

eman ta zabal zazu



Universidad del País Vasco Euskal Herriko Unibertsitatea

Quantitative analyses in basic, translational and
clinical biomedical research: metabolism, vaccine
design and preterm delivery prediction

Iker Malaina

PhD Thesis

Department of Physiology

University of the Basque Country (UPV/EHU)

Bilbao, Spring 2017

Tome II


Contents of Tome II

Introduction	2
Dynamic properties of calcium-activated chloride currents in <i>Xenopus laevis</i> oocytes	3
On the dynamics of the adenylate energy system: homeostasis vs homeorhesis	17
A combinatorial approach to the design of vaccines	35
Montevideo units vs autoregressive models on preterm labor detection	67

Introduction

With the purpose of facilitating the consultation of methods, tables and other supplementary data, in this tome we include the four articles in which this thesis is based, thus complementing Tome I.

SCIENTIFIC REPORTS



OPEN

Dynamic properties of calcium-activated chloride currents in *Xenopus laevis* oocytes

Received: 22 July 2016
Accepted: 30 December 2016
Published: 13 February 2017

Ildefonso M. De la Fuente^{1,2}, Iker Malaina², Alberto Pérez-Samartín³, María Dolores Boyano⁴, Gorka Pérez-Yarza⁴, Carlos Bringas⁴, Álvaro Villarroel⁵, María Fedetz⁶, Rogelio Arellano⁷, Jesus M. Cortes^{4,8,9} & Luis Martínez²

Chloride is the most abundant permeable anion in the cell, and numerous studies in the last two decades highlight the great importance and broad physiological role of chloride currents mediated anion transport. They participate in a multiplicity of key processes, as for instance, the regulation of electrical excitability, apoptosis, cell cycle, epithelial secretion and neuronal excitability. In addition, dysfunction of Cl⁻ channels is involved in a variety of human diseases such as epilepsy, osteoporosis and different cancer types. Historically, chloride channels have been of less interest than the cation channels. In fact, there seems to be practically no quantitative studies of the dynamics of chloride currents. Here, for the first time, we have quantitatively studied experimental calcium-activated chloride fluxes belonging to *Xenopus laevis* oocytes, and the main results show that the experimental Cl⁻ currents present an informational structure characterized by highly organized data sequences, long-term memory properties and inherent “crossover” dynamics in which persistent correlations arise at short time intervals, while anti-persistent behaviors become dominant in long time intervals. Our work sheds some light on the understanding of the informational properties of ion currents, a key element to elucidate the physiological functional coupling with the integrative dynamics of metabolic processes.

Chloride (Cl⁻) is thought to be the most abundant free anion in the cell¹, and its movement through the cellular membranes is mainly mediated by Cl⁻ channels, which seem to be widespread in nearly all cellular organisms, from bacteria to mammals^{2,3}.

Chloride-conducting anion channels are localized both in the plasma membrane and in intracellular organelles such as the endoplasmic reticulum, the Golgi apparatus, the nucleus, the mitochondria, the lysosomes, the endosomes and the cell vesicles⁴⁻⁷. They participate in a multiplicity of key functions like, for instance, the stabilization of the membrane potential, the regulation of cell volume and electrical excitability, and the acidification of intracellular organelles^{4,8}. In addition, different studies have recognized the Cl⁻ channels' contributions to apoptosis⁹, signal transduction¹⁰, cell cycle¹¹, cell adhesion and motility¹², among other complex cellular processes.

Intracellular chloride currents also play important roles in a variety of physiological processes¹³, including epithelial secretion¹⁴, neuronal excitability¹⁵, repolarization of the cardiac action potential¹⁶, modulation of light responses¹⁷ and olfactory transduction¹⁸. It can be noted that, under physiological conditions, certain types of Cl⁻ channels participate in the regulation of the action potentials and synaptic responses, which are important for learning and memory¹⁹. In fact, dramatic changes in intracellular Cl⁻ currents occur both during development and in response to synaptic activity^{20,21}.

¹Department of Nutrition, CEBAS-CSIC Institute, Espinardo University Campus, Murcia, Spain. ²Department of Mathematics, Faculty of Science and Technology, University of the Basque Country, UPV/EHU, Leioa, Spain. ³Department of Neurosciences, Faculty of Medicine and Dentistry, University of the Basque Country, UPV/EHU, Leioa, Spain. ⁴Department of Cell Biology and Histology, Faculty of Medicine and Dentistry, University of the Basque Country, UPV/EHU, Leioa, Spain. ⁵Biophysics Unit, CSIC, University of the Basque Country, UPV/EHU, Leioa, Spain. ⁶Department of Biochemistry and Pharmacology, Institute of Parasitology and Biomedicine “López-Neyra”, CSIC, Granada, Spain. ⁷Laboratory of Cellular Neurophysiology, Neurobiology Institute, UNAM, Querétaro, México. ⁸BioCruces Health Research Institute, Cruces University Hospital, Barakaldo, Spain. ⁹IKERBASQUE: The Basque Foundation for Science, Bilbao, Spain. Correspondence and requests for materials should be addressed to I.M.d.l.F. (email: mtpmadei@ehu.eus)

At a protein metabolism level, there are numerous examples of proteins whose activity is dependent on, or regulated by Cl^- ^{22–24}. For instance, the $\text{Na}^+/\text{K}^+/\text{2Cl}^-$ cotransporter NKCC1 is activated by low intracellular Cl^- via a Cl^- -sensitive protein kinase²⁵.

The importance of chloride channels was also evidenced through studies of human diseases. In fact, the dysfunction of certain types of chloride channels is involved in a variety of diseases such as epilepsy, male infertility, cystic fibrosis, myotonia, lysosomal storage disease, deafness, kidney stones, and osteoporosis^{1,26,27}.

Moreover, different oncogenic processes such as the high rate of proliferation, active migration, and invasiveness of malignant cells into normal tissue have been shown to require the involvement of determined chloride channel activity in a variety of cancer types^{22,23}.

In general, some chloride channels are activated only by voltage i.e., voltage-gated, while others are activated by various ions e.g., H^+ (pH), or Ca^{2+} , or by the phosphorylation of intracellular residues by several protein kinases^{4,28}. Based on these and other characteristics, chloride channels have been classified into five main functional groups: (i) extracellular ligand-gated channels, (ii) calcium-activated chloride channels, (iii) volume-regulated anion channels, (iv) cAMP-PKA activated channels, and (v) voltage-gated chloride channels²⁹.

Calcium-activated chloride channels (CaCCs) are a key family of chloride channels that regulate the flow of chloride and other monovalent anions across cellular membranes in response to intracellular calcium levels³⁰. These channels are ubiquitously expressed, in both excitable and non-excitable cells³¹.

Currents mediated by CaCCs were first observed in 1981 in *Rana pipiens* eggs where the injection of Ca^{2+} initiated a transient shift to positive membrane potentials in a Cl^- -dependent manner³². Later studies in *Xenopus laevis* oocytes and salamander photoreceptors characterized these calcium-activated chloride currents^{33,34}.

The relationship between chloride currents and intracellular calcium fluctuations gives CaCCs a crucial role in many cellular processes, and numerous studies show the great importance and broad physiological role of these channels³⁵.

Historically, chloride channels have been less studied than cation channels. Considerable progress has been made in the knowledge of their molecular structures and functions³⁰, but there seems to be practically no quantitative studies of the dynamics of chloride currents. On the contrary, there are a significant number of studies made from the perspective of systems biology on free cations such as calcium. For instance, from the perspective of systems biology, different studies have shown that information might be encoded in the amplitude, the frequency, the duration, the waveform or the timing of the calcium oscillations^{36,37}. Moreover, the mutual information method was used to calculate the amount of information transferred through a calcium signaling channel³⁸ and long-term correlations were also observed in calcium-activated potassium channels³⁹.

Here, we present a pioneer quantitative study of the dynamic properties of the chloride currents belonging to calcium-activated chloride channels (CaCCs) of *Xenopus laevis* oocytes, analyzed under different external pH environments (acid, neutral and basic). *Xenopus* oocytes have long been a model system for studying CaCCs because these channels are the predominant channels expressed at extremely high levels (0.5 mA/cm²)⁴⁰.

The calcium-activated chloride currents were measured by the patch-clamp technique and the experimental series were analyzed by means of non-linear approaches. Our main result shows that the currents present a structure characterized by highly organized data sequences, long-term memory and inherent “crossover” dynamics with transitions from persistent to anti-persistent behaviors. In this dynamic structure, short memory time periods with a mean of 7.6 seconds arise from the experimental data, which correspond to non-trivial correlations that encompass around 4,000 experimental chloride values.

In this paper, for the first time, we have addressed essential aspects of calcium-activated chloride channels (CaCCs), and the informational properties herein analyzed seem to be intrinsic characteristics of the dynamics involved in these physiological ion currents.

Results

In order to study some of the dynamic properties of the chloride channels we have recorded calcium-activated chloride currents in *Xenopus laevis* oocytes, which have been evoked by serum under different external pH stimuli (pH = 0.5, pH = 0.7 and pH = 0.9). Thus, we had 21 time series in total, each one of them formed by 130,000 discrete data points. Figure 1 shows three representative experimental signals obtained by means of the patch-clamp technique, under three different pH conditions, Ringer’s solution at pH 5.0, 7.0 and 9.0 (acid, neutral and basic pH).

To confirm that oscillations monitored in *Xenopus* oocytes by application of Fetal Bovine Serum corresponded with Ca^{2+} -dependent Cl^- currents, three different experiments were performed. First, oocytes generating oscillations were voltage-clamped at 4 different voltages (either -60 , -40 , -20 or at 0 mV). As it is illustrated in Fig. 2a, currents reversed near to -20 mV, in accordance with the reversal potential of Cl^- in oocytes. Second, the reversal potential observed was shifted toward more positive potentials when the external Cl^- concentration was reduced, this is shown in Fig. 2b. In this case, oocytes were held to either -30 mV (first column) or 0 mV (second column), while they were superfused with solutions containing 100%, 50% or 0% of Cl^- (NaCl was substituted proportionally by Na_2SO_4 in Ringer solution and, osmolarity compensated adding sucrose). It is clear that reversal potential is close to -30 mV in 100% Cl^- , while in 0% Cl^- oscillations continued being in inward direction at 0 mV, indicating that reversal potential in this condition is more positive. An intermediate case occurs with 50% Cl^- solution, where the shift in reversal potential by reducing external Cl^- is predicted by the Nernst equation. And finally, it was demonstrated that Cl^- currents were Ca^{2+} -dependent. Intraoocyte injection of the calcium chelator ethylene glycol-bis(2-aminoethylether)N,N,N',N'-tetraacetic acid (EGTA) abolished completely oscillatory currents, according to Ca^{2+} -dependent Cl^- currents.

First, to test for the presence of long-term correlations in the experimental chloride data we have used the root-mean square (rms) fluctuation $F(l)$. For uncorrelated data, the exponent α for the relationship $F(l) \sim l^\alpha$ is equal to 0.5; in contrast $\alpha > 0.5$ indicates the presence of positive long-range correlations and $\alpha < 0.5$ implies

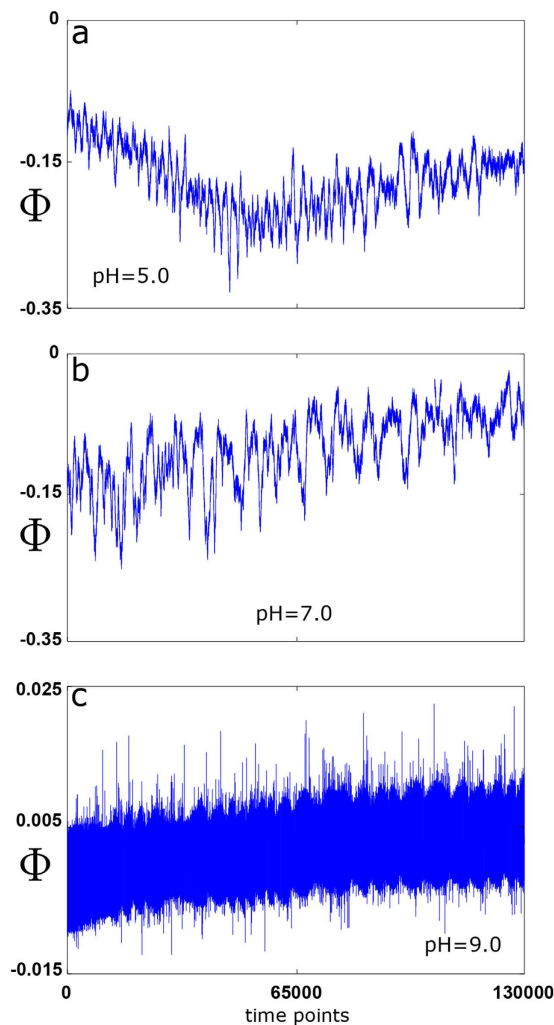


Figure 1. Calcium-activated chloride currents in *Xenopus laevis* oocyte. Three prototype experimental Cl^- currents obtained from the same cell at different conditions: (a) pH 5.0 (n10), (b) pH 7.0 (n11), (c) pH 9.0 (n12). Each chloride time series has 130,000 points (sampling interval 2 milliseconds), which correspond to a period of time of 260,000 milliseconds duration. The vertical axis (Φ) corresponds to the measures of currents in nanoamperes (nA).

long-term anti-correlations. According to this method, we have divided the 130,000 data points of each time series in 6 non-overlapping windows with $k = 5$, performing the rms fluctuation method on every window for each of the 21 experimental chloride series and fitting $F(l)$ within the range $l = 1, \dots, l_{max}$ (see Methods for more details). The values of l_{max} were systematically increased in 100 points, which correspond to 1 second, and the reliability of the rms correlation exponent α was calculated by means of the R^2 parameter, which measures the goodness fit (also called the coefficient of determination).

Second, in order to discern whether the experimental Cl^- currents exhibit non-trivial correlations, we have fixed a threshold criterion of $R^2 \geq 0.99$. The obtained α values were calculated for every window on each time series, and the results ranged between 0.75 and 1, being 0.927 ± 0.048 (mean \pm SD) the global mean $\bar{\alpha}$ of all the experimental chloride series. These non-trivial correlations encompassed between 1,500 and 6,500 evoked chloride values (mean of $3,809.5 \pm 1,298.8$), which correspond to periods of time ranging between 3 and 13 seconds (mean of 7.66 ± 2.6). Boundary times where achieved on the series n17 (experiment 6, pH = 7.0) and n2 (experiment 1, pH = 7.0) respectively. The mean rms correlation coefficients (α), as well as the number of evoked chloride values under the non-trivial correlation regimen (N), with their respective correlation times (T_c) for all the experimental series are given in Table 1. Figure 3 shows an example of rms fluctuation analysis applied to three calcium-activated chloride responses of the same oocyte (n1, n2 and n3 time series belonging to the experiment 1) for their T_c times on a single window. In all three cases, the obtained α values were significantly different to 0.5, and for at least 10, 13 and 12 seconds respectively, the evoked chloride dynamics presented non-trivial long-term correlations. Alternatively, long term correlations were also observed by calculating the autocorrelation function from the time series (Supplementary Information).

