Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2011, Article ID 572876, 5 pages
doi:10.1155/2011/572876

Research Article

Inference of Kinetic Parameters of Delayed Stochastic Models of Gene Expression Using a Markov Chain Approximation

Henrik Mannerstrom,[1] Olli Yli-Harja,[1,2] and Andre S. Ribeiro[1]

[1] *Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland*
[2] *Institute for Systems Biology, Seattle, WA 98103, USA*

Correspondence should be addressed to Henrik Mannerstrom, henrik.mannerstrom@tut.fi

Received 21 October 2010; Accepted 4 December 2010

Academic Editor: Carsten Wiuf

We propose a Markov chain approximation of the delayed stochastic simulation algorithm to infer properties of the mechanisms in prokaryote transcription from the dynamics of RNA levels. We model transcription using the delayed stochastic modelling strategy and realistic parameter values for rate of transcription initiation and RNA degradation. From the model, we generate time series of RNA levels at the single molecule level, from which we use the method to infer the duration of the promoter open complex formation. This is found to be possible even when adding external Gaussian noise to the RNA levels.

1. Introduction

Gene expression dynamics is influenced by even small fluctuations on the levels of various molecular species, such as RNA polymerases and transcription factors. In some cases, even the presence of a single molecule can cause phenotypic switching [1]. This makes the cellular metabolism inherently stochastic [2].

The stochasticity in the abundance of a substance is in general thought of being noise that obscures a signal that carries information relevant to the cell. However, recent evidence suggests that cells may be able to use the noise component in benefit of their survival [3]. Due to this, several modelling strategies have been proposed for accurately accounting for noise in the dynamics of gene regulatory networks (GRNs) [2, 4–7].

The chemical master equation is a probabilistic description of the dynamics of interacting molecules that fully captures the stochasticity of their kinetics. However, it is intractable to solve in the biologically relevant cases.

The stochastic simulation algorithm [8] (SSA) is a Monte Carlo simulation of the chemical master equation, allowing the study of complex models of gene expression. In the SSA, all chemical reactions are assumed instantaneous. However, several processes during the transcription and translation of a gene are highly complex, either involving many molecular species or involving reactions that are not bimolecular (e.g., the promoter open complex formation). To account for the effects of these events on the dynamics of RNA and proteins, the delayed SSA (DSSA) was proposed [5]. The ability of the DSSA to model chemical reactions with noninstantaneous events makes it a good tool to model GRN [6].

Assessing a model's accuracy and validity is important [9]. Even if experimental data has been used in model building, one must also be able to quantitatively rank the models based on the data. This ranking can be used to determine realistic parameter values, if these have not been measured directly, and to choose between models. As single molecule measurements of gene expression are becoming available [10], even the most detailed stochastic models can now be ranked.

Inference methods have been proposed to assess stochastic models of gene expression based on the SSA [11, 12]. Such methods are still lacking for the DSSA. Here, we present a method that, while requiring additional developments for analyzing complex gene networks, can be used to determine underlying features of single gene expression when simulated by the DSSA.

One feature in gene expression that has been proposed to influence noise in RNA and protein levels is the promoter open complex formation [13]. We use the proposed method to determine the duration of the promoter open complex formation from the dynamics of RNA levels of a delayed stochastic model of transcription.

2. Methods

2.1. Stochastic and Delayed Stochastic Simulation Algorithms. The Stochastic Simulation Algorithm (SSA) is a Monte Carlo simulation of the chemical master equation and, thus, is an exact procedure for numerically simulating the time evolution of a well-stirred reacting system [8]. Each chemical species quantity is treated as an independent variable, and each reaction is executed explicitly. Time is advanced by stepping from one reaction event to the next. At each step, the number of molecules of each affected species is updated according to the reaction formula.

For each reaction r, the stochastic rate constant, c_r, depends on the reactive radii of the molecules involved in the reaction and their relative velocities. The velocities depend on the temperature and molecular masses. After setting the initial species populations, X_i, the SSA calculates the propensities $a_r = c_r \cdot h_r$, for all possible reactions, where h_r is the number of distinct molecular reactants combinations available at a given moment. Then, it generates two random numbers, $\tau \sim \text{Exp}(\sum a_r)$, the time until the next reaction occurs, and μ, the reaction to occur. The probability for $\mu = r$ is $a_r / \sum a_r$. Finally, the system time t is increased by τ, and the X_i quantities are adjusted to account for the occurrence of reaction μ, assuming it to be an instantaneous reaction. This process is repeated until no more reactions can occur or for a defined time interval.

