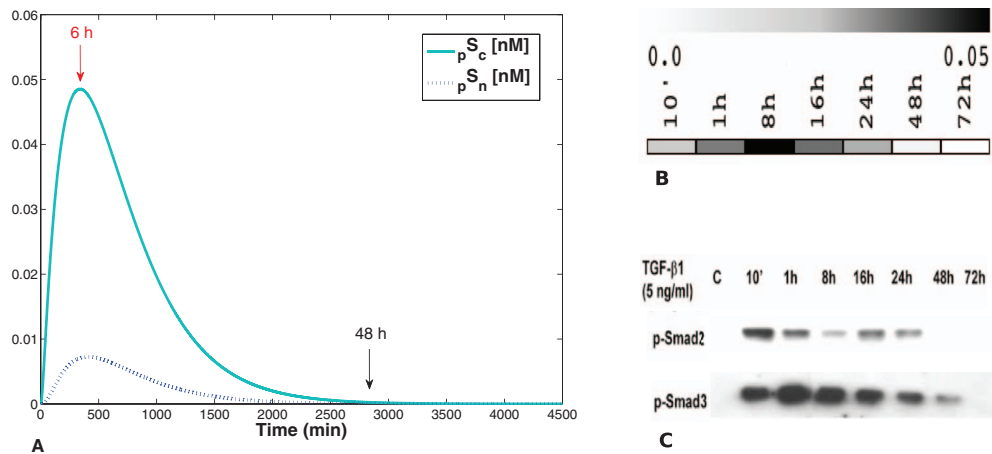


FIG. 3. Panel A: The phosphorylation kinetic of the Smad proteins as predicted by the model. The solid line stands for the concentration of cytoplasmic Smad. The dashed line refers instead to the nuclear Smad. A peak in the concentration is displayed after about 6 hours from the insertion of $TGF-\beta$. Panel B: the cumulative concentration of Smad (cytoplasmic + nuclear) is depicted using a different graphical representation. Darker patches are associated to higher concentration values (see color code annexed as a legend). The signal is acquired after 10 minutes and 1, 8, 16, 24, 48, 72 hours. Panel C: Western immunoblotting results adapted with permission from Venkateshwar G. Keshamouni *et al.*, "Differential Protein Expression Profiling by iTRAQ-2DLC-MS/MS of Lung Cancer Cells Undergoing Epithelial-Mesenchymal Transition Reveals a Migratory/Invasive Phenotype," *Journal of Proteome Research* 5, 1143-1154 (2006). Copyright (2006) by the American Chemical Society. The time evolution of both pS_2 and pS_3 (the two units that result into pS , as defined in the model) is reported. The comparison between theory and prediction is satisfying (see also comments in the main body of the paper).

is simply motivated by the need to facilitate the comparison with the experiments of Keshamouni and colleagues¹³ where western immunoblotting measurements were performed using the same cell line (A549) as employed in the microarray investigations of Jin J.Y. *et al.*,¹⁰ and in response to the exactly the same quantity of administered $TGF-\beta$. The western immunoblotting results reported in Keshamouni *et al.*¹³ are enclosed for convenience in Fig.3, panel C. Indeed, the experiments monitor the concentrations of both pS_2 and pS_3 species (with no distinction between cytoplasmic and nuclear contribution), the two phosphorylated components that, following successful encounters, result in the complex here object of investigations. Several comments are mandatory at this point. When comparing panel B and C of Fig. 3, one can appreciate a qualitative degree of correspondence between experimental and theoretical profiles. In both cases, the concentrations of the cumulated Smad grow and then disappear in about 48 hours. The experimental peak seems to slightly precede the theoretical one. Indeed, from the experiments, one can solely argue that the maximum in the Smad concentration is probably located between 1 and 8 hours, a fact that agrees, within the accessible accuracy, with the theoretical prediction that positions the peak at 6 hours. As a side remark, we recall that the model looks at the behavior of the complex, formed by the two units pS_2 and pS_3 , and termed pS . The modest shift (if any) between the two maxima can be therefore interpreted as the natural time delay, that has to be allowed for the system to eventually establish the sought complex. Finally, we also reconstructed the Smad dynamics working within the context of the model in Schmierer *et al.*³ (data not shown). The qualitative profile is correctly reproduced, but both the location of the peak and the global time of convergence to the asymptotic steady state regime appear to be significantly overestimated.

V. STATISTICAL ANALYSIS

In this section we shall complement the fitting procedure with a statistical based analysis aimed at assessing the parameters sensitivity and model robustness.¹⁸ In other words, we wish to quantitatively evaluate the system response as follows an imposed modulation of key control parameters.