Next, we have studied the long-range correlations for $\alpha \geq 0.6$. The analysis showed exponents ranging between 0.6008 and 0.9718, which respectively correspond to the time series n1 (pH = 5.0, $l_{max} = 2,200$) and n17 (pH = 7.0,

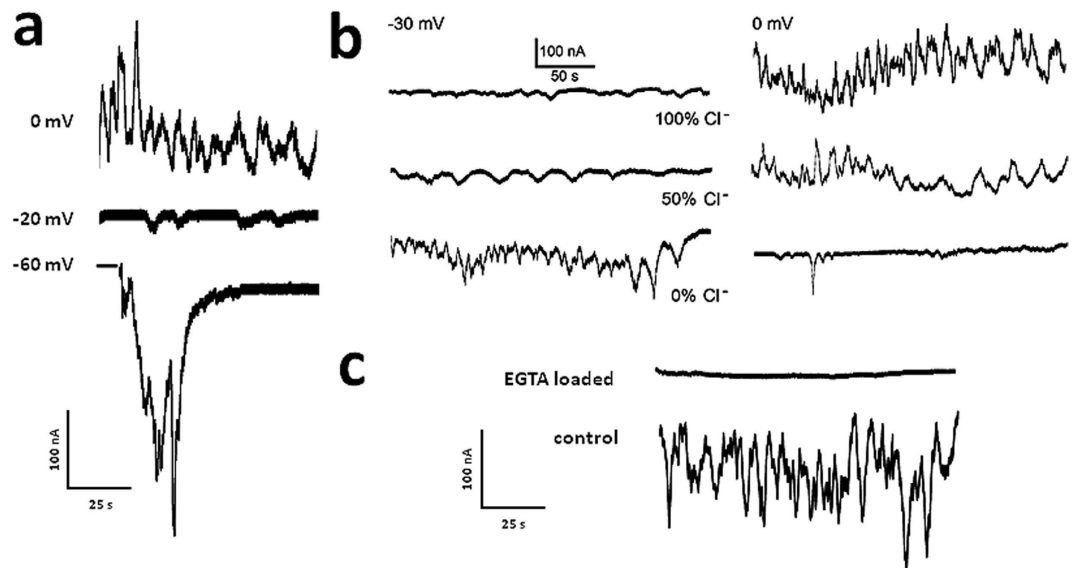


Figure 2. Ca^{2+} -dependent Cl^- current validation. (a) *Xenopus* oocyte held at either -60 , -40 , -20 or 0 mV. Reversal potential of oscillatory currents corresponded to a value close to -23 mV. (b) Oscillatory current reversal potential were dependent on external Cl^- concentration, traces show currents in oocytes held at -30 mV or 0 mV in 3 different solutions containing 100%, 50% or 0% Cl^- , reversal potential shifted toward more positive potentials as external Cl^- concentration decreased. (c) Cytoplasmic injection of EGTA, a Ca^{2+} chelator, completely eliminated the oscillatory Cl^- current.

Experiment	Stimulus	Number	α	N	T_c
1	pH5.0	n1	0.9137 ± 0.051	5,000	10
	pH7.0	n2	0.9286 ± 0.009	6,500	13
	pH9.0	n3	0.9118 ± 0.053	6,000	12
2	pH5.0	n4	0.9339 ± 0.035	4,500	9
	pH7.0	n5	0.9177 ± 0.031	5,000	10
	pH9.0	n6	0.9182 ± 0.056	5,000	10
3	pH5.0	n7	0.9226 ± 0.032	3,500	7
	pH7.0	n8	0.9002 ± 0.041	5,500	11
	pH9.0	n9	0.9471 ± 0.078	3,500	7
4	pH5.0	n10	0.9364 ± 0.030	3,500	7
	pH7.0	n11	0.9300 ± 0.037	4,000	8
	pH9.0	n12	0.9295 ± 0.050	2,500	5
5	pH5.0	n13	0.9301 ± 0.036	4,000	8
	pH7.0	n14	0.9199 ± 0.083	4,000	8
	pH9.0	n15	0.9096 ± 0.096	2,000	4
6	pH5.0	n16	0.9420 ± 0.049	3,000	6
	pH7.0	n17	0.9480 ± 0.062	1,500	3
	pH9.0	n18	0.9372 ± 0.049	3,000	6
7	pH5.0	n19	0.9372 ± 0.023	2,500	5
	pH7.0	n20	0.9208 ± 0.021	2,500	5
	pH9.0	n21	0.9433 ± 0.043	3,500	7

Table 1. The first column shows the number of the experiment, each one corresponding to a single oocyte. The second column contains the pH stimuli applied to each specific experiment. The third one shows the number assigned to each obtained chloride series. The rest of the data corresponds to the values of mean rms correlation coefficient (α), number of concentration measurements under the correlation regimen (N), and regime correlation time in seconds for non-trivial correlations (T_c).

$l_{max} = 1,200$). The global average $\bar{\alpha}$ was 0.774 ± 0.108 . All the means of α values, R^2 adjustments, and the l_{max} are given in Table 2. It can be observed that the values of α decrease slowly as l_{max} increases. This behavior is illustrated in Fig. 4a, where the average $\bar{\alpha}$ for the 21 time series, as a function of l_{max} , are represented; all the corresponding values of the Fig. 4 are displayed on Table 3.

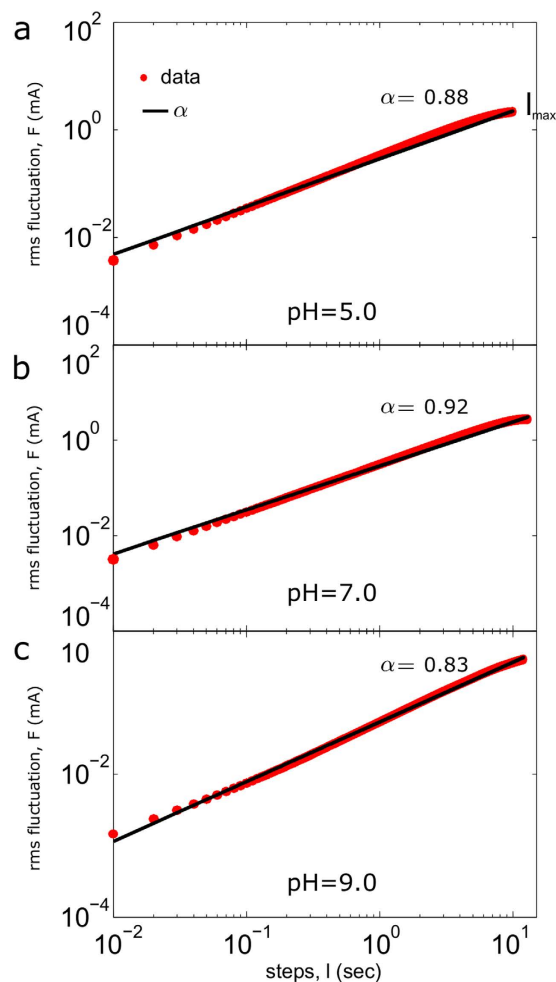


Figure 3. Root mean square fluctuation analysis applied to experiment 1 on a single window. Log-log plot of the rms fluctuation F versus l step. The red points depict the results of the original data for each value of l , while the black lines represent the regression lines. (a) $\alpha = 0.88$ (n1), (b) $\alpha = 0.92$ (n2) and (c) $\alpha = 0.83$ (n3). Corresponding (respectively) R^2 adjustment coefficients were 0.9915, 0.9921 and 0.9976. The high values of α and R^2 indicate non-trivial long-term correlations for each chloride time series during 10, 13 and 12 seconds respectively.

In addition, we have observed a critical transition around $l_{max} = 28$ seconds, where the behavior of the Cl^- currents changes from positive to negative correlations (Fig. 4b). It can be observed that as l_{max} increases, all the α exponent values decreased, and for the maximum window length ($l_{max} = 40$, corresponding to 20,000 time points), the α values were lower than 0.5 ($\bar{\alpha} = -0.051 \pm 0.283$) indicating anti-correlations in all cases; concretely, α values ranged between -0.885 and 0.349 , which belong to n2 (experiment1, pH = 7.0) and n7 time series (experiment 3, pH = 5.0) respectively.

Finally, we performed a rms fluctuation analysis without the separation of the data in shorter windows, thus considering all the points for each experimental time series, observing anti-correlations for all the cases ($\bar{\alpha} = -0.01 \pm 0.1$).

Moreover, we have examined whether the chloride currents are described by a fractional Gaussian noise (fGn) or a fractional Brownian motion (fBm) by calculating the slope of the Power Spectral Density plot⁴¹. The signal exhibits power law scaling if the relationship between its Fourier spectrum and the frequency is approximated asymptotically by $S(f) \approx S(f_0)/f^\beta$, where $S(f_0)$ and β are constant values. If $-1 < \beta < 1$ the signal corresponds to an fGn. In particular, when $\beta = 0$, the power spectrum is flat, as is the case for white noise in which the time series is composed of a sequence of independent random values. If $1 < \beta < 3$ the signal corresponds to a fBm. The analysis of the Power Spectral Density plot revealed that the experimental series are characterized by a power-law scaling with β ranging within 1.507 and 2.991, which suggests that all the series are described by fBm (β values are given in Table 4).

Additionally, an analysis of the classical descriptive statistics of the experimental data has been included in the Supplementary Information).

Next, we have checked whether the chloride time series show persistent or anti-persistent long-term memory by calculating the Hurst exponent. Although several tools exist for estimating the long-term memory from

Experiment	Stimulus	Number	α	R^2	$\max I_{max}$
1	pH5.0	n1	0.7094 \pm 0.100	0.8896 \pm 0.070	2,200
	pH7.0	n2	0.7587 \pm 0.092	0.8844 \pm 0.120	2,200
	pH9.0	n3	0.7145 \pm 0.080	0.8730 \pm 0.107	2,300
2	pH5.0	n4	0.7245 \pm 0.093	0.8955 \pm 0.085	2,200
	pH7.0	n5	0.7638 \pm 0.107	0.9243 \pm 0.084	2,000
	pH9.0	n6	0.7390 \pm 0.103	0.9085 \pm 0.059	2,000
3	pH5.0	n7	0.7974 \pm 0.089	0.9430 \pm 0.084	1,300
	pH7.0	n8	0.7614 \pm 0.081	0.9315 \pm 0.066	2,000
	pH9.0	n9	0.8474 \pm 0.133	0.9721 \pm 0.033	1,500
4	pH5.0	n10	0.7918 \pm 0.111	0.9288 \pm 0.099	1,400
	pH7.0	n11	0.7543 \pm 0.128	0.8913 \pm 0.113	1,600
	pH9.0	n12	0.7406 \pm 0.142	0.9309 \pm 0.050	1,500
5	pH5.0	n13	0.7905 \pm 0.110	0.9313 \pm 0.077	1,500
	pH7.0	n14	0.7792 \pm 0.111	0.9551 \pm 0.028	1,800
	pH9.0	n15	0.8190 \pm 0.136	0.9784 \pm 0.034	1,000
6	pH5.0	n16	0.8550 \pm 0.119	0.9675 \pm 0.061	1,200
	pH7.0	n17	0.8734 \pm 0.138	0.9836 \pm 0.030	900
	pH9.0	n18	0.7264 \pm 0.079	0.9206 \pm 0.051	2,000
7	pH5.0	n19	0.7262 \pm 0.067	0.9142 \pm 0.048	1,300
	pH7.0	n20	0.7837 \pm 0.066	0.9435 \pm 0.054	900
	pH9.0	n21	0.8137 \pm 0.107	0.9549 \pm 0.055	1,500

Table 2. The first column shows the number of the experiment, each one corresponding to a single oocyte. The second column contains the pH stimuli applied to each specific experiment. The third one shows the number assigned to each obtained Cl^- series. The rest of the data corresponds to the values of mean rms correlation coefficient (α), coefficient of adjustment (R^2), and maximum regime correlation points ($\max I_{max}$).

fBm time series, one of the most reliable methods is the bridge detrended Scaled Windowed Variance analysis (bdSWV) (see Methods for more details). After bdSWV analysis, the resulting Hurst exponents had a mean value of 0.191 ± 0.101 , implying long-range memory and an anti-persistence effect in all the experimental data sets (Table 4). In addition, an ANOVA test revealed that Hurst exponent values were significantly different for time series corresponding to pH = 9.0 in comparison to pH = 7.0 (p-value = 10^{-5}) and pH = 5.0 (p-value = 10^{-4}), but no significant distinction was found between pH = 7.0 and pH = 5.0 (p-value = 0.42). Notice that the obtained values of H are very low, showing a high degree of anti-persistence (strong trend-reversing), so that an increasing trend in the experimental data values will tend markedly to be followed by a decreasing trend, or a decrease on average will be followed by a robustly increasing trend.

In order to estimate the significance of our results, we have performed a shuffling procedure that defines the null-hypothesis. If the original time series exhibits a memory structure ($H \neq 0.5$), after the shuffling such structure will disappear, thus re-applying a new Hurst analysis on the shuffled data should provide values of H close to 0.5. According to this procedure, for each experimental time series (21 in total), we performed a thousand random permutations, which allowed building the null-hypothesis of no correlations. In total, we generated 21,000 random series from the original data belonging to the seven experiments with *Xenopus laevis* oocytes. After shuffling, the results show a mean Hurst exponent of 0.499 ± 0.01 , indicating the absence of long-term memory i.e., the informational memory structures in all shuffled series was completely lost. Notice that after shuffling, the series became Gaussian white noise (fGn series with $\beta = -0.0006 \pm 0.004$, and for this case the use of bdSWV is not justified. Instead, Dispersion Analysis is the most recommendable tool for this kind of series^{41,42} (for more details see Methods).

Figure 5a illustrates the regression lines of a bdSWV process applied to an example of experimental series giving $H = 0.104$ (experiment 5, n13, pH = 5.0), which indicates a strong anti-persistent memory. After randomly permuting all the 130,000 points contained in this time series n13, the Dispersion Analysis gave $H = 0.492$, which indicates a breakdown for the long-term memory (Fig. 5b). In Fig. 5c, we represent 100 Hurst exponent values corresponding to 100 shuffled series, obtained from shuffling the experimental data. It can be observed that, after shuffling, the long-term memory disappears completely in all the time series ($\bar{H} = 0.498 \pm 0.01$). For illustration purposes, Fig. 5c shows, rather than the 21,000 obtained values of Hurst exponent, only 100 of them. The informational memory structures in all shuffled series were completely broken-down, and therefore, the memory structure that characterizes the experimental data could not be found by chance. Finally, in order to calculate the values of Hurst exponent from short data periods, we used the Detrended Fluctuation Analysis (DFA), because the bdSWV is recommended for data sizes greater than 2^{12} , whilst for data sets with less than 2^8 points bdSWV has been shown to be unreliable⁴³. The DFA analysis showed that for time periods ranging between 2 and 5 seconds all the experimental time series exhibit persistent behavior with $H > 0.5$ being the global mean of $\bar{H} = 0.697 \pm 0.11$, which indicates that the properties of persistent memory dominate at short time intervals of the calcium-activated chloride currents in *Xenopus laevis* oocytes.