Several steps in gene expression, such as transcripts assembly, are time consuming [14]. Such complex processes involve many reactions and events that cannot be modelled as uni- or bimolecular reaction events. To account for these events, the "delayed SSA" was proposed [5]. It uses a "waitlist" to store delayed output events. Multidelayed reactions are represented as $A \rightarrow B + C(\tau_1) + D(\tau_2)$. In this reaction, B is instantaneously produced and C and D are placed on a waitlist until they are released, after τ_1 and τ_2 seconds, respectively.

The delayed SSA proceeds as follows.

(1) Set $t = 0$, $t_{\text{stop}} =$ stoptime, set initial number of molecules and reactions, and create empty waitlist L. Go to step (2).

(2) Generate an SSA step for reacting events to get the next reacting event R_1 and the corresponding occurrence time $t + t_1$. Go to step (3).

(3) Compare t_1 with the least time in L, t_{min}. If $t_1 < t_{\text{min}}$ or L is empty, set: $t = t + t_1$. Update the number of molecules by performing R_1, adding to L both any delayed products and the time delay for which they have to stay in L. This time can be chosen from a defined distribution. Go to step (4).

(4) If L is not empty and if $t_1 \geq t_{\text{min}}$, set $t = t + t_{\text{min}}$. Update the number of molecules and L, by releasing the first element in L; otherwise go to step (5).

(5) If $t < t_{\text{stop}}$, go to step (2); otherwise stop.

2.2. Delayed Stochastic Model of Transcription. A delayed stochastic model of transcription that includes the promoter open complex formation was proposed in Ribeiro et al. [6]. This model was shown to match the dynamics of transcription at the single RNA molecule level [15].

Our model is identical, except that it does not include an explicit representation of the RNA polymerase. This simplification is valid when the number of RNA polymerases does not vary significantly over time in the cell, which is likely to be the case in normal conditions in *E. coli* (Reaction (1)):

$$\text{Pro} \xrightarrow{k_t} \text{Pro}(\tau_{\text{Pro}}) + \text{RNA}(\tau_{\text{RNA}}), \tag{1}$$

$$\text{RNA} \xrightarrow{k_d} \varnothing. \tag{2}$$

In Reaction (1), Pro (set to 1 in the begin of the simulation) is the promoter region of the gene while k_t is the stochastic rate constant of transcription initiation and its value is set to $0.5\,\text{s}^{-1}$. This value assumes that the number of RNA polymerases available for transcription is always 40 [6] and that the binding affinity between RNA polymerase and transcription start site equals the one measured for the lac promoter [16]. The promoter delay, τ_{Pro}, is set to 40 s, in agreement with measurements for the lac Promoter [17]. Also, RNA stands for a fully transcribed RNA molecule, and τ_{RNA} is the time that it takes for the transcription process to be completed, once initiated. This delay accounts for the promoter open complex formation (40 s), transcription elongation (mean value 60 s), and termination. Its value is randomly generated from a Gaussian distribution with a mean of 102 s and a standard deviation of 14 s. These values assume a lac promoter and a gene 2445 nucleotides long [16, 18].

Note that while Reaction (1) has a rate of k_t, each activation cycle includes the open complex formation delay of τ_{Pro} seconds, making the effective mean cycle duration equal to $k_t^{-1} + \tau_{\text{Pro}}$.

Reaction (2) models RNA degradation. k_d is the rate of degradation and is set to $0.0017\,\text{s}^{-1}$ (10 min mean lifetime), which is within realistic parameter values for *E. coli* [19].

In Figure 1 are shown, as examples, levels of RNA molecules produced by independent simulations. The simulator ran for 6000 s from which the data from the last 3000 s was used as "steady state" data.

2.3. Approximative Inference. The system is approximated as a Markov chain with stationary distribution P and transition matrix T. As we are only considering steady state conditions, P and T can be built by thoroughly sampling ($\approx 1 \times 10^5$ samples) from the simulated model. To compensate for the sampling error both P and T are "smeared out" with a kernel of $N(0, 0.2)$. For example, if the raw sampling yields $T_\theta(i, j) = p$, then after the smearing $T_\theta(i, j) = 0.98p$, $T_\theta(i, j-1) = 0.0062p$, $T_\theta(i, j+1) = 0.0062p$.