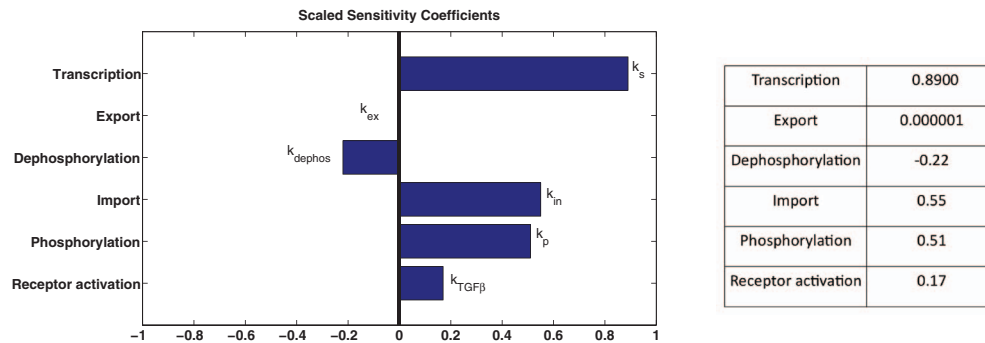


FIG. 4. The results of the sensitivity analysis are displayed. Left panel: the bars represent the scaled sensitivity coefficients, calculated for each of the considered reaction parameters. The table on the right reports the values of each scaled coefficients.

A. Sensitivity analysis

The susceptibility of the mRNA concentration of *hmga2* gene to the variation of the model's parameter is here inspected. Define $[X]$ as the mRNA concentration of *hmga2* gene and K the scalar control parameter. The sensitivity coefficient is calculated as $S = \frac{\partial [X]}{\partial K} \frac{K}{[X]}$. The calculation is performed numerically by means of the FME package¹⁹ made available in R.²⁰ The results reported in fig 4 allow us to appreciate a degree of positive correlation between the mRNA concentration of *hmga2* gene on the one side and the transcriptional rate (k_s), the phosphorylation rate (k_p) and nuclear import rate (k_{in}), on the other. Conversely, a negative interference is observed with the dephosphorylation rate (k_{dephos}): by increasing the latter, the concentration of $[hmga2]$ gets reduced. These tendencies, clearly consistent with the mathematical scheme, as one can readily qualitatively appreciate, are quantitatively characterized through the coefficients enclosed in the Table of fig 4. We notice in particular that the model is very robust to changes in the export rate (k_{ex}).

B. Markov chain Monte Carlo analysis

As a final point, we addressed the problem of parameters estimation in a statistical framework by using a Bayesian Markov Chain Monte Carlo technique.¹⁹ This technique enables one to reconstruct a posteriori the parameter distributions (not just the associated mean expected values), accounting for the noise component which certainly affects the experimental data.²¹ In our setting, we use the MCMC analysis to calculate the best values of the parameters k_p and k_s and their associated probability distribution. In Fig 5, the results of the analysis are displayed (in log-lin scale). The average values agree with those obtained by means of conventional optimization algorithms.

The analysis is again performed by using the FME package.¹⁹

VI. DISCUSSION

The EMT is an important biological process which induces the cells to change their phenotype. This process, although implicated in physiological conditions, plays also a key role in cancer pathogenesis.²² In cancer, in fact, the acquisition of the mesenchymal phenotype makes it possible for the cells to migrate toward nearby organs and so promote metastasis. From a molecular and cellular viewpoint, no specific features have been pinpointed that allow to distinguish physiological from pathological EMT. Several experiments established that, in both cases, the *TGF- β* pathway is directly involved in the activation of EMT. Starting from this setting, one needs to gain insight into the implicated processes to possibly help developing novel strategies for clinical oncology practice. In this paper we present a mathematical model which describes quantitatively the dynamics of *TGF- β* pathway in cancer cells during EMT. The model is constrained to fit experimental microarray dataset, a task that it is accomplished by adjusting a limited number of free parameters. The degree of correspondence between the experimental time series and the interpolating theoretical profile