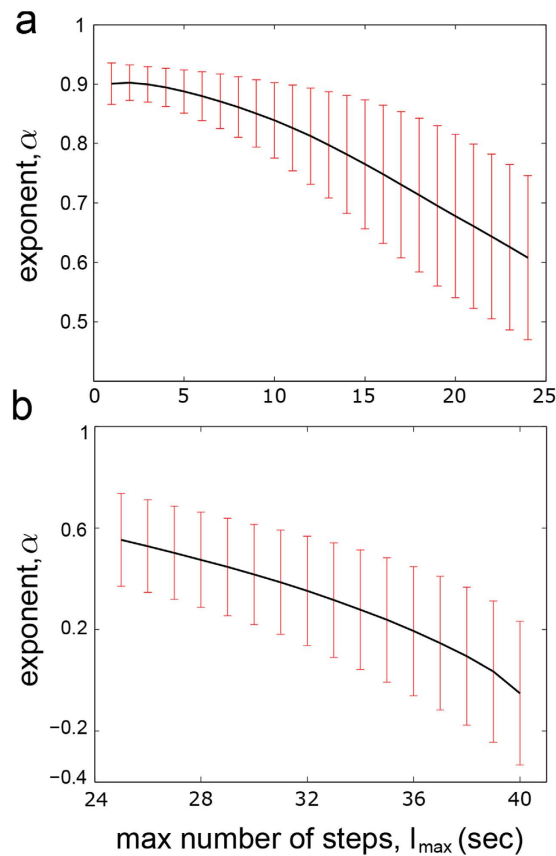


Figure 4. Long-term correlations across different windows lengths. (a) Global average $\bar{\alpha}$ versus different values of l_{max} (varying from 1 to 24 seconds). (b) $\bar{\alpha}$ as a function of l_{max} (varying from 25 to 40 seconds). The error bars represent the standard deviation at each step. It can be observed that all Cl^- time series change from positive to negative correlation near $l_{max} = 28$ seconds.

l_{max} (sec)	$\bar{\alpha}$	l_{max} (sec)	$\bar{\alpha}$
1	0.967 ± 0.04	21	0.647 ± 0.18
2	0.970 ± 0.03	22	0.624 ± 0.18
3	0.966 ± 0.03	23	0.600 ± 0.18
4	0.959 ± 0.04	24	0.577 ± 0.18
5	0.950 ± 0.04	25	0.553 ± 0.18
6	0.939 ± 0.05	26	0.528 ± 0.18
7	0.928 ± 0.06	27	0.502 ± 0.18
8	0.915 ± 0.06	28	0.474 ± 0.18
9	0.901 ± 0.07	29	0.446 ± 0.19
10	0.885 ± 0.08	30	0.416 ± 0.19
11	0.868 ± 0.09	31	0.385 ± 0.20
12	0.849 ± 0.10	32	0.351 ± 0.21
13	0.830 ± 0.11	33	0.316 ± 0.22
14	0.808 ± 0.13	34	0.278 ± 0.23
15	0.786 ± 0.14	35	0.238 ± 0.24
16	0.764 ± 0.15	36	0.193 ± 0.25
17	0.741 ± 0.16	37	0.146 ± 0.26
18	0.717 ± 0.17	38	0.094 ± 0.27
19	0.693 ± 0.17	39	0.034 ± 0.27
20	0.670 ± 0.18	40	-0.051 ± 0.28

Table 3. The first and third columns represent different values of l_{max} , ranging from 1 to 40 seconds. The second and fourth columns show the values of global mean rms correlation coefficients ($\bar{\alpha}$) for each l_{max} values.

Experiment	Stimulus	Number	β	H
1	pH5.0	n1	2.0248	0.2455 ± 0.0012
	pH7.0	n2	1.9723	0.1744 ± 0.0016
	pH9.0	n3	2.3749	0.0950 ± 0.0019
2	pH5.0	n4	1.8460	0.2501 ± 0.0009
	pH7.0	n5	1.6835	0.3521 ± 0.0007
	pH9.0	n6	1.5669	0.1076 ± 0.0017
3	pH5.0	n7	1.8741	0.1830 ± 0.0012
	pH7.0	n8	1.5075	0.2681 ± 0.0010
	pH9.0	n9	1.7512	0.1071 ± 0.0015
4	pH5.0	n10	1.5834	0.2962 ± 0.0012
	pH7.0	n11	1.6043	0.3174 ± 0.0011
	pH9.0	n12	1.9086	0.0616 ± 0.0023
5	pH5.0	n13	2.0533	0.1040 ± 0.0019
	pH7.0	n14	2.0738	0.1725 ± 0.0016
	pH9.0	n15	1.8572	0.0589 ± 0.0022
6	pH5.0	n16	2.3889	0.3339 ± 0.0009
	pH7.0	n17	2.4520	0.2949 ± 0.0011
	pH9.0	n18	2.1698	0.0526 ± 0.0022
7	pH5.0	n19	2.8441	0.2174 ± 0.0016
	pH7.0	n20	2.5817	0.2718 ± 0.0013
	pH9.0	n21	2.9913	0.0621 ± 0.0026

Table 4. The first column shows the number of the experiment, each one corresponding to a single oocyte. The second column contains the pH stimuli applied to each experiment. The third one shows the number assigned to each obtained chloride series. The rest of the data corresponds to the values of Power Spectral Density slope (β) and Hurst exponent (H) calculated by the bdSWV method.

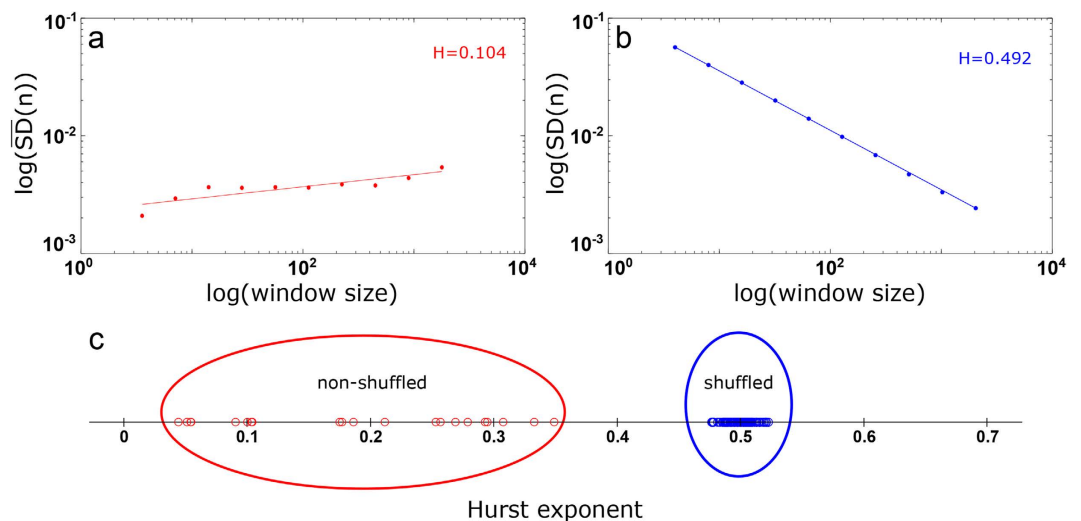


Figure 5. Hurst exponents obtained by the bdSWV analysis. (a) The slope of a log-log plot of the $\overline{SD}(n)$ versus the window size for a bdWSV applied to an evoked chloride series (n13, experiment 5, pH = 5.0) gives $H = 0.104$, indicating the presence of long-term memory. (b) The slope of a log-log plot of the $SD(n)$ versus the window size for a Dispersion Analysis applied to shuffled time series obtained by randomly permuting all the 130,000 time points for each Cl^- time series (n13). After shuffling, H was close to 0.5, indicating the disappearance of the memory structure. (c) In red, Hurst exponent values of all the experimental chloride time series; in blue, 100 Hurst exponent values obtained from shuffled series.

Discussion

Chloride (Cl^-) thought to be the most abundant permeable anion in the cell; it participates in a wide variety of important local and systemic physiological processes, while also being involved in a variety of human diseases. Historically, chloride anions have been of less interest than most other free cations. In fact, many molecular aspects of the chloride channels have been well studied, but the characterization of their dynamic properties is still unknown.

Here, we have quantitatively studied experimental Ca^{2+} -dependent Cl^- currents belonging to *Xenopus laevis* oocytes, which have been evoked by serum under different external pH environments. These Cl^- currents were measured by the patch-clamp technique and the data series have been mainly analyzed by means of non-linear dynamic tools.

First, we have applied an analysis based on the root mean square fluctuation and the results revealed non-trivial correlations in all experimental time series. The α exponent has a mean of 0.927 ($R^2 \geq 0.99$) and these strong long-range correlations encompasses concentration values between 1,500 and 6,500, which correspond to time periods ranging between 3 and 13 seconds (with a mean of 7.66 sec). Therefore, the chloride currents present a dynamical structure characterized by long range correlations, and this occurred independently of the experimental condition (here defined by the pH of the cellular external medium).

In addition, transitions from negative to positive correlations were found in the Ca^{2+} -dependent Cl^- data. Positive long-range correlations arise in short time intervals while negative correlations become dominant over longer ones. This dynamic behavior has been observed in all experimental chloride series.

Moreover, we have calculated the slope of the Power Spectral Density plot concluding that the Cl^- data sets can be categorized as fractional Brownian motion i.e., non-stationary series with time-dependent variance.

To test the presence of persistent or anti-persistent memory properties for long time intervals in the experimental data, we have applied the bridge detrended Scaled Windowed Variance analysis, a specific method to obtain Hurst exponent values in fBm signals. We have found that the Hurst exponents satisfy $0.05 < H < 0.35$, indicating the existence of anti-persistent long-term memory during long time intervals, in all the series. Values of $H < 0.5$ have been interpreted as a characteristic for “trend-reversing”, which means that a decreasing trend in the past usually implies an increasing trend (on average) in the future and vice versa, an increase over a set of values in the past is likely to be followed by a decrease in the future.

Our obtained Hurst exponent values are very small ($\bar{H} = 0.191 \pm 0.101$), which shows a high degree of negative dependence between experimental values, indicating strong “trend-reversing”. The strength of this reversion tendency increases as H approaches 0; consequently, when the evoked calcium values spike in one direction, there is a very strong probability that they will subsequently revert back. This important anti-persistent property indicates a self-correcting effect in the experimental data, which describes a situation where tendencies to increase or decrease will tend to reverse themselves.

The high reliability of our Hurst analysis for long time intervals was tested by applying a shuffling procedure (21,000 shuffled time series in total), showing that the Hurst exponent values measured from the original experimental series ($\bar{H} = 0.191 \pm 0.101$) were significantly different from the ones obtained after shuffling ($\bar{H} = 0.498 \pm 0.01$), implying that the correlation structure in all shuffled series was completely broken-down, and therefore, the memory structure that characterizes the original experimental data could not be found by chance.

Finally, in agreement with the observed transitions from negative to positive correlations in the rms fluctuation analysis, we have verified that persistent memory properties arise for short time intervals in all the experimental data sets, while anti-persistent behaviors become dominant in longer intervals. This “crossover phenomenon”, a dynamical property characterized by transitions from persistent to anti-persistent behaviors at a physiological level, seems to show a highly complex regulation of the intracellular chloride currents which exhibit persistence at short time scales (i.e., a trend to increasing in the past will likely be followed by an increasing trend in the future and, vice versa, a trend to decreasing in the past will likely be followed by a decreasing trend in the future), while strong anti-persistence arises in long time scales (when the chloride currents present a determinate trend in the past, there is a high probability to subsequently revert back); this “trend-reversing” behavior suggest that, at long time intervals, the intracellular chloride dynamics are bounded, and reflects the consequences of an inherent self-correcting effect in the system⁴⁴. Similar crossover phenomena have also been observed in some other numerical and experimental physiological processes⁴⁴.

Long-term memory properties found in the calcium-activated chloride behaviors might be related to the dynamic metabolic memory recently proposed to exist in the Cellular Metabolic Structure (CMS in short)^{45,46}. At a systemic level, cells seem to display a CMS, which behaves as a very complex decentralized information processing system with the capacity to store metabolic memory. According to this framework, the CMS exhibits an essential dynamic informational mechanism by which Hopfield-like attractor dynamics regulate the enzymatic activities. These attractors have the capacity to store functional catalytic patterns that can be correctly recovered by specific input stimuli. The Hopfield-like metabolic dynamics are stable and can be maintained as a long-term functional memory^{45,46}.

Moreover, since the beginning of the neuronal network modeling of associative memory, the connectivity matrix in the Hopfield network was assumed to result from a long-term memory learning process, occurring over a much slower time scale than neuronal dynamics^{47–49}. Therefore, it is well accepted that the attractors emerging in neuronal dynamics described by Hopfield networks are the result of a long-term memory process. Besides, extensive physiological recordings of neuronal processes have revealed the presence of long range correlations in plasticity dynamics for measured synaptic weights. For instance, long tails in the synaptic distribution of weights have been interpreted as short-term memory in neural dynamics⁵⁰.

These studies and others support the thesis that neuronal dynamics exhibit both long-term and short-term memory, and the same may happen with the metabolic processes. In fact, long-term correlations (mimicking short-term memory in neuronal systems) have also been analyzed in different metabolic processes not belonging to the neuronal lineage. One of the most studied is the calcium-activated potassium channels, existing in Leydig cells⁵¹, kidney Vero cells⁵² and human bronchial epithelial cells⁵³. Other biochemical processes also present long-term correlations for example, the intracellular transport pathway of *Chlamydomonas reinhardtii*⁵⁴, the NADPH series of mouse liver cells⁵⁵, and the mitochondrial membrane potential of cardiomyocytes⁵⁶. Similar to what happens in the brain, we believe that the observed long-term memory in the calcium-activated chloride

responses might correspond to the short-term memory of the metabolic system involved in these physiological dynamics, and in accordance with our analysis for the non-trivial correlation regimes, this short-term memory could correspond to times around 7 seconds.

In brief, here, we have addressed some essential aspects of calcium-activated chloride currents, in which the concentration dynamics are strongly conditioned by previous concentration measurements over time. Indeed, non-trivial correlations were observed within time-windows of 4,000 experimental concentration values, which correspond approximately to time memory periods with a mean of 7.6 seconds. The analyzed experimental series exhibit fractional Brownian motion, with an informational structure characterized by highly organized data sequences, memory properties and inherent “crossover” dynamics, in which persistent behaviors exist within short time intervals, while anti-persistent dynamics become dominant within long time intervals. In addition, the anti-persistent behavior that encompasses all the points of the time series suggests self-correcting effects in the experimental data. These properties seem to be intrinsic characteristics of the dynamics involved in these physiological processes.

Our work opens up new perspectives for quantitative analysis of the dynamics involved in the dysfunction of calcium-activated chloride channels and sheds some light on the understanding of the informational properties of intracellular signals, a key element to elucidate the physiological functional coupling of the cell with the integrative dynamics of metabolic processes.

Methods

Calcium-activated chloride currents in *Xenopus laevis* oocytes. Adult *Xenopus laevis* frogs were obtained from Blades Biological (Cowden, Kent, UK). Oocytes at stage V were plucked from the ovaries and defolliculated by collagenase treatment (type 1, Sigma-Aldrich Quimica, S.A., Madrid, Spain) at 80–630 units/ml in frog Ringer’s solution (115 mM NaCl, 2 mM KCl, 1.8 mM CaCl₂, 5 mM HEPES at pH 7.0) for 20 min in order to remove the surrounding follicular and epithelial cell layers. Oocytes were maintained at 18 °C in sterile unsupplemented modified Barth’s medium containing (mM): 88 NaCl, 0.2 KCl, 2.4 NaHCO₃, 0.33 Ca(NO₃)₂, 0.41 CaCl₂, 0.82 MgSO₄, 0.88 KH₂PO₄, 2.7 Na₂HPO₄, with gentamicin 70 µg/ml and adjusted to pH 7.4.

Xenopus oocytes have long been a model system for the study of calcium-activated chloride currents because they express extremely high levels of chloride channels whose activation depends on Ca²⁺⁴⁰.

For this activation we have used Fetal Bovine Serum (FBS). Serum is known to promote oscillations due to alterations of Ca²⁺ concentrations in the cytoplasm, which, as a consequence, evoke Cl[−] movements across the oocyte membrane⁵⁷ through different calcium-activated chloride channels. According to this procedure, FBS (Sigma-Aldrich) diluted 1:1000 in Ringer’s solution was used for the oocytes’ perfusion to achieve the generation of chloride currents oscillations. The membrane was usually voltage clamped at −60 mV, and in the experiments, three different pH conditions were considered, Ringer’s solution at pH 5.0, 7.0 and 9.0. The sampling interval time scale in the experiments was 2 milliseconds.