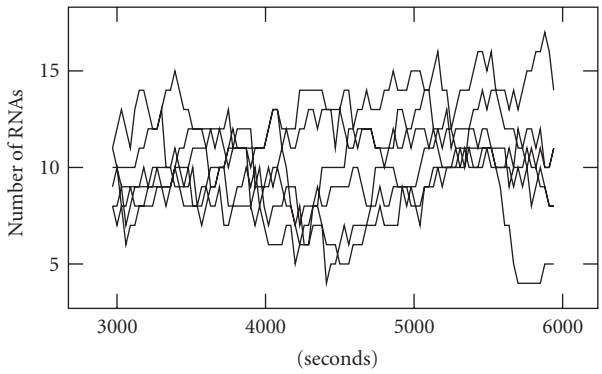

Figure 1: RNA levels from six independent simulations.

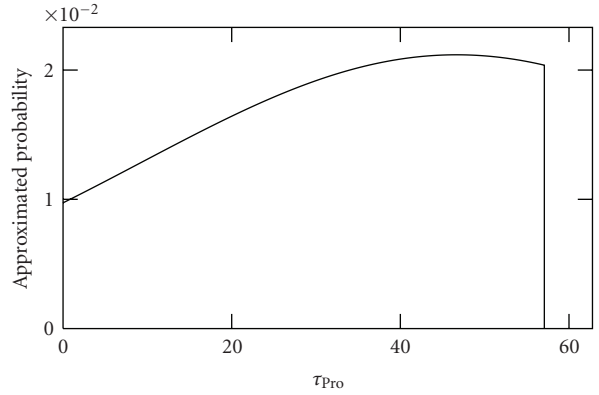

Figure 2: Approximated probabilities for values of τ_{Pro} inferred using simulated noiseless data from 10 cells. The true value is 40, the maximum likelihood value is 46.7 and the expected value is 31.8.

The log likelihood $L(\theta; X)$ of the parameter $\theta = (\tau_{\text{Pro}})$, given a time series X can then be computed by

$$\log L(\theta; X) = \log P_\theta(X_1) + \sum_{i=1}^{N} \log T_\theta(X_i, X_{i+1}), \qquad (3)$$

where X_i is the RNA level at time i.

The likelihood term is evaluated at suitable points over the full range of possible τ_{Pro} values, ranging from zero to the maximum determined by dividing the mean RNA life time by the mean RNA level (in our case study, this ratio around 60). Due to the approximation of P_θ and T_θ, the likelihood term will be nonsmooth and cannot be used as such. Instead, a quadratic polynomial is fitted to the point samples. The quadratic fit was chosen because it gives a likelihood proportional to a truncated normal distribution. Similar to the application of Bayes' theorem with a flat, non informative prior, the likelihood is converted to a probability distribution by normalizing it to unit probability.

2.4. Error Model. To simulate measurement error, normally distributed noise with zero mean and 0.5 standard deviation was added to the simulated time series used for inference. Any negative values were zeroed.

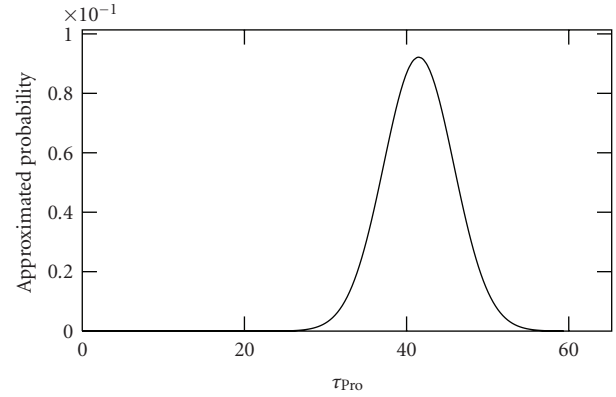

Figure 3: Approximated probabilities for values of τ_{Pro} inferred using simulated noiseless data from 100 cells. The true value is 40 and the expected value is 41.5.