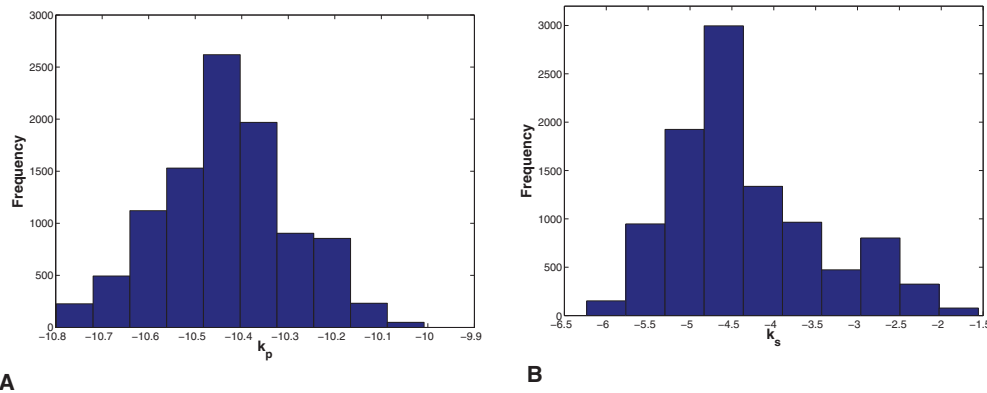


FIG. 5. Results of the MCMC analysis. In panels A and B the distributions of the parameters k_p and k_s is respectively displayed in log-lin scale. Here, log stands for natural logarithm. The best values of the parameters as calculated via the MCMC technique, are $k_p = 0.0000289 \text{ nM}^{-1} \text{ s}^{-1}$ $k_s = 0.012 \text{ s}^{-1}$.

is indeed excellent. Importantly, the model predicts the kinetic of phosphorylation of the Smad proteins, with no further adjustments in the parameters involved. We elaborate on the correctness of the prediction, benchmarking our theory to the existing experimental literature. Again, the correspondence is satisfying, a result which in turn provides an a posteriori validation of the proposed mathematical model. The model calibration step is complemented by a dedicated statistical based analysis to assess the sensitivity and robustness of the assigned parameters. A Bayesian Markov Chain Monte Carlo is also performed and shown to return a generalized scenario which agrees with that obtained working with more conventional optimization protocols.

ACKNOWLEDGMENTS

This work was supported by the Association for International Cancer Research (grant no. 06-0491), Associazione Italiana per la Ricerca sul Cancro (grant no. 1662), and Istituto Toscano Tumori (DD Regione Toscana no. 6888) to A.A., Ente Cassa di Risparmio di Firenze and PRIN2009 to D.F.. The Associazione Genitori contro le Leucemie e Tumori Infantili “Noi per Voi” for the financial support to A.A. and P.L. is also acknowledged.

- ¹ Jean Paul Thiery, Hervé Acloque, Rubi Y. J. Huang, and M. Angela Nieto, “Epithelial-mesenchymal transition in development and disease,” *Cell* **21**, 166-176 (2009).
- ² JCarl-Henrik Heldin, Maréne Landström, and Aristidis Moustakas, “Mechanism of TGF- β signaling to growth arrest, apoptosis, and epithelial mesenchymal transition,” *Current Opinion in Cell Biology* **21**, 166-176 (2009).
- ³ Bernhard Schmierer, Alexander L. Tournier, Paul A. Bates, and Caroline S. Hill, “Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system,” *PNAS* **105** no. 18, 6608-6613 (2008).
- ⁴ D. C. Clarke, M. D. Betterton, and X. Liu, “Systems theory of Smad signalling,” *IEE Proc.-Syst. Biol.* **153** no. 6, 412-424 (2006).
- ⁵ Zhike Zi and Edda Klipp, “Constraint-Based Modeling and Kinetic Analysis of the Smad Dependent TGF- β Signaling Pathway,” *PLOS one* **9**, e936 (2007).
- ⁶ Jose M. G. Vilar, Ronald Jansen, and Chris Sander, “Signal Processing in the TGF-b Superfamily Ligand-Receptor Network,” *PLOS Computational Biology* **1**, e3 (2006).
- ⁷ Linsey E. Lindley and Karoline J. Briegel, “Molecular characterization of TGFb-induced epithelial-mesenchymal transition in normal finite lifespan human mammary epithelial cells,” *Biochemical and Biophysical Research Communications* **399**, 659-664 (2010).
- ⁸ Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown, “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray,” *Science* **270** no. 5235, 467-470 (1995).
- ⁹ Ting Liu and Xin-Hua Feng, “Regulation of TGF- β signalling by protein phosphatases,” *Biochem. J.* **430**, 191-198 (2010).
- ¹⁰ Jin J. Y., Richard R. Almon, Debra C. Dubois, and William J. Jusko, “Modeling of Corticosteroid Pharmacogenomics in Rat Liver Using Gene Microarrays,” *The journal of pharmacology and experimental therapeutics* **307** no. 1, 93-109 (2003).