All the procedures followed the guidelines of regulation 1201/2005 of Ministerio de Agricultura, Pesca y Alimentación and the experimental protocols were approved by the University of the Basque Country (UPV/EHU) ethics committee (code: CEBA/8/2009).

Root mean square fluctuation. An important measure for quantifying long-range correlations in time series is the root mean square (rms) fluctuation⁵⁸, a technique initially developed for random walk studies⁵⁹. Before calculating it, we define the move-step length at time point i ; here, for the evoked calcium-activated chloride time series, it simply corresponds to electrical current variations, i.e., $u^k(i) \equiv \Phi(i+k) - \Phi(i)$ which are given in nanoamperes (nA). Without loss of generality, hereon, we denote for a fixed k , $u^k(i) \equiv u(i)$. Next, defining the net displacement after l steps as

$$y(l) \equiv \sum_{i=1}^l u(i), \quad (1)$$

the rms fluctuation of the average displacement is given by:

$$F(l) \equiv \sqrt{\langle \Delta y(l)^2 \rangle - \langle \Delta y(l) \rangle^2}, \quad (2)$$

where $\Delta y(l) \equiv y(l+l_0) - y(l_0)$, and brackets denote average over all possible values of l_0 . Thus, $F(l)$ is defined as the square root of the difference between the average of the square of $\Delta y(l)$ minus the square of its average.

For many processes, $F(l)$ scales asymptotically with l , i.e., $F(l) \sim l^\alpha$ ⁵⁸, and the relationship can be observed by representing F as a function of l in a log-log plot, fitting $F(l)$ in the range $l = 1, \dots, l_{max}$. Here, three important regimes can be distinguished, depending on the exponent α (rms correlation coefficient)⁵⁸; when $\alpha = 0.5$ the random walk is time uncorrelated and no memory exists. Markov processes initially decay exponentially with l , but also give $\alpha = 0.5$ for sufficiently large l . If $\alpha > 0.5$, it indicates the presence of positive long-range correlations and $\alpha < 0.5$ implies long-term anti-correlations.

When the method is applied directly to large data sets, there is a risk of concluding that there are no correlations from long-term correlated data. To avoid this issue, data can be subdivided into smaller windows. In our case, the chloride data consisted of 130,000 time points, which we divided into 6 non-overlapping windows of 20,000 points each, leaving the last 10,000 values out of the analysis. The final α was calculated averaging over the 6 individual values of α , each one calculated within a different window. To estimate the duration of the long-term correlation regime (T_c), we increased l_{max} systematically until the value of R^2 (the goodness fit in the log-log scale) was smaller than 0.99.

Hurst exponent Scaled Windowed Variance Analysis. The calculation of the Hurst exponent is a classical method to detect long-term memory in time series introduced by the hydrologist H.E. Hurst in 1951 to study the annual discharges of the Nile River⁶⁰. Afterwards, this method was developed by Mandelbrot in order to apply it to different dynamic processes⁶¹.

The Hurst exponent, H , is referred to as the index of long-range dependence, which characterizes how the variance depends on a time interval, and also provides information about autocorrelations. The H exponent is also related to the fractal dimension for self-affine series⁶², and for one-dimensional series, $H = 2 - D$, where D is the fractal dimension and satisfies $1 < D < 2$ ⁶³.

The Hurst exponent H satisfies $0 \leq H \leq 1$. For a random process with independent increments, H is 0.5. When H differs from 0.5, the process is properly fractional and indicates the existence of long-term memory, in which future events have long-term correlations with past events. If $H > 0.5$, it indicates a biased random process which exhibits persistent behavior. In this case, for several previous transitions, an increment on the average value implies an increasing trend in the future. Conversely, a previously decreasing trend for a sequence of values usually implies a decrease for a similar sequence. Anti-persistent behavior is obtained for $0 \leq H < 0.5$; in this case, a previously decreasing trend implies a probable increasing trend in the future and vice versa, an increase in the past is usually followed by a decrease in the future^{41,53}. Persistent behavior carries out a superdiffusion, which is faster than in a normal random walk; and, conversely, anti-persistent behavior carries out an abnormal diffusion that is slower than in a normal random walk. In some dynamic processes a transition from persistent to anti-persistent correlation regimes over different time scales, which is known as a “crossover phenomenon”, may emerge⁴⁴.

Two fundamental classes of fractal time series are fractional Brownian motion (fBm) and fractional Gaussian noise (fGn). The fBm is a continuous-time Gaussian process $B^H(t)$ with $t \geq 0$ such that it satisfies $B^H(0) = 0$ with probability 1, the expectation $E[B^H(t)]$ is 0 for every t , and the covariance function is given by $E[B^H(t_1)B^H(t_2)] = \frac{1}{2}(t_1^{2H} + t_2^{2H} - |t_1 - t_2|^{2H})$ for every t_1, t_2 in \mathfrak{R}^+ , where the parameter H is the Hurst exponent. The fractional Gaussian noise (fGn) is the process $W^H(t)$, with $t \geq 0$, obtained from the fBm increments for discrete time, that is, $W^H(t) = B^H(t+1) - B^H(t)$.

The two main, most robust methods to calculate the Hurst exponent are the Dispersion Analysis applied on fractional Gaussian noise (fGn) and Scaled Windowed Variance Analysis for fBm signals⁴¹.

The Scaled Windowed Variance Analysis (SWVA) is a reliable method for the estimation of the Hurst exponent (H) that has been thoroughly tested on fractional Brownian motion (fBm) signals⁴³. In particular, we have used the bridge detrended Scaled Windowed Variance analysis (bdSWV) for the study of calcium-signal time series⁴¹. To define the SWVA method, let the time series signal be represented by x_t , with $t = 1, \dots, N$, time points. Next, the following steps are carried out for each one of the window sizes $n = 2, 4, \dots, N/2, N$ (if N is not a power of 2, then n takes the values $2, 4, \dots, 2^k$, where k is the integer part of $\log_2 N$):

- (1) Partition of the data points in N/n adjacent non-overlapping windows $\{W_1, \dots, W_{N/n}\}$ of size n , where $W_i = \{x_{(i-1)n+1}, \dots, x_{in}\}$. If N is not a power of 2 and N is not divisible by n , then the last remaining points are ignored for this value of n . For instance, if $N = 31$ and $n = 4$, the first 28 points are partitioned into seven windows.
- (2) Subtraction of the line between the first and last points in the n -th window.
- (3) For each $i = 1, \dots, N/n$, calculation of the standard deviation SD_i of the points in each window, by using the formula

$$SD_i = \sqrt{\frac{\sum_{t=(i-1)n+1}^{i \cdot n} (x_t - \bar{x}_i)^2}{n-1}}, \quad (3)$$

where \bar{x}_i is the average in the window W_i .

- (4) Evaluation of the average $\overline{SD}(n)$ of the N/n standard deviations corresponding to equation (3).
- (5) Observation of the range of the window sizes n over which the regression line of $\log(\overline{SD}(n))$ versus $\log(n)$ gives a good fit (usually some initial and end points are excluded).
- (6) In this range, the slope of the regression line gives the estimation of the Hurst coefficient H .

Here, to calculate SWVA, we have made use of the program bdSWV, available on the web of the Fractal Analysis Programs of the National Simulation Resource⁶⁴.

Dispersion Analysis. The Dispersion Analysis (DA) method is applied for the estimation of the Hurst exponent (H) on fractional Gaussian noise (fGn)⁴².

For different bins of length n , with n varying from 2 to $N/2$, one can define the standard deviation $SD(n)$ of the series formed by the mean of the n consecutive values of the original series x_t . That is, $SD(n)$ is the standard deviation of the series $y_{n,i}$, where

$$y_{n,i} = \frac{x_i + \dots + x_{(i+n-1)}}{n}. \quad (4)$$

Now, the relation between $\log(SD(n))$ and $\log(n)$ is approximately linear:

$$SD(n) = SD(1) \cdot n^{H-1}, \quad (5)$$

with slope $H-1$, where H is the Hurst coefficient and $SD(1)$ the standard deviation calculated on the first window.

Detrended Fluctuation Analysis. Detrended Fluctuation Analysis (DFA) is a method that allows for the detection of long-memory processes on non-stationary time series that can be used properly for small data sizes⁶⁵.

The method is summarized as follows: first, given the time series $y(t)$ we obtain a signal profile by computing the cumulative sum

$$x(k) = \sum_{i=1}^k (y(i) - \bar{y}), \quad (6)$$

of the time series. Then, the obtained time series is divided into boxes of equal length n . Next, the local trend $x_n(k)$ in each box is subtracted and the fluctuations of this detrended and integrated signal is calculated by

$$F(n) = \sqrt{\frac{1}{n} \sum_{k=1}^N [x(k) - x_n(k)]^2}, \quad (7)$$

This computation is repeated over all box sizes obtaining a relationship between fluctuations $F(n)$ and box sizes n . A linear relationship on a log-log graph indicates the presence of scaling, and under such conditions, fluctuations can be characterized by a scaling exponent γ , related to the Hurst exponent⁶⁶. Mainly, if $0 < \gamma < 0.5$, the process is anti-correlated and exhibits anti-persistent behavior, which can be modeled by fGn with $H = \gamma$. When $0.5 < \gamma < 1$, the process exhibits positive correlations and persistent behavior which can be modeled by fGn with $H = \gamma$, and for a random process with independent increments, γ is 0.5 ($H = \gamma$). Other scenarios also can be considered in DFA⁶⁶. Besides, we would like to highlight some of the recent progress in nonlinear time series analysis^{67–71}.

Use of experimental cells. *Xenopus laevis* frogs (Guy Pluck, Xenopus Express, France) were anaesthetized by hypothermia. Ovary lobules (4–8) were surgically removed under sterile conditions. After surgery, frogs were sutured, and allowed to recover and then returned to housing. No further oocytes were taken for at least 2 months. All the procedures followed the guidelines of regulation 1201/2005 of Ministerio de Agricultura, Pesca y Alimentacion and the experimental protocols were approved by the University of the Basque Country (UPV/EHU) ethics committee (code: CEBA/8/2009).

References

- Huang, F., Wong, X. & Jan, L. Y. International Union of Basic and Clinical Pharmacology. LXXXV: calcium-activated chloride channels. *Pharmacol. Rev.* **64**, 1–15 (2012).
- Jentsch, T. J. & Günther, W. Chloride channels: an emerging molecular picture. *Bioessays* **19**, 117–126 (1997).
- Jentsch, T. J., Neagoe, I. & Scheel, O. CLC chloride channels and transporters. *Curr. Opin. Neurobiol.* **15**, 319–325 (2005).
- Jentsch, T. J., Stein, V., Weinreich, F. & Zdebik, A. A. Molecular structure and physiological function of chloride channels. *Physiol. Rev.* **82**, 503–568 (2002).
- Nilius, B. & Droogmans, G. Amazing chloride channels: an overview. *Acta Physiol. Scand.* **177**, 119–147 (2003).
- O'Rourke, B. Mitochondrial ion channels. *Annu. Rev. Physiol.* **69**, 19–49 (2007).
- Stauber, T. & Jentsch, T. J. Chloride in vesicular trafficking and function. *Annu. Rev. Physiol.* **75**, 453–477 (2007).
- Tang, C. Y. & Chen, T. Y. Physiology and pathophysiology of CLC-1: mechanisms of a chloride channel disease, myotonia. *J. Biomed. Biotechnol.* **2011**, 685328 (2011).
- Okada, Y. *et al.* Volume-sensitive chloride channels involved in apoptotic volume decrease and cell death. *J. Membr. Biol.* **209**, 21–29 (2006).
- Gonzalez-Silva, C. *et al.* Ca²⁺-activated Cl[−] channels of the ClCa family express in the cilia of a subset of rat olfactory sensory neurons. *PLoS One* **9**, e69295 (2013).
- Mao, J. *et al.* Volume-activated chloride channels contribute to cell-cycle-dependent regulation of HeLa cell migration. *Biochem. Pharmacol.* **77**, 159–68 (2009).
- Kim, M. J., Cheng, G. & Agrawal, D. K. Cl[−] channels are expressed in human normal monocytes: a functional role in migration, adhesion and volume change. *Clin. Exp. Immunol.* **138**, 453–459 (2004).
- Berg, J., Yang, H. & Jan, L. Y. Ca²⁺-activated Cl[−] channels at a glance. *J. Cell. Sci.* **125**, 1367–1371 (2012).
- Frizzell, R. A. & Hanrahan, J. W. Physiology of epithelial chloride and fluid secretion. *Cold. Spring. Harb. Perspect. Med.* **6**, a009563 (2012).
- Voglis, G. & Tavernarakis, N. The role of synaptic ion channels in synaptic plasticity. *EMBO Rep.* **7**, 1104–1110 (2006).
- Duan, D. D. Phenomics of cardiac chloride channels. *Compr. Physiol.* **3**, 667–692 (2013).
- Endeman, D. *et al.* Chloride currents in cones modify feedback from horizontal cells to cones in goldfish retina. *J. Physiol.* **590**, 5581–5595 (2012).
- Pifferi, S., Cenedese, V. & Menini, A. Anoctamin 2/TMEM16B: a calcium-activated chloride channel in olfactory transduction. *Exp. Physiol.* **97**, 193–199 (2012).
- Huang, W. C. *et al.* Calcium-activated chloride channels (CaCCs) regulate action potential and synaptic response in hippocampal neurons. *Neuron* **74**, 179–192 (2012).
- Kuner, T. & Augustine, G. J. A genetically encoded ratiometric indicator for chloride: capturing chloride transients in cultured hippocampal neurons. *Neuron* **27**, 447–459 (2000).
- Isomura, Y. *et al.* Synaptically activated Cl[−] accumulation responsible for depolarizing GABAergic responses in mature hippocampal neurons. *J. Neurophysiol.* **90**, 2752–2756 (2003).
- Li, M., Wang, Q., Lin, W. & Wang, B. Regulation of ovarian cancer cell adhesion and invasion by chloride channels. *Int. J. Gynecol. Cancer* **19**, 526–530 (2009).
- Peretti, M. *et al.* Chloride channels in cancer: Focus on chloride intracellular channel 1 and 4 (CLIC1 AND CLIC4) proteins in tumor development and as novel therapeutic targets. *Biochim. Biophys. Acta* **1848**, 2523–2531 (2015).
- Duran, C., Thompson, C. H., Xiao, Q. & Hartzell, H. C. Chloride channels: often enigmatic, rarely predictable. *Annu. Rev. Physiol.* **72**, 95–121 (2010).
- Ponce-Coria, J. *et al.* Regulation of NKCC2 by a chloride-sensing mechanism involving the WNK3 and SPAK kinases. *Proc. Nat. Acad. Sci.* **105**, 8458–8463 (2008).
- Jentsch, T. J. CLC chloride channels and transporters: from genes to protein structure, pathology and physiology. *Crit. Rev. Biochem. Mol. Biol.* **43**, 3–36 (2008).
- Planells-Cases, R. & Jentsch, T. J. Chloride channelopathies. *Biochim. Biophys. Acta* **1792**, 173–189 (2009).