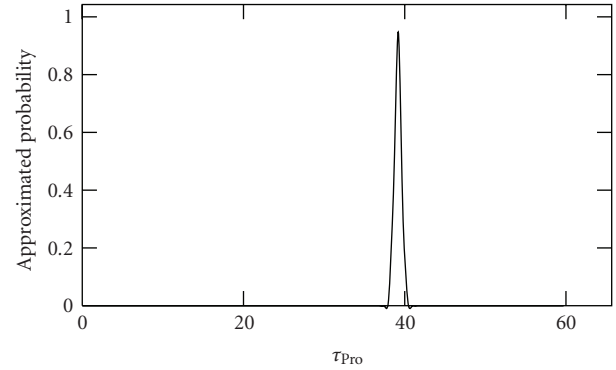

Figure 4: Approximated probabilities for values of τ_{Pro} inferred using simulated noiseless data from 1000 cells. The true value is 40 and the expected value is 39.2.

3. Results

In all simulations we set the sample interval to 30 s, as this is currently the shortest interval possible in real measurements of RNA numbers at the single molecule level [10]. The inference was made using these point samples.

We applied the method to sample sizes of 10, 100, and 1000 independent time series of length 2970 s (100 time points). As no external noise sources are applied to these data, we refer to it as "noiseless" data. Results are shown in Figures 2, 3, and 4, respectively. As seen, as the sample size is increased, the better becomes the inference of the true value of τ_{Pro}.

Interestingly, as seen from these results, using this method it is possible to show, even using a small sample size of 10, that the time length of the promoter open complex formation measurably affects the dynamics of RNA levels as previously shown by confronting numerical simulations with a null model [13].

We now test the robustness of the method to experimental measurement error. For this, to the previous time series we add Gaussian noise "noisy data" as described in the Methods section. Results of the inference, using 10, 100

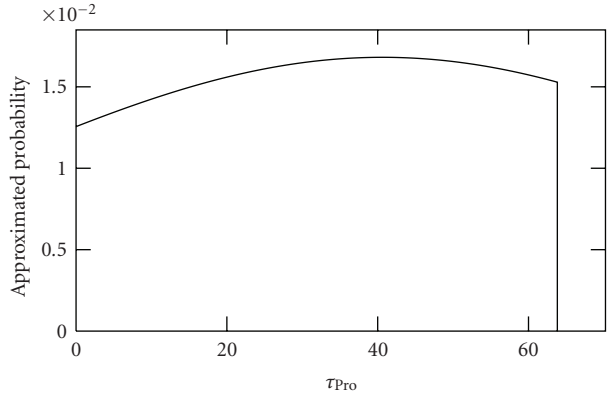

Figure 5: Approximated probabilities for values of τ_{Pro} inferred using simulated noisy data from 10 cells. The true value is 40, the maximum likelihood value is 40.6 and the expected value is 32.9.

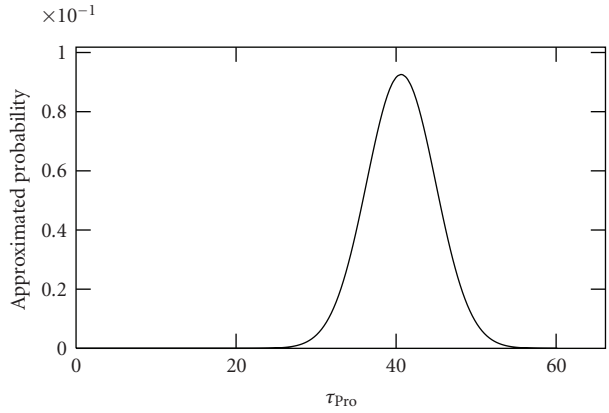

Figure 6: Approximated probabilities for values of τ_{Pro} inferred using simulated noisy data from 100 cells. The true value is 40 and the expected value is 40.6.