- ¹¹ Zhenling Yao, Eric P. Hoffman, Svetlana Ghimbovski, Debra C. DuBois, Richard R. Almon, and William J. Jusko, "Mathematical Modeling of Corticosteroid Pharmacogenomics in Rat Muscle following Acute and Chronic Methylprednisolone Dosing," *Molecular Pharmaceutics* 5 no. 2, 328-339 (2008).
- ¹² Tianhai Tian, Songlin Xu, Junbin Gao, and Kevin Burrage, "Simulated maximum likelihood method for estimating kinetic rates in gene expression," *Bioinformatics* 23, 84-91 (2007).
- ¹³ Venkateshwar G. Keshamouni, George Michailidis, Catherine S. Grasso, Shalini Anthwal, John R. Strahler, Angela Walker, Douglas A. Arenberg, Raju C. Reddy, Sudhakar Akulapalli, Victor J. Thannickal, Theodore J. Standiford, Philip C. Andrews, and Gilbert S. Omenn, "Differential Protein Expression Profiling by iTRAQ-2DLC-MS/MS of Lung Cancer Cells Undergoing Epithelial-Mesenchymal Transition Reveals a Migratory/Invasive Phenotype," *Journal of Proteome Research* 5, 1143-1154 (2006).
- ¹⁴ Shi Y and Massague J, "Mechanisms of TGF-beta signaling from cell membrane to the nucleus," *Cell* 113, 685-700 (2003).
- ¹⁵ Feng XH and Derynck R, "Specificity and versatility in TGF- β signaling through Smads," *Annu Rev Cell Dev Bio* 21, 659-693 (2005).
- ¹⁶ Incidentally, we are now in a position to clarify why the families of fixed points (i), (ii), (iii) are not compatible with the selected initial data. First, let us note that $TGF-\beta$ and R get progressively consumed at exactly the same pace, see the first two equations of system (7), as they are both implicated in the creation of the R^* species. However, at $t = 0$, the quota of free receptors R is significantly larger than the number of injected $TGF-\beta$ molecules ($R(t = 0) = 1nM$ vs. $TGF-\beta(t = 0) = 0.113nM$). As a consequence, and because R and $TGF-\beta$ obey to an identical kinetic, it is the $TGF-\beta$ that vanish first, reaching its asymptotic state $TGF-\beta = 0$ when a residual quota of R is still present. From here on, the receptors R cannot be mutated in R^* any longer and are therefore indefinitely frozen to the asymptotic value $R(t = 0) - TGF-\beta(t = 0)$, which is positive and different from zero. Both solutions (i) and (ii) are therefore to be rejected because they require $R = 0$. A similar reasoning can be invoked to exclude solution (iii). This latter would in fact imply $S_c = 0$. However, as it can be readily appreciated by inspection of system (7), the rate of loss of both R^* and S_c is governed by the same term, namely $-k_p[R^*][S_c]$. The maximum amount of bound receptors R^* is equal to the number of $TGF-\beta(t = 0)$ molecules, while the S_c elements at time $t = 0$ are definitely many more ($121.1nM$). Moreover, the population of S_c gets also re-integrated, via a source term controlled by the reaction rate k_{exp} , which acts as long as S_n is different from zero. Based on the above, we can therefore conclude that S_c molecules are still present when R^* becomes zero. From here on the S_c cannot decrease any longer and stay frozen to the value that they have eventually attained when the condition $R^* = 0$ is met. Hence, solution (iii) cannot apply to the scrutinized setting, as it assumes the asymptotic condition $S_c = 0$.
- ¹⁷ Katharine H Wrighton, Xia Lin, and Xin-Hua Feng, "Phospho-control of TGF- β superfamily signaling," *Cell Research* 19, 8-20 (2009).
- ¹⁸ Andreas Wagner, "Robustness and Evolvability in Living Systems," Princeton University Press (2007).
- ¹⁹ Karlina Soetaert and Thomas Petzoldt, "Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME," *Journal of Statistical Software* 33(3), 1-28 (2010).
- ²⁰ R Development Core Team, Vienna, Austria, "R: A Language and Environment for statistical Computing," ISBN 3-900051-07-0, (<http://www.R-project.org>).
- ²¹ Daniel Sorensen and Daniel Gianola, "Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics," Springer, New York (2002).
- ²² Kornelia Polyak and Robert A. Weinberg, "Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits," *Nature Reviews* 9, 265-273 (2009).