28. Suzuki, M., Morita, T. & Iwamoto, T. Diversity of Cl^- Channels. *Cell. Mol. Life Sci.* **63**, 12–24 (2006).
29. Verkman, A. S. & Galletta, L. J. Chloride channels as drug targets. *Nat. Rev. Drug Discov.* **8**, 153–171 (2009).
30. Kane-Dickson, V., Pedit, L. & Long, S. B. Structure and insights into the function of a $\text{Ca}(2+)$ -activated $\text{Cl}(-)$ channel. *Nature* **516**, 213–218 (2014).
31. Hartzell, C., Putzier, I. & Arreola, J. Calcium-activated chloride channels. *Annu. Rev. Physiol.* **67**, 719–58 (2005).
32. Cross, N. L. Initiation of the activation potential by an increase in intracellular calcium in eggs of the frog, *Rana pipiens*. *Dev. Biol.* **85**, 380–384 (1981).
33. Miledi, R. A calcium-dependent transient outward current in *Xenopus laevis* oocytes. *Proc. R. Soc. Lond. B. Biol. Sci.* **215**, 491–497 (1982).
34. Bader, C. R., Bertrand, D. & Schwartz, E. A. Voltage-activated and calcium-activated currents studied in solitary rod inner segments from the salamander retina. *J. Physiol.* **331**, 253–284 (1982).
35. Hoffmann, E. K., Holm, N. B. & Lambert, I. H. Functions of volume-sensitive and calcium-activated chloride channels. *IUBMB Life* **66**, 257–267 (2014).
36. Larsen, A. Z., Olsen, L. & Kummer, U. On the encoding and decoding of calcium signals in hepatocytes. *Biophys. Chem.* **107**, 83–99 (2004).
37. Prank, K., Gabbiani, F. & Brabant, G. Coding efficiency and information rates in transmembrane signaling. *Biosystems* **55**, 15–22 (2000).
38. Nakano, T. & Liu, J.-Q. Information transfer through calcium signaling In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (ed. Schmid, A., Goel, S., Wang, W., Beiu, V., Carrara, S.) 29–33 (Springer, 2009).
39. Kazachenko, V. N., Astashev, M. E. & Grinevic, A. A. Multifractal analysis of K^+ channel activity. *Biochemistry* **2**, 169–175 (2007).
40. Dascal, N. The use of *Xenopus* oocytes for the study of ion channels. *CRC. Crit. Rev. Biochem.* **22**, 317–387 (1987).
41. Eke, A. *et al.* Physiological time series: distinguishing fractal noises from motions. *Pflugers Arch.* **439**, 403–415 (2000).
42. Caccia, D. C., Percival, D. B., Cannon, M. J., Raymond, G. M. & Bassingthwaight, J. B. Analyzing exact fractal time series: evaluating dispersional analysis and rescaled range methods. *Physica A* **246**, 609–632 (1997).
43. Cannon, M. J., Percival, D. B., Caccia, D. C., Raymond, G. M. & Bassingthwaight, J. B. Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A* **241**, 606–626 (1997).
44. Liebovitch, L. S. & Yang, W. Transition from persistent to anti-persistent correlation in biological systems. *Phys. Rev. E* **56**, 4557–4566 (1997).
45. De la Fuente, I. M. Elements of the cellular metabolic structure. *Front. Mol. Biosci.* **2**, 16 (2015).
46. De la Fuente, I. M., Cortes, J. M., Pelta, D. A. & Veguillas, J. Attractor metabolic networks. *PLoS One* **8**, e58284 (2013).
47. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. USA.* **79**, 2554–2558 (1982).
48. Amit, D. J. *Modeling Brain Function. – The World Of Attractor Neural Networks* (Cambridge University Press, 1989).
49. Hertz, J., Krogh, A. & Palmer, R. G. *Introduction To The Theory Of Neural Computation.* (Addison-Wesley Longman Publishing Co, 1991).
50. Barbour, B., Brunel, N., Hakim, V. & Nadal, J. P. What can we learn from synaptic weight distributions? *Trends Neurosci.* **30**, 622–629 (2007).
51. Bandeira, H. T., Barbosa, C. T., De Oliveira, R. A., Aguiar, J. F. & Nogueira, R. A. Chaotic model and memory in single calcium-activated potassium channel kinetics. *Chaos* **18**, 033136 (2008).
52. Varanda, W. A., Liebovitch, L. S., Figueiroa, J. N. & Nogueira, R. A. Hurst analysis applied to the study of single calcium-activated potassium channel kinetics. *J. Theor. Biol.* **206**, 343–353 (2000).
53. Wawrzkiwicz, A., Pawelek, K., Borys, P., Dworakowska, B. & Grzywna, Z. J. On the simple random-walk models of ion-channel gate dynamics reflecting long-term memory. *Eur. Biophys. J.* **41**, 505–526 (2012).
54. Ludington, W. B., Wemmer, K. A., Lechtreck, K. F., Witman, G. B. & Marshall, W. F. Avalanche-like behavior in ciliary import. *Proc. Nat. Acad. Sci. USA* **110**, 3925–3930 (2013).
55. Ramanujan, V. K., Biener, G. & Herman, B. A. Scaling behavior in mitochondrial redox fluctuations. *Biophys. J.* **90**, L70–L72 (2006).
56. Aon, M. A. *et al.* The scale-free dynamics of eukaryotic cells. *PLoS One* **3**, e3624 (2008).
57. Tigyi, G., Dyer, D., Matute, C. & Miledi, R. A serum factor that activates the phosphatidylinositol phosphate signaling system in *Xenopus* oocytes. *Proc. Nat. Acad. Sci. USA* **87**, 1521–1525 (1990).
58. Stanley, H. E. *et al.* Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. *Physica A* **224**, 302–321 (1996).
59. G. H. Weiss. *Aspects And Applications Of The Random Walk.* (North-Holland, 1994).
60. Hurst, H. Long term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers* **6**, 770–799 (1951).
61. Mandelbrot, B. & Wallis, J. R. Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. *Water Resources Research* **5**, 967–988 (1969).
62. Voss, R. F. *Fundamental Algorithms in Computer Graphics.* (Earnshaw, R. A. ed.) 805–835 (Berlin: Springer-Verlag, 1985).
63. Bassingthwaight, J. B. & Raymond, G. M. Evaluation of the dispersional analysis method for fractal time series. *Ann. Biomed. Eng.* **23**, 491–505 (1995).
64. Raimond, G. Fractal analysis programs of the national simulation resource. Physiome project <http://www.physiome.org/software/fractal/> (2013).
65. Peng, C. K., Havlin, S., Stanley, H. E. & Goldberger, A. L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **5**, 82–87 (1995).
66. Hardstone, R. *et al.* Detrended fluctuation analysis: a scale-free view on neuronal oscillations. *Frontiers in physiology* **30**, 450 (2012).
67. Gao, Z. K. & Jin, N. D. A directed weighted complex network for characterizing chaotic dynamics from time series. *Nonlinear Anal. Real.* **13**, 947–952 (2012).
68. Gao, Z. K. *et al.* Multi-frequency complex network from time series for uncovering oil-water flow structure. *Sci. Rep.* **5**, 8222 (2015).
69. Gao, Z. K. *et al.* Characterizing slug to churn flow transition by using multivariate pseudo Wigner distribution and multivariate multiscale entropy. *Chem. Eng. J.* **291**, 74–81 (2016).
70. Gao, Z. K. *et al.* Multiscale limited penetrable horizontal visibility graph for analyzing nonlinear time series. *Sci. Rep.* **6**, 35622 (2016).
71. Gao, Z. K., Yang, Y. X., Zhai, L. S., Jin, N. D. & Chen, G. R. A four-sector conductance method for measuring and characterizing low-velocity oil-water two-phase flows. *IEEE Transactions on Instrumentation and Measurement* **65**, 1690–1697 (2016).

Author Contributions

Conceived and designed the experiments: I.M.D.F. Performed the experiments: A.P.S., A.V., R.A. Analyzed the data: I.M.D.F., I.M., A.P.S., J.M.C., C.B., L.M., R.A. Contributed reagents/materials/analysis tools/biological discussion: I.M.D.F., I.M., A.P.S., J.M.C., M.D.B., G.P.Y., C.B., M.F., A.V., L.M., R.A. Wrote the paper: I.M.D.F., I.M., A.P.S., J.M.C., L.M., R.A. Supervised the research: I.M.D.F.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: De la Fuente, I. M. *et al.* Dynamic properties of calcium-activated chloride currents in *Xenopus laevis* oocytes. *Sci. Rep.* 7, 41791; doi: 10.1038/srep41791 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017



On the Dynamics of the Adenylate Energy System: Homeorhesis vs Homeostasis

Ildefonso M. De la Fuente^{1,2,3,4*}, Jesús M. Cortés^{4,5}, Edelmira Valero⁶, Mathieu Desroches⁷, Serafim Rodrigues⁸, Iker Malaina^{9,4}, Luis Martínez^{2,4}

1 Institute of Parasitology and Biomedicine "López-Neyra", CSIC, Granada, Spain, **2** Department of Mathematics, University of the Basque Country UPV/EHU, Leioa, Spain, **3** Unit of Biophysics (CSIC, UPV/EHU), and Department of Biochemistry and Molecular Biology University of the Basque Country, Bilbao, Spain, **4** Biocruces Health Research Institute, Hospital Universitario de Cruces, Barakaldo, Spain, **5** Ikerbasque: The Basque Foundation for Science, Bilbao, Basque Country, Spain, **6** Department of Physical Chemistry, School of Industrial Engineering, University of Castilla-La Mancha, Albacete, Spain, **7** INRIA Paris-Rocquencourt Centre, Paris, France, **8** School of Computing and Mathematics, University of Plymouth, Plymouth, United Kingdom, **9** Department of Physiology, University of the Basque Country UPV/EHU, Bilbao, Spain

Abstract

Biochemical energy is the fundamental element that maintains both the adequate turnover of the biomolecular structures and the functional metabolic viability of unicellular organisms. The levels of ATP, ADP and AMP reflect roughly the energetic status of the cell, and a precise ratio relating them was proposed by Atkinson as the adenylate energy charge (AEC). Under growth-phase conditions, cells maintain the AEC within narrow physiological values, despite extremely large fluctuations in the adenine nucleotides concentration. Intensive experimental studies have shown that these AEC values are preserved in a wide variety of organisms, both eukaryotes and prokaryotes. Here, to understand some of the functional elements involved in the cellular energy status, we present a computational model conformed by some key essential parts of the adenylate energy system. Specifically, we have considered (I) the main synthesis process of ATP from ADP, (II) the main catalyzed phosphotransfer reaction for interconversion of ATP, ADP and AMP, (III) the enzymatic hydrolysis of ATP yielding ADP, and (IV) the enzymatic hydrolysis of ATP providing AMP. This leads to a dynamic metabolic model (with the form of a delayed differential system) in which the enzymatic rate equations and all the physiological kinetic parameters have been explicitly considered and experimentally tested *in vitro*. Our central hypothesis is that cells are characterized by changing energy dynamics (*homeorhesis*). The results show that the AEC presents stable transitions between steady states and periodic oscillations and, in agreement with experimental data these oscillations range within the narrow AEC window. Furthermore, the model shows sustained oscillations in the Gibbs free energy and in the total nucleotide pool. The present study provides a step forward towards the understanding of the fundamental principles and quantitative laws governing the adenylate energy system, which is a fundamental element for unveiling the dynamics of cellular life.

Citation: De la Fuente IM, Cortés JM, Valero E, Desroches M, Rodrigues S, et al. (2014) On the Dynamics of the Adenylate Energy System: Homeorhesis vs Homeostasis. PLoS ONE 9(10): e108676. doi:10.1371/journal.pone.0108676

Editor: Marie-Joelle Virolle, University Paris South, France

Received: January 21, 2014; **Accepted:** September 3, 2014; **Published:** October 10, 2014

Copyright: © 2014 De la Fuente et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was funded by the University of Basque Country (UPV/EHU): University-Society grant US11/13 and Ministerio de Economía y Competitividad (Spain), Project No. BFU2013-44095-P. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: mtpmadei@ehu.es

Introduction

Living cells are essentially highly evolved dynamic reactive structures, in which the most complex known molecules are synthesized and destroyed by means of a sophisticated metabolic network characterized by hundreds to thousands of biochemical reactions, densely integrated, shaping one of the most complex dynamic systems in nature [1,2].

Energy is the fundamental element for the viability of the cellular metabolic network. All cells demand a large amount of energy to keep the entropy low in order to ensure their self-organized enzymatic functions and to maintain their complex biomolecular structures. For instance, during growth conditions it has been observed that in microbial cells the protein synthesis accounts for 75% of the total energy, and the cost of DNA replication accounts for 2% of the energy [3,4].

Although different nucleosides can bind to three phosphates which may serve to store biochemical energy i.e., GTP, (d)CTP, (d)TTP and (d)UTP [5], there exists a consensus that adenosine 5'-

triphosphate (ATP) is the principal molecule for storing and transferring energy in cells. All organisms, from the simplest bacteria to human cells, use ATP (Mg-ATP) as their major energy source for metabolic reactions [6–8], and the levels of ATP, ADP and AMP reflect roughly the energetic status of the cell [7]. ATP is originated from different classes of metabolic reactions, mainly substrate-level phosphorylation, cellular respiration, photophosphorylation and fermentation, and it is used by enzymes and structural proteins in all main cytological processes, i.e., motility, cell division, biosynthetic reactions, cell cycle, allosteric regulations, and fast synaptic modulation [7–9].

In the living cell, practically all bioenergetic processes are coupled with each other via adenosine nucleotides, which are consumed or regenerated by the different enzymatic reactions. In fact, the most important regulatory elements involved in the coupling of catabolic and anabolic reactions are ATP, ADP and AMP [7]. The adenosine nucleotides are not only tied to the metabolic pathways involved in the cell's energetic system but also

act as allosteric control of numerous regulatory enzymes allowing that changes in ATP, ADP and AMP levels can practically regulate the functional activity of the overall multienzymatic network of cell [10–13].

A characteristic of the temporal evolution of ATP, ADP and AMP concentrations is their complexity [14]. Extensive experimental studies have shown that metabolism exhibits extremely large and complex fluctuations in the concentrations of individual adenosine nucleotides, which are anything but stationary [14–16]. In fact, under normal conditions inside the cell, the time evolution of the adenosine-5'-triphosphate is subjected to marked variations presenting transitions between quasi-steady states and oscillatory behaviors [15,16]. For instance, complex ATP rhythms were reported to occur in: myxomycetes [17,18], neurons [19], yeast [16], embryonic cells [20,21], myocytes [22], islet β -cells [23,24], keratinocytes [25], hepatocytes [26], red blood cells [27] and L and MEL cells [28]. Many of these oscillations have clearly non-periodic behaviors [19,26], and ADP and AMP also exhibit complex oscillatory patterns [29–31]. In addition to ATP ultradian oscillations, specific circadian rhythms have also been reported, which occur with a period close to 24 hours (the exogenous period of the Earth's rotation) [27,32,33].

Oscillatory behavior is a very common phenomenon in the temporal dynamics of the concentration for practically all cell metabolites. Indeed, during the last four decades, the studies of biochemical dynamical behaviors, both in prokaryotic and eukaryotic organisms, have shown that in cellular conditions spontaneous molecular oscillations emerge in most of the fundamental metabolic processes. For instance, specific biochemical oscillations were reported to occur in: free fatty acids [34], NAD(P)H concentration [35], biosynthesis of phospholipids [36], cyclic AMP concentration [37], actin polymerization [38], ERK/MAPK metabolism [39], mRNA levels [31], intracellular free amino acid pools [40], cytokinins [41], cyclins [42], transcription of cyclins [43], gene expression [44–47], microtubule polymerization [48], membrane receptor activities [49], membrane potential [50,51], intracellular pH [52], respiratory metabolism [53], glycolysis [54], intracellular calcium concentration [55], metabolism of carbohydrates [56], beta-oxidation of fatty acids [57], metabolism of mRNA [58], tRNA [59], proteolysis [60], urea cycle [61], Krebs cycle [62], mitochondrial metabolic processes [63], nuclear translocation of the transcription factor [64], amino acid transports [65], peroxidase-oxidase reactions [66], protein kinase activities [67] and photosynthetic reactions [68]. In addition, experimental observations in *Saccharomyces cerevisiae* during continuous culture have shown that the majority of metabolome also shows oscillatory dynamics [69].