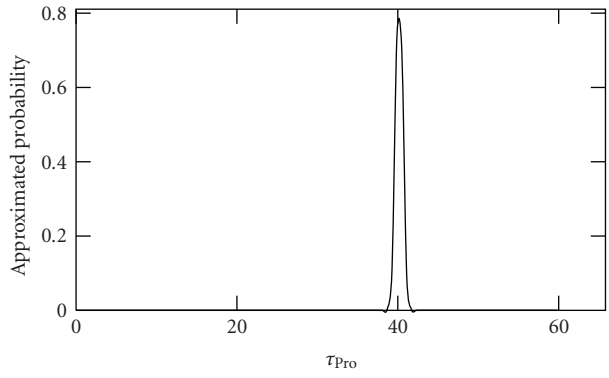

Figure 7: Approximated probabilities for values of τ_{Pro} inferred using simulated noisy data from 1000 cells. The true value is 40 and the expected value is 40.6.

and 1000 time series, are shown in Figures 5, 6, and 7, respectively. As the results show, the accuracy of the method is not significantly affected when the standard deviation of the external noise is in the range 0 to 0.5. If the noise level in the data is increased beyond this, the results become biased.

Finally, we note that using 1000 time series for the inference procedure, the method takes 15 min to be completed on a contemporary personal computer.

4. Conclusions

We tested an inference method for inferring, from time series data, kinetic parameters affecting the dynamics of RNA levels subject to degradation. When inferring the duration of the promoter open complex formation, we showed that, for known values of the RNA degradation rate, the method is accurate and fast. When a reasonable amount of noise is added to the data the performance is not significantly affected.

The inference was shown possible when considering only one previous sample point, by approximating it with a time-homogeneous Markov chain. This is especially relevant as, in *E. coli*, most RNA mean levels are from 1 to a few [19], implying that the system may have very little memory of far past events.

While experimentally challenging, it is already possible to collect time series of RNA levels of living cells close to the accuracy assumed by the model. This can be done using a technique that is based on the ability of the MS2d-GFP protein complex to bind to a target RNA [20]. This system possesses some limitations, such as the need to maintain weak transcription rate so as to distinguish individual RNA molecules [10].

While the present approximative method proposed is still far from an analytical likelihood, it can serve as a crude statistical tool to analyze experimental time series data. In the future, we aim to extend this method to infer other kinetic parameters associated with the dynamics RNA and protein levels in prokaryotes. Also, we will apply this method to determine from real measurements of RNA levels, if these are influenced by currently unknown processes.

Acknowledgment

This work was supported by Academy of Finland and FiDiPro program of Tekes.

References

[1] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie, "A stochastic single-molecule event triggers phenotype switching of a bacterial cell," *Science*, vol. 322, no. 5900, pp. 442–446, 2008.

[2] H. H. McAdams and A. Arkin, "It's a noisy business! Genetic regulation at the nanomolar scale," *Trends in Genetics*, vol. 15, no. 2, pp. 65–69, 1999.

[3] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, "Stochasticity in gene expression: from theories to phenotypes," *Nature Reviews Genetics*, vol. 6, no. 6, pp. 451–464, 2005.

[4] D. Bratsun, D. Volfson, L. S. Tsimring, and J. Hasty, "Delay-induced stochastic oscillations in gene regulation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 41, pp. 14593–14598, 2005.

[5] M. R. Roussel and R. Zhu, "Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression," *Physical Biology*, vol. 3, no. 4, pp. 274–284, 2006.

[6] A. Ribeiro, R. Zhu, and S. A. Kauffman, "A general modeling strategy for gene regulatory networks with stochastic dynamics," *Journal of Computational Biology*, vol. 13, no. 9, pp. 1630–1639, 2006.

[7] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008.

[8] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[9] D. J. Wilkinson, "Stochastic modelling for quantitative description of heterogeneous biological systems," *Nature Reviews Genetics*, vol. 10, no. 2, pp. 122–133, 2009.

[10] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, "Real-time kinetics of gene activity in individual bacteria," *Cell*, vol. 123, no. 6, pp. 1025–1036, 2005.

[11] R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood, "Bayesian inference for a discretely observed stochastic kinetic model," *Statistics and Computing*, vol. 18, no. 2, pp. 125–135, 2008.

[12] Y. Wang, S. Christley, E. Mjolsness, and X. Xie, "Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent," *BMC Systems Biology*, vol. 4, article 99, 2010.

[13] A. S. Ribeiro, A. Häkkinen, H. Mannerström, J. Lloyd-Price, and O. Yli-Harja, "Effects of the promoter open complex formation on gene expression dynamics," *Physical Review E*, vol. 81, no. 1, Article ID 011912, 2010.