Persistent properties in oscillatory behaviours have also been observed in other studies, e.g., DNA sequences [70–71], NADPH series [72], K^+ channel activity [73], biochemical processes [74,75], physiological time series [76,77], and neural electrical activity [78,79].

Likewise, it has been observed that genomic activity shows oscillatory behavior. For instance, under nutrient-limited conditions yeast cells have at least 60% of all gene expressions oscillating with an approximate period of 300 min [80]. Other experimental observations have shown that practically the entire transcriptome exhibits low-amplitude oscillatory behavior [81] and this phenomenon has been described as a genomewide oscillation [47,81–83].

At a global metabolic level, experimental studies have shown that the cellular metabolic system resembles a complex multi-oscillator system [69,81,83], what allows for interpretation that the cell is a complex metabolic network in which multiple autonomous

oscillatory and quasi-stationary activity patterns simultaneously emerge [84–89].

Cells are open dynamic systems [90,91], and when they are exposed to unbalanced conditions, such as metabolic stress, physiological processes produce drastic variations both in the concentration of the adenosine nucleotides [15,16,92,93] and in their molecular turnovers [94]. Tissues such as skeletal and cardiac muscles must sustain very large-scale changes in ATP turnover rate during equally large changes in work. In many skeletal muscles, these changes can exceed 100-fold [95].

The ratio of ATP, ADP and AMP is functionally more important than the absolute concentration of ATP. Different ratios have been used as a way to test the metabolic pathways which produce and consume ATP. In 1967, Atkinson proposed a simple index to measure the energy status of the cell, defined as $AEC = ([ATP] + 0.5[ADP]) / ([ATP] + [ADP] + [AMP])$ [96].

The AEC is a scalar index ranging between 0 and 1. When all adenine nucleotide pool is in form of AMP the energy charge (AEC) is zero, and the system is completely discharged (zero concentrations of ATP and ADP). With only ADP, the energy charge is 0.5. If all adenine nucleotide pool is in form of ATP the AEC is 1.

The first experimental testing of this equation showed that (despite of extremely large fluctuations in the adenosine nucleotide concentrations), many organisms under optimal growth conditions maintained their AEC within narrow physiological values, between $AEC = 0.7$ and $AEC = 0.95$, stabilizing in many cases at a value close to 0.9. Atkinson and coauthors concluded that for these values of AEC, the major ATP-producing reactions are in balance with the major ATP-consuming reactions; for very unfavorable conditions the AEC drops off provoking cells to die [97–101].

During the last four decades, extensive biochemical studies have shown that the narrow margin of the AEC values is preserved in a wide variety of organisms, both eukaryotes and prokaryotes. For instance, AEC values between 0.7 and 0.95 have been reported to occur in cyanobacteria [102,103], mollicutes (mycoplasmas) [104], different bacteria both gram positive and gram negative as *Dimoroseobacter shibae* [105], *Streptococcus lactis* [106], *Bacillus licheniformis* [107], *Thermoactinomyces vulgaris* [108], *Escherichia coli* [109], *Myxococcus xanthus* [110] and *Myxococcus Coralloides* [111], different eukaryotic cells as zooplankton [112], algae [113], yeast [114], neurons [115,116], erythrocytes [117], astrocytes [118], platelets [119], spermatozoa [120], embryonic kidney cells (HEK) [121], skeletal muscle [122], liver tissue [123], fungi [124], different microorganisms of mangrove soils and water (fungi, bacteria and algae) [125] and plants [126–128].

Studies of different species of plants, over long periods of time, have demonstrated a close relationship between AEC and cellular growth, e.g. leaf tissue collected bimonthly from *Spartina patens*, *S. cynosuroides*, *S. alterniflora* and *Distichlis spicata* showed that the adenylate energy charge peaked in spring and summer at 0.78–0.85 and then declined in late summer and early fall [129]. In the case of organisms better adapted to cold, such as winter wheat cells (*Triticum aestivum*) that are cultivated from September to December in the Northern Hemisphere, the ATP levels were shown to decrease gradually when the cells were exposed to various low temperature stresses (ice encasement at -1°C); however, even after 5 weeks of icing when cell viability was severely reduced, AEC values remained high, about 0.8 [130].

There is a long history of quantitative modelling of ATP production and turnover, dating back to Sel'kov's model on glycolytic energy production from 1968 [131], later developed by Goldbeter [132], as well as by Heinrich and Rapoport [133]. In

this context, Sel'kov also published a kinetic model of cell energy metabolism with autocatalytic reaction sequences for glycolysis and glycogenolysis in which oscillations of the adenylate energy charge were observed [134].

However, the first adenylate energy system was developed by Reich and Sel'kov in 1974 [135]. This system was modeled with first-order kinetics by using ordinary differential equations.

Here, in order to further understanding of the elements that determine the cellular energy status of cells we present a computational model conformed by some key essential parts of the adenylate energy system. Specifically, the model incorporates (I) the main synthesis process of ATP for cell from ADP (ATP synthase), (II) the catalyzed phosphotransfer reaction for interconversion of adenine nucleotides (ATP, ADP and AMP) (adenylate kinase), (III) the enzymatic hydrolysis of ATP yielding ADP (kinase and ATPase reactions) and (IV) the enzymatic hydrolysis of ATP providing AMP (enzymatic processes of synthetases). The metabolic model has been analyzed by using a system of delay differential equations in which the enzymatic rate equations and all the physiological kinetic parameters have been explicitly considered and experimentally tested *in vitro* by other groups. We have used a system of delay-differential equations fundamentally to model the asynchronous metabolite supplies to the enzymes.

The numerical analysis shows that the AEC can perform transitions between oscillations and steady state patterns in a stabilized way, similar to what happens in the prevailing conditions inside the cell. The max and min values of the oscillations range within a physiological window validated by experimental data.

We finally suggest that rather than a permanent physiological stable state (*homeostasis*), the living systems seem to be characterized by changing energy dynamics (*homeorhesis*).

Methods

Cells require a permanent generation of energy flow to keep the functionality of its complex metabolic structure which integrates a large ensemble of enzymatic processes, interconnected by a network of substrate fluxes and regulatory signals [4].

To understand some elements that determine the energy status of cells we have studied the dynamics of the main biochemical reactions interconverting ATP, ADP and AMP. Specifically, we have developed a model for the basic structure of the adenylate energy system which represents the fundamental biochemical reactions interconverting ATP, ADP and AMP coupled to the main fluxes of adenine nucleotides involved in catabolic and anabolic processes (Figure 1).

The essential metabolic processes incorporated into the adenylate energy model are the following:

I. First, we have assumed the oxidative phosphorylation as the main synthesis source of ATP in the cell.

As is well known, the enzymatic oxidation of nutrients generates a flow of electrons to O_2 through protein complexes located in the mitochondrial inner membrane in eukaryotes, and in the cell intermembrane space in prokaryotes, that leads to the pumping of protons out of the matrix. The resulting uneven distribution of protons generates a pH gradient that creates a proton-motive force. This proton gradient is converted into phosphoryl transfer potential by ATP synthase which uses the energy stored in the electrochemical gradient to drive the synthesis of ATP from ADP and phosphate (P_i) [7]. Thus, oxidative phosphorylation is the culmination of a series of complex enzymatic transformations whose final phase is carried out by ATP synthase.

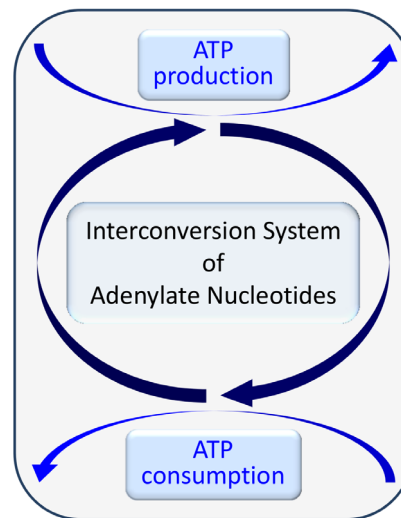
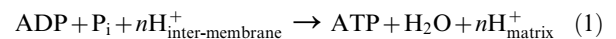


Figure 1. Elemental biochemical processes involved in the energy status of cells. The synthesis sources of ATP are coupled to energy-consumption processes through a network of enzymatic reactions which, interconverting ATP, ADP and AMP, shapes a permanent cycle of synthesis-degradation for the adenine nucleotides. This dynamic functional structure defines the elemental processes of the adenylate energy network, a thermodynamically open system able to accept, store, and supply energy to cells. doi:10.1371/journal.pone.0108676.g001

Experimental studies in non-pathologic cells have shown that ATP synthase generates the vast majority of cellular energy in the form of ATP (more than 90% in human cells) [136]; consequently, it is one of the central enzymes in energy metabolism for most cellular organisms, both prokaryotes and eukaryotes. This sophisticated rotatory macromolecular machine is embedded in the inner membrane of the mitochondria, the thylakoid membrane of chloroplasts, and the plasma membrane of bacteria [137].

The overall reaction sequence for the ATP synthase is:



where n indicates the H^+/ATP ratio with values between 2 and 4 which have been reported as a function of the organelle under study [138].

II. Besides the oxidative phosphorylation, we have also considered that in optimal growth conditions a small part of ATP is generated through substrate-level phosphorylation [7].

III. Another essential metabolic process for cellular energy is the catalyzed phosphotransfer reaction performed by the enzyme adenylate kinase, which is required for interconversion of adenine nucleotides.

Almost since its discovery, about 60 years ago, adenylate kinase (phosphotransferase with a phosphate group as acceptor) has been considered to be a key enzyme in energy metabolism for all organisms [139–140]. This enzyme catalyzes the following reversible reaction for the interconversion of ATP, ADP and AMP:

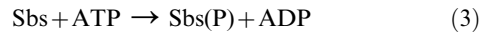


Adenylate kinase catalyzes the interconversion of the adenine nucleotides and so it is an important factor in the regulation of the

adenine nucleotide ratios in different intracellular compartments, i.e. it contributes to regulate the adenylate energy charge in cells. The equilibrium will be shifted to the left or right depending on the relative concentrations of the adenine nucleotides. In contrast, ATP synthase catalyzes the *de novo* synthesis of the vast majority of ATP from ADP and Pi [136].

IV. The next catalytic process that we have considered corresponds to the enzymes implied in the hydrolysis of ATP to form ADP and orthophosphate (P_i). The chemical energy that is stored in the high-energy phosphoanhydridic bonds in ATP is released, ADP being a product of its catalytic activity.

The basic reaction sequence for the enzymatic process is:



where Sbs and Sbs(P) are the substrate and the product of the catalytic process, respectively. In this kind of metabolic reaction different groups of enzymes are involved, mainly kinases and ATPases. Particularly, kinases catalyze the transfer of a phosphoryl group from ATP to a different class of specific molecules, which may be also a protein. By adding phosphate groups to substrate proteins, the kinases enzymes shape the activity, localization and overall function of many proteins and pathways, which orchestrate the activity of almost all cellular processes. Up to 30% of all human proteins may be modified by a kinase activity, and they regulate the majority of cellular pathways, especially those involved in signal transduction [141]. These enzymes are fundamental for the functional regulation of the cellular metabolic network and they constitute one of the largest and most diverse gene families. The human genome contains about 500 protein kinase genes and they constitute about 2% of all human genes [141].

V. Finally, we have taken into account the ligase enzymes that catalyze the joining of smaller molecules to make larger ones, coupling the breakdown of a pyrophosphate bond in ATP to provide AMP and pyrophosphate as main products.

The basic reaction sequence for the ligases is:



The enzymes belonging to the family of ligases involve different groups as DNA ligases, aminoacyl tRNA synthetases, ubiquitin ligase, etc. They are very important catalytic machines for anabolic processes and for the molecular architecture of the cell. Most ligases are mainly implied in the protein synthesis consuming a large part of the cellular ATP. Thus, for microbial cells, the protein synthesis accounts for 75% of the total energy during growth conditions [3,4].

Protein synthesis uses energy mainly from ATP at several stages such as the attachment of amino acids to transfer RNA, and the movement of mRNA through ribosomes, resulting in the attachment of new amino acids to the chain. In these processes, aminoacyl tRNA synthetases constitute an essential enzyme super-family, providing fidelity of the translation process of mRNA to proteins in living cells and catalyzing the esterification of specific amino acids and their corresponding tRNAs. They are common to all classes of organisms and are of utmost importance for all cells [142]. In the present model we have considered aminoacyl tRNA synthetase as a representative enzyme of the ligases group.

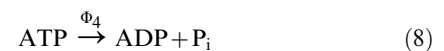
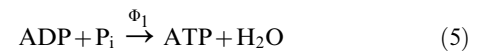
Figure 2 schematically shows the enzymatic processes of the ATP consuming-generating system. First, a permanent input of nutrients is considered to be the primary energy source. In the

final phase of oxidative phosphorylation, the ATP synthase uses the energy stored in the proton gradient, generated by the enzymatic oxidation of nutrients, to drive the synthesis of ATP from ADP and phosphate (P_i). The flow of protons thus behaves like a gear that turns the rotary engine of ATP synthase. Likewise, a small part of ATP is also incorporated into the system via substrate-level phosphorylation. The ATP synthesized is fundamentally consumed by two different enzymatic reactions: (i) the ligase processes which provide the system with AMP molecules and (ii) the kinase and ATPase reactions which mainly generate ADP. The interconversion of ATP, ADP and AMP is performed by the enzyme adenylate kinase, which regenerates them according to the dynamic needs of the system.

The ATP consuming-generating system is open and consequently some AMP molecules are *de novo* biosynthesized [143]; whilst a part of AMP does not continue in the reactive system due to its hydrolysis, forming adenine and ribose 5-phosphate [144]. Finally, according to experimental observations, we have considered that a very small part of ATP does not remain in the system, but is drained out from the cell [145–149].

We want to emphasize that the biochemical energy system depicted in Figure 2 represents some key essential parts of the adenylate energy system (see for more details the end of the “Model Section”), which constitute a thermodynamically open system able to accept, store, and supply energy to cells.

This metabolic network of crucial biochemical processes for the cell can be rewritten in a simplified way to gain a better understanding about the dynamic behavior of the model:



where Φ_i ($i=1-5$) are the rates of the enzymatically-catalyzed reactions (5) to (9), v_1 is the rate of the ATP input into the system by substrate-level phosphorylation, v_2 is the rate of the ATP output from the cell [145–149], being $v_2 = k_2[\text{ATP}]$, v_3 is the rate of the biosynthesis *de novo* of AMP and v_4 is the rate of the sink of AMP, being $v_4 = k_4[\text{AMP}]$. The reversible adenylate kinase reaction (2) has been described by its corresponding reactions (6) and (7) linked by a control parameter (see below for more details) allowing to move the reactive process to either of the two reactions according

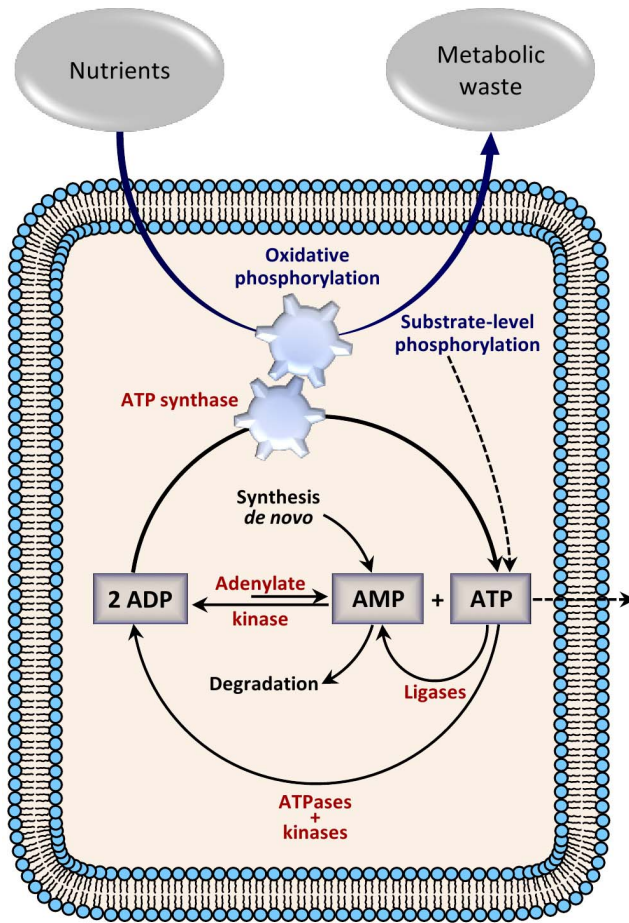


Figure 2. The Adenylate energy system. Oxidative phosphorylation and substrate-level phosphorylation generate ATP which is degraded by kinases (also ATPases) and ligases yielding ADP and AMP, respectively. The three adenine nucleotides are catalytically interconverted by adenylate kinase according to the needs of the metabolic system. AMP is also subjected to processes of synthesis-degradation, some AMP molecules are *de novo* biosynthesized, and a part of AMP is hydrolyzed. According to experimental observations, a very small number of ATP molecules may not remain in the adenylate reactive structure. The system (thermodynamically open) needs a permanent input of nutrients as primary energy source and a consequent output of metabolic waste. The biochemical energy system depicted in the figure represents some key essential parts of the adenylate energy system.

doi:10.1371/journal.pone.0108676.g002

to the physiological needs of the system, i.e. the synthesis or the consumption of ATP or ADP. According to the stoichiometry of this set of chemical equations, there is a net consumption of ATP in the system, which can be regulated by reactions (6), (8), (9) and (10), as well as a production of AMP, which is regulated by steps (7), (9) and (11).

Although the kinetic behavior *in vivo* of most enzymes is unknown, *in vitro* studies can provide both adequate kinetic parameters and enzymatic rate functions. We have used this strategy to implement the dynamical model of the adenylate energy system. Thus, for ATP synthase we have assumed Michaelis–Menten kinetics with competitive inhibition by the product [150]. An iso-random Bi Bi mechanism has been reported for adenylate kinase kinetics [151–152]. We have also considered that a fraction of the adenylate kinases exhibit the balance shifted to the left and simultaneously the rest of the adenylate kinase

macromolecules present a balance shifted to the right, depending their catalytic activities on the system demand. For the kinase family we have selected phosphofructokinase, whose rate equation was developed in the framework of concerted transition theory of Monod and Changeux [153,154], and finally, for the ligase family we have chosen threonyl-tRNA synthetase, which shows Michaelis–Menten kinetics [155].

The time-evolution of the ATP consuming-generating system (Figure 2) can be described by the following three differential equations:

$$\frac{d\alpha}{dt} = v_1 + \lambda\sigma_1\Phi_1 - \Delta'\sigma_2\Phi_2 + \Delta''\sigma_3\Phi_3 - \sigma_4\Phi_4 - \sigma_5\Phi_5 - v_2,$$

$$\frac{d\beta}{dt} = -\lambda\sigma_1\Phi_1 + \Delta'\sigma_2\Phi_2 - \Delta''\sigma_3\Phi_3 + \sigma_4\Phi_4,$$

$$\frac{d\gamma}{dt} = v_3 - \Delta'\sigma_2\Phi_2 + \Delta''\sigma_3\Phi_3 + \sigma_5\Phi_5 - v_4 \quad (12)$$

where the variables α , β and γ denote the ATP, ADP and AMP concentrations respectively, $\sigma_1, \dots, \sigma_5$ correspond to the maximum rates of the reactions (5) – (9), respectively, the nutrients are injected at a constant rate and λ is a control parameter related to the energy level stored in the proton gradient generated by the enzymatic oxidation of input nutrients. Δ' and Δ'' are also control parameters in the system regulating adenylate kinase activity towards the synthesis or the consumption of ADP, respectively, with $\Delta' = 2\Delta''$.

The enzymatic rate functions are the following:

$$\Phi_1 = \frac{\beta}{\beta + K_{m,1} \left(1 + \frac{\alpha}{K_{I,1}}\right)} \quad (13)$$

$$\Phi_2 = \frac{\alpha\gamma}{K_2 + K_{m,2}^{ATP}\gamma + K_{m,2}^{AMP}\alpha + \alpha\gamma} \quad (14)$$

$$\Phi_3 = \frac{\beta^2}{K_3 + 2K_{m,3}^{ADP}\beta + \beta^2} \quad (15)$$

$$\Phi_4 = \frac{\alpha(1+\alpha)(1+\beta)^2}{L_4 + (1+\alpha)^2(1+\beta)^2} \quad (16)$$

$$\Phi_5 = \frac{\alpha}{K_{m,5} + \alpha} \quad (17)$$

where $K_{m,1}, K_{m,2}^{ATP}, K_{m,2}^{AMP}, K_{m,3}^{ADP}$ and $K_{m,5}$ are the Michaelis constants for each respective enzyme, $K_{I,1}$ is the dissociation constant of the ADP-ATP synthase complex, K_2 and K_3 are kinetic parameters of the adenylate kinase, α and β in Eq. (16) are divided by $1 \mu\text{M}$ so that this equation is dimensionally homogeneous, and L_4 is the allosteric constant of phosphofructokinase. More details about the kinetic parameters and experimental references are given in Table 1.

Table 1. Values of the kinetic parameters used to simulate some of the dynamics of the adenylate energy system.

Parameter	Value	Reference
σ_1	7.14 $\mu\text{mol s}^{-1}$	[166]
$K_{m,1}$	30 μmol	[150]
$K_{I,1}$	25 μmol	[150]
σ_2	800 $\mu\text{mol s}^{-1}$	[167]
K_2	71000 μmol^2	[151]
$K_{m,2}^{ATP}$	25 μmol	[151]
$K_{m,2}^{AMP}$	110 μmol	[151]
σ_3	800 $\mu\text{mol s}^{-1}$	[168]
K_3	1360 μmol^2	[152]
$K_{m,3}^{ADP}$	29 μmol	[152]
σ_4	100 $\mu\text{mol s}^{-1}$	[132]
L_4	10^6	[169]
σ_5	0.43 $\mu\text{mol s}^{-1}$	[155]
$K_{m,5}$	100 μmol	[155]

doi:10.1371/journal.pone.0108676.t001

These equations are simplified expressions, but they are particularly useful in the analysis of models of dynamic behavior [154]. For simplification, we do not consider orthophosphate molecules, nor the H_2O involved in the reaction (5), which has been omitted because the solvent has a standard state of 1M.

To study the system dynamics, the model here described has been analyzed by means of a system of delay differential equations accounting for the delays in the supplies of adenine nucleotides to the specific enzymes involved in the biochemical model.

Generally in the cellular metabolic networks the enzymatic processes are not coupled instantaneously between them. The metabolic internal medium is a complex, crowded environment [156], where the dynamic behavior of intracellular metabolites is controlled by a wide mixture of specific interactions and physical constraints mainly imposed by the viscosity of the cellular plasma, mass transport across membranes and variations in the diffusion times which are dependent on the physiological cellular context [157–160].

For example, there is a time-running from the instant in which ATP molecules are produced in the mitochondria until they come to the place where they are used by the target enzymes. Sometimes the spatial separations may involve long intracellular macroscopic distances. As a result of these intracellular phenomena (transport across membranes, diffusion, long macroscopic distances, interactions with the internal molecular crowded, etc.), the supply of metabolites to the enzymes (substrates and regulatory molecules) occurs in different time scales, and with different delays.

Time scales in biochemical systems mean an asynchronous temporal structure characterized by different magnitudes of metabolite supply delays associated to specific enzymatic processes.

Moreover, experimental studies have shown that metabolism exhibits complex oscillations in the concentrations of individual adenine nucleotides, with periods from seconds to several minutes [15,16], which shape a complex temporal structure for intracellular ATP/ADP/AMP concentrations. The phase shifts in this temporal structure also originate delays in the supplies of

substrates and regulatory molecules to the specific enzymes [161–165].

Consequently, metabolic reactions involving ATP/ADP/AMP may occur at different characteristic time scales, ranging from seconds to minutes, originating a temporal structure for intracellular ATP/ADP/AMP concentrations within the cell.

Dynamic processes with delay cannot be modeled using systems of ordinary differential equations. The different time scales can be considered with delay differential equations, which are not ordinary differential equations. In these systems, some dependent variables can be evaluated in terms of $(t-r_i)$ where r_i are the delays and t the time, and consequently the metabolite supplies to the enzymes (substrates and regulatory molecules) are not instantaneous; other dependent variables may be evaluated in terms of t ($r_i=0$), if metabolite supplies are considered instantaneous.

According to these regards, we have analyzed our system with three delayed variables $\alpha(t-r_1)$, $\beta(t-r_2)$ and $\gamma(t-r_3)$. r_1 models the delay in the supply of ATP to its specific enzymes; r_2 does the same for ADP and r_3 for AMP. Nevertheless, we have assumed that ATP concentration ($\alpha(t)$) in the equation corresponding to ATP synthase (Eq (18)) is not delayed, as this product formation can be considered instantaneous with respect to the competitive inhibition of the enzyme by the same ATP. Likewise, since the adenylate kinase enzyme is reversible, the ADP formed from ATP and AMP in the reaction (6) is used by the reaction (7) in the same place, and therefore, we have also considered that ADP concentration ($\beta(t)$) is not delayed in this process (Eq (20)).

Therefore, the adenylate energy system exhibits several time scales and we have used the system of delay-differential equations to model the asynchronous metabolite supplies to the enzymes. In some processes it can be considered that the substrate or regulatory molecules instantly reach the enzyme and in other processes there are delays for substrate supplies to them.

According to these kinds of dependent variables in the system, the enzymatic rate functions are written as follows:

$$\Phi_1 = \frac{\beta(t-r_2)}{\beta(t-r_2) + K_{m,1} \left(1 + \frac{\alpha(t)}{K_{I,1}}\right)} \quad (18)$$

$$\Phi_2 = \frac{\alpha(t-r_1)\gamma(t-r_3)}{K_2 + K_{m,2}^{ATP}\gamma(t-r_3) + K_{m,2}^{AMP}\alpha(t-r_1) + \alpha(t-r_1)\gamma(t-r_3)} \quad (19)$$

$$\Phi_3 = \frac{\beta(t)^2}{K_3 + 2K_{m,3}^{ADP}\beta(t) + \beta(t)^2} \quad (20)$$

$$\Phi_4 = \frac{\alpha(t-r_1)(1 + \alpha(t-r_1))(1 + \beta(t-r_2))^2}{L_4 + (1 + \alpha(t-r_1))^2(1 + \beta(t-r_2))^2} \quad (21)$$

$$\Phi_5 = \frac{\alpha(t-r_1)}{K_{m,5} + \alpha(t-r_1)} \quad (22)$$

Our differential equations system with delay (12) takes the following particular form, up to a permutation of the indexes of the variables:

$$\begin{cases} y'_1(t) = f_1(y_1(t-r_1), y_1(t), \dots, y_j(t-r_j), y_j(t), y_{j+1}(t), \dots, y_n(t)) \\ \vdots \\ y'_n(t) = f_n(y_1(t-r_1), y_1(t), \dots, y_j(t-r_j), y_j(t), y_{j+1}(t), \dots, y_n(t)) \end{cases} \quad (23)$$

where the dependent variable is a n -dimensional vector of the form $y = (y_1, \dots, y_n)$, t being the independent variable. In system (23), the derivatives of y_1, \dots, y_n , evaluated in t , are related to the variables y_1, \dots, y_j , where each y_i with $i \leq j$ appears evaluated in $t - r_i$, being r_i the corresponding delay, and might appear evaluated also in t , and the derivatives are also related to the variables y_{j+1}, \dots, y_n evaluated in t .

Unlike ODE systems, in delayed differential equations, in order to determine a particular solution, it is necessary to give the initial solution in the interval $[t_0, t_0 + \delta]$ with $\delta = \max\{r_1, \dots, r_j\}$. That involves the consideration, in the solution of the system, of the function $f_0[t_0, t_0 + \delta] \rightarrow R^n$ called initial function. It can be observed therefore that infinite degrees of freedom exist in the determination of the particular solutions.

Since in our system simple oscillatory behavior of period 1 emerges from numerical integration, an acceptable approximation to the initial function is a periodic solution.

In the system described by (23), it is possible to take the initial function f_0 equal to any $y(t)$ and, in particular, it can be a periodic function.

With this type of systems, it is possible to take into account dynamic behaviours related to parametric variations linked to the independent variable. The parametric variations r_i affect the independent variable; they represent time delays and can be related to the domains of the initial functions.

Table 1 shows the values of the kinetic parameters involved in the system chosen to run the model. All of these values have been obtained from *in vitro* experiments reported in the scientific literature and they are within the range of the values published in the enzyme database Brenda (<http://www.brenda-enzymes.info/>). For these values, the preliminary integral solutions of the differential equations system (12) show a simple oscillatory behavior of period 1 and as an approximation we have assumed that the initial functions present simple harmonic oscillations in the following form:

$$\alpha_0(t) = C + D \sin(2\pi/P), \quad (24)$$

$$\beta_0(t) = E + F \sin(2\pi/P), \quad (25)$$

$$\gamma_0(t) = G + H \sin(2\pi/P), \quad (26)$$

with $C = 6 \mu\text{mol}$, $D = 2 \mu\text{mol}$, $E = 4 \mu\text{mol}$, $F = 1 \mu\text{mol}$, $G = 7 \mu\text{mol}$, $H = 3 \mu\text{mol}$ and $P = 200$ s. The other parameter values used were $v_1 = 35 \times 10^{-3} \mu\text{mol s}^{-1}$, $k_2 = 9 \times 10^{-5} \text{s}^{-1}$, $v_3 = 1.4 \mu\text{mol s}^{-1}$, $k_4 = 0.69 \text{s}^{-1}$, $\Delta' = 1.98$, $r_1 = 5$ s, $r_2 = 27$ s and $r_3 = 50$ s.

In this paper, we have studied the dynamic behavior of the system under two parametric scenarios:

- In Scenario I, λ is the control parameter, which is related to the energy level stored in the proton gradient generated by the enzymatic oxidation of input nutrients. This scenario represents the main analysis of the paper, and the values used for the kinetic parameters involved in the model are those set out above.

- In Scenario II, the delay r_2 is the control parameter, modeling the time constants for the time delays of ADP, with $v_1 = 3 \times 10^{-3} \mu\text{mol s}^{-1}$, $k_2 = 2 \times 10^{-4} \text{s}^{-1}$, $v_3 = 2.1 \mu\text{mol s}^{-1}$, $\lambda = 1.09$, $r_1 = 3$ s, and all other parameters as indicated in Scenario I.

The extracellular ATP concentration [145–149] is considerably much lower than its intracellular concentration [147], which makes accurate quantification of extracellular levels of ATP an extremely difficult task. Therefore, k_2 ($v_2 = k_2 [\text{ATP}]$) must be a sufficiently small value. The values of k_2 here used have been: $9 \times 10^{-5} \text{s}^{-1}$ in Scenario I and $2 \times 10^{-4} \text{s}^{-1}$ in Scenario II. If we now take an intermediate value for the ATP concentration, 10 nmol for example, the following data are obtained: $v_2 = 9 \times 10^{-7} \mu\text{mol s}^{-1}$ under Scenario I, and $2 \times 10^{-6} \mu\text{mol s}^{-1}$ under Scenario II, which are significantly lower than the values considered for v_1 (the rate of the ATP input into the system by substrate-level phosphorylation): $3.5 \times 10^{-2} \mu\text{mol s}^{-1}$ under Scenario I, and $3.0 \times 10^{-3} \mu\text{mol s}^{-1}$ under Scenario II.