[14] K. Ota, T. Yamada, and Y. Yamanishi, "Comprehensive analysis of delay in transcriptional regulation using expression profiles," *Genome Informatics*, vol. 14, pp. 302–303, 2003.

[15] A. S. Ribeiro, "Stochastic and delayed stochastic models of gene expression and regulation," *Mathematical Biosciences*, vol. 223, no. 1, pp. 1–11, 2010.

[16] R. Zhu, A. S. Ribeiro, D. Salahub, and S. A. Kauffman, "Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models," *Journal of Theoretical Biology*, vol. 246, no. 4, pp. 725–745, 2007.

[17] W. R. McClure, "Rate-limiting steps in RNA chain initiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 10 II, pp. 5634–5638, 1980.

[18] JI. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie, "Probing gene expression in live cells, one protein molecule at a time," *Science*, vol. 311, no. 5767, pp. 1600–1603, 2006.

[19] J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao, and S. N. Cohen, "Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 15, pp. 9697–9702, 2002.

[20] D. Fusco, N. Accornero, B. Lavoie et al., "Single mRNA molecules demonstrate probabilistic movement in living mammalian cells," *Current Biology*, vol. 13, no. 2, pp. 161–167, 2003.

Preliminary call for papers

The 2011 European Signal Processing Conference (EUSIPCO-2011) is the nineteenth in a series of conferences promoted by the European Association for Signal Processing (EURASIP, www.eurasip.org). This year edition will take place in Barcelona, capital city of Catalonia (Spain), and will be jointly organized by the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC) and the Universitat Politècnica de Catalunya (UPC).

EUSIPCO-2011 will focus on key aspects of signal processing theory and applications as listed below. Acceptance of submissions will be based on quality, relevance and originality. Accepted papers will be published in the EUSIPCO proceedings and presented during the conference. Paper submissions, proposals for tutorials and proposals for special sessions are invited in, but not limited to, the following areas of interest.

Areas of Interest

- Audio and electro-acoustics.
- Design, implementation, and applications of signal processing systems.
- Multimedia signal processing and coding.
- Image and multidimensional signal processing.
- Signal detection and estimation.
- Sensor array and multi-channel signal processing.
- Sensor fusion in networked systems.
- Signal processing for communications.
- Medical imaging and image analysis.
- Non-stationary, non-linear and non-Gaussian signal processing.

Submissions

Procedures to submit a paper and proposals for special sessions and tutorials will be detailed at www.eusipco2011.org. Submitted papers must be camera-ready, no more than 5 pages long, and conforming to the standard specified on the EUSIPCO 2011 web site. First authors who are registered students can participate in the best student paper competition.

Important Deadlines:

Proposals for special sessions	15 Dec 2010
Proposals for tutorials	18 Feb 2011
Electronic submission of full papers	**21 Feb 2011**
Notification of acceptance	23 May 2011
Submission of camera-ready papers	6 Jun 2011

Webpage: www.eusipco2011.org

Organizing Committee

Honorary Chair
Miguel A. Lagunas (CTTC)

General Chair
Ana I. Pérez-Neira (UPC)

General Vice-Chair
Carles Antón-Haro (CTTC)

Technical Program Chair
Xavier Mestre (CTTC)

Technical Program Co-Chairs
Javier Hernando (UPC)
Montserrat Pardàs (UPC)

Plenary Talks
Ferran Marqués (UPC)
Yonina Eldar (Technion)

Special Sessions
Ignacio Santamaría (Unversidad de Cantabria)
Mats Bengtsson (KTH)

Finances
Montserrat Nájar (UPC)

Tutorials
Daniel P. Palomar (Hong Kong UST)
Beatrice Pesquet-Popescu (ENST)

Publicity
Stephan Pfletschinger (CTTC)
Mònica Navarro (CTTC)

Publications
Antonio Pascual (UPC)
Carles Fernández (CTTC)

Industrial Liaison & Exhibits
Angeliki Alexiou (University of Piraeus)
Albert Sitjà (CTTC)

International Liaison
Ju Liu (Shandong University-China)
Jinhong Yuan (UNSW-Australia)
Tamas Sziranyi (SZTAKI -Hungary)
Rich Stern (CMU-USA)
Ricardo L. de Queiroz (UNB-Brazil)