In this paper, we have studied the bifurcation analysis for the two control parameters (λ and r_2) here considered (Scenarios I and II, resp.). Further future studies, beyond the scope of the present work, might consider including other control parameters to understand how the stability of the solutions change along parameters space. Furthermore, the presence of “molecular noise” might also be included as a possibility to achieve non-periodic variability in the ATP/ADP/AMP oscillations.

An important feature of metabolism is the wide range of time scales in which cellular processes occur.

Generally enzymatic reactions take place at high speed e. g., carbonic anhydrase has a turnover number (k_{cat}) of 400,000 to 600,000 s^{-1} [170] and the turnover number for RNA polymerase II is less rapid, about 0.16 s^{-1} [171].

However, many cellular processes occur on a time scale of minutes. For instance, studies in glucose-limited cultures by up- and downshifts of the dilution rate in *Escherichia coli* K-12 have shown time delays of minutes in the metabolic mechanisms involved in the dynamics of the adenylate energy charge exhibiting drastic changes within 2 min after the nutrients dilution [109]. Intracellular concentrations of the adenine nucleotides and inorganic phosphate may present sustained oscillations in the concentrations of the adenine nucleotides with periods around a minute which can originate large temporal variations in the supplies of these substrates and regulatory molecules to the specific enzymes [30]. In addition to the temporal oscillations, sustained chemical redox waves (NAD(P)H– NAD(P)+) are a rather general feature of some cells [90] which may exhibit qualitative changes with wavefronts traveling in opposite directions within ≈ 2 min after the start [172].

It is also known that ATP can evoke fast currents by activation of different purinergic receptors expressed in the plasma membranes of many cells [173]. However, ATP exposure for several minutes can lead to the formation of a high conductance pore permeable for ions and molecules up to 900 Da [174,175]. The activation of some kinases, such as MAPK, occurs with a time scale of minutes [176,177]. Furthermore fructose-2,6-bisphosphate levels are also regulated through cyclic-AMP-based signalling, which occurs on the timescale of minutes [178].

According to these experimental observations, we have analyzed the dynamic behavior of the adenylate system taking into account both instantaneous substrate input conditions and delay times for metabolite supplies, between 1 to 120 seconds, which covers a wide range of cellular physiological processes.

The numerical integration of the system was performed with the package ODE Workbench, developed by Dr. Aguirregabiria

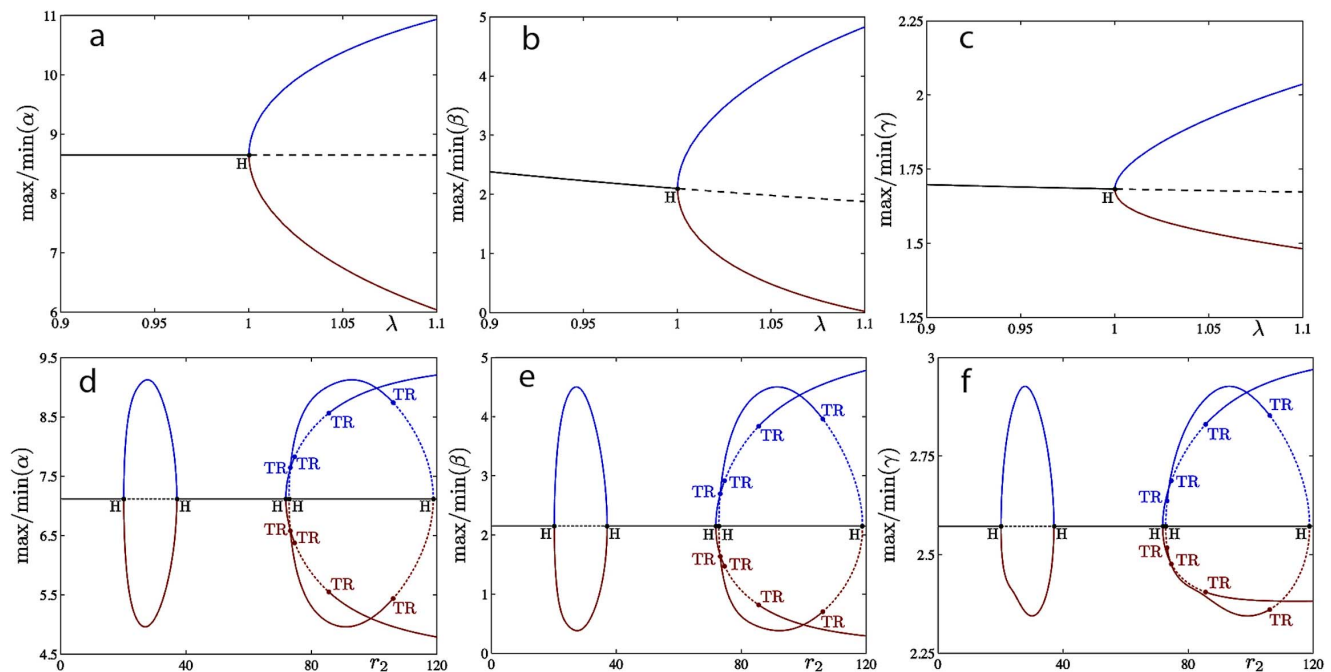


Figure 3. Numerical analysis for the model of the adenylate energy system. a–c: (cf. Scenario I in text) In y-axis we are plotting the max and the min of the different variables α , β and γ . For situations with no oscillations (stable fixed point colored in solid black lines) the max and the min are coincident. For situations with oscillations, the max and the min of the oscillations are plotted separately; in blue we are coloring the max of the oscillation, in red, its minimum value. λ is the control parameter. The numerical integration shows simple solutions. For small λ values ($0.9 \leq \lambda < 1$) the adenine nucleotide concentrations present different stable steady states which lose stability at a Hopf bifurcation at $\lambda \sim 1$. For $\lambda > 1$, the attractor is a stable limit cycle. d–f: (Scenario II) The delay r_2 is the control parameter. The numerical bifurcation analysis reveals that the temporal structure is complex, emerging 5 Hopf bifurcations as well as a secondary bifurcation of Neimark-Sacker type. Two pairs of Hopf bifurcations are connected in the parameter space. A third supercritical Hopf bifurcation occurs at $r_2 \sim 71.94$, rapidly followed by another Hopf bifurcation, subcritical, at $r_2 \sim 72.83$. This marks the beginning of the region where the system is multi-stable. The last Hopf bifurcation, born at $r_2 \sim 72.83$, which is subcritical exhibiting the presence of several Torus bifurcations, occurs on a branch of limit cycles when a pair of complex-conjugated Floquet multipliers, leave the unit circle. Branches of stable (resp. unstable) steady states are represented by solid (resp. dashed) black lines; branches of stable (resp. unstable) limit cycles are represented by the max of the oscillation in blue and the minimum in red and by solid (resp. dashed). Hopf bifurcation points are black dots labeled H; Torus bifurcation points are blue dots labeled TR. The bifurcation parameters λ (Scenario I) and r_2 (Scenario II) are represented on the horizontal axis. The max and min values of each variable are represented on the vertical axis.
doi:10.1371/journal.pone.0108676.g003

which is part of the Physics Academic Software. Internally this package uses a Dormand-Prince method of order 5 to integrate differential equations (http://archives.math.utk.edu/software/msdos/diff.equations/ode_workbench/).

The use of differential equations in the study of metabolic processes is widespread nowadays and different biochemical regulation processes have been quantitatively analyzed using time delayed simulations, e.g., in the phosphorylation–dephosphorylation pathways [179], in the endocrine metabolism [180], in the Lactose Operon [181], in the regulation of metabolic pathways [182], in cell signaling pathways [183], and in metabolic networks [184].

Finally, we want to again emphasize that our model only represents some key essential parts of the adenylate energy system. As has previously been indicated, each living cell is essentially a sophisticated metabolic network characterized by hundreds to thousands of biochemical reactions, densely integrated, shaping one of the most complex dynamic systems in nature. The cellular metabolic network functionally integrates all their catalytic processes as a whole. For instance, in a cellular eukaryotic organism the systemic metabolic network includes the enzymatic reactions linked to the plasma membrane, the catabolic and anabolic processes of cytoplasm, the metabolism developed by organelles and subcellular structures, the processes of cell signaling, the adenylate energy system, the metabolism of the

nuclear membrane and the nucleoplasm, the enzymatic processes for genetic expression, etc.

A fundamental property of this cellular metabolic network is their modularity. Metabolism is organized in a modular fashion and the emergence of modules is a genuine characteristic of the functional metabolic organization in all cells [185,186].

Energy is the essential element for the viability of the cellular metabolic network, and practically all bioenergetic processes are coupled with each other via adenosine nucleotides, which are consumed or regenerated by the different enzymatic reactions of the network.

The adenosine nucleotides also act as allosteric control of numerous regulatory enzymes allowing that changes in ATP, ADP and AMP levels can practically regulate the functional activity of the overall metabolic network of cell [10–13].

Accordingly, the cellular energetic system is an integral part of the systemic metabolic network and also shapes a super-complex dynamical system which consists of thousands of biochemical reactions.

In addition, the cellular energy system is involved as well in the set of catabolic and anabolic reactions of the systemic metabolism exhibiting specific processes, e.g., the oxidative phosphorylation, the glycolytic metabolism and other catalytic reactions of substrate-level phosphorylation, the regulatory modular sub-networks of adenosine nucleotide signals, the AMPK system

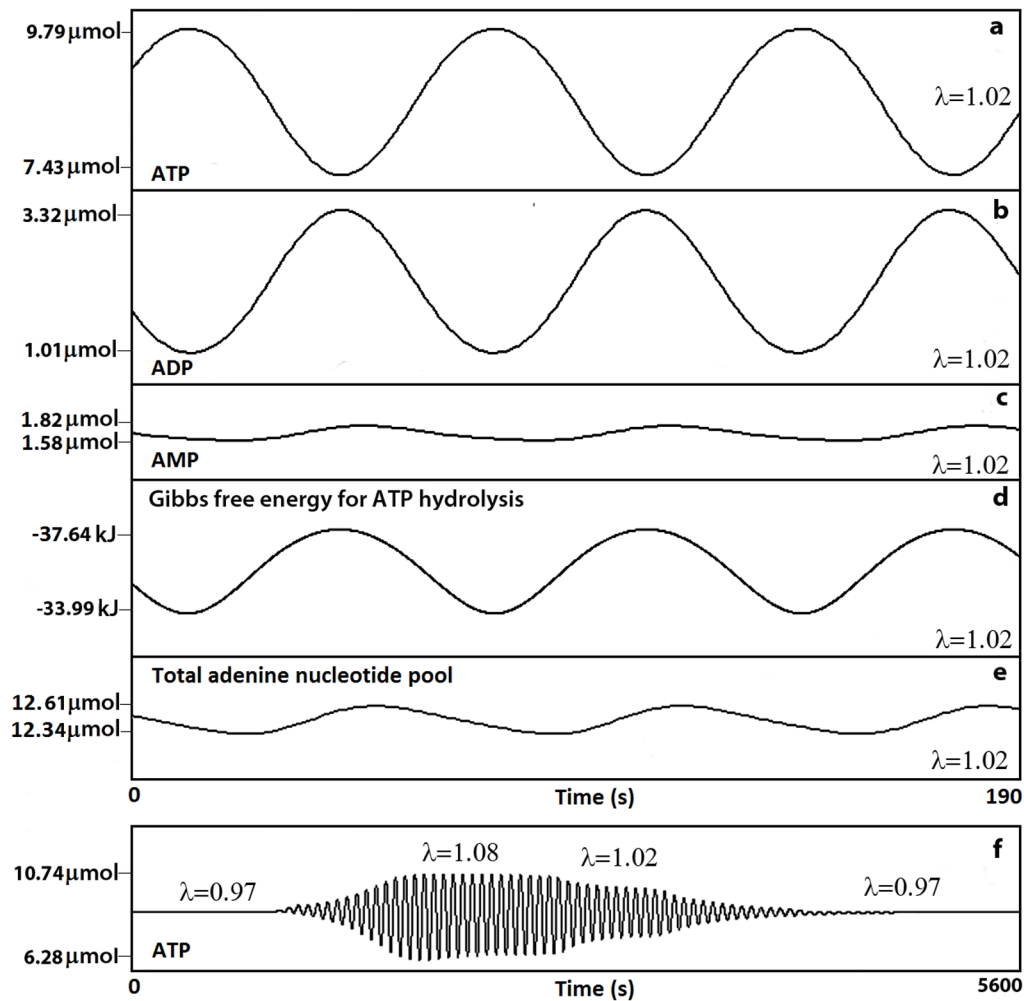


Figure 4. Dynamical solutions of Scenario I. For $\lambda = 1.02$ (normal activity for the ATP synthesis), periodic oscillations emerge. (a) ATP concentrations. (b) ADP concentrations. (c) AMP concentrations. (d) The Gibbs free energy change for ATP hydrolysis to ADP. (e) The total adenine nucleotide (TAN) pool. It can be observed that ATP and ADP oscillate in anti-phase (the ATP maximum concentration corresponds to the ADP minimum concentration). Likewise, it is noted that the total adenine nucleotide pool shows very small amplitude of only $0.27 \mu\text{mol}$ and a period around 65 s. (f) ATP transitions between different periodic oscillations and a steady state pattern for several values of λ (0.97, 1.08, 1.02, 0.97). Maxima and minima values per oscillation are shown in y-axis.
doi:10.1371/journal.pone.0108676.g004

which acts as a metabolic master switch, the degradation processes of the adenosine nucleotides, the allosteric and covalent modulations of enzymes involved in bioenergetic processes, the role of AMP, AMPK and adenylate kinase in nucleotide-based metabolic signaling, the principles of dissipative self-organization of the bioenergetic processes and the significance of metabolic oscillations in the adenosine nucleotide propagation inside the cell.

Results

To understand the dynamics of the main enzymatic reactions interconverting the adenine nucleotides we have analyzed a biochemical model for the adenylate energy system using the system of delay differential equations (12) to account for the asynchronous conditions inside the cell.

Scenario I

Scenario I represents the fundamental analysis of the paper, being λ the main control parameter, which models the energy level stored in the proton gradient generated by the enzymatic

oxidation of input nutrients, and therefore, represents the modifying factor for the ATP synthesis in the system due to substrate intake.

The numerical integration illustrated in Figure 3a-c shows that the temporal structure of the biochemical model is simpler than Scenario II (see below). At small λ values, for $0.9 \leq \lambda < 1$ the adenine nucleotide concentrations display a family of stable steady states (notice that $\lambda = 0.9$ represents a 10% reduction of the ATP synthesis). These steady states lose stability at a Hopf bifurcation detected numerically for $\lambda \sim 1$ which corresponds to a normal activity of ATP synthase with a maximum rate of $7.14 \mu\text{mol s}^{-1}$ [166]. For values of λ bigger than 1 the attractor of the system is a stable limit cycle (therefore, the Hopf bifurcation is supercritical). Concretely, the amplitude of adenine nucleotide oscillations augments as λ increases, e. g., for $\lambda = 1.02$, which represents a 2% of increment in ATP synthesis, the adenine nucleotides exhibit new oscillations with amplitude values of $2.36 \mu\text{mol}$ (ATP), $2.21 \mu\text{mol}$ (ADP) and $0.24 \mu\text{mol}$ (AMP). With an 8% of increase in the ATP synthesis ($\lambda = 1.08$) the amplitudes show higher values, namely $4.47 \mu\text{mol}$ (ATP), $4.41 \mu\text{mol}$ (ADP) and $0.5 \mu\text{mol}$ (AMP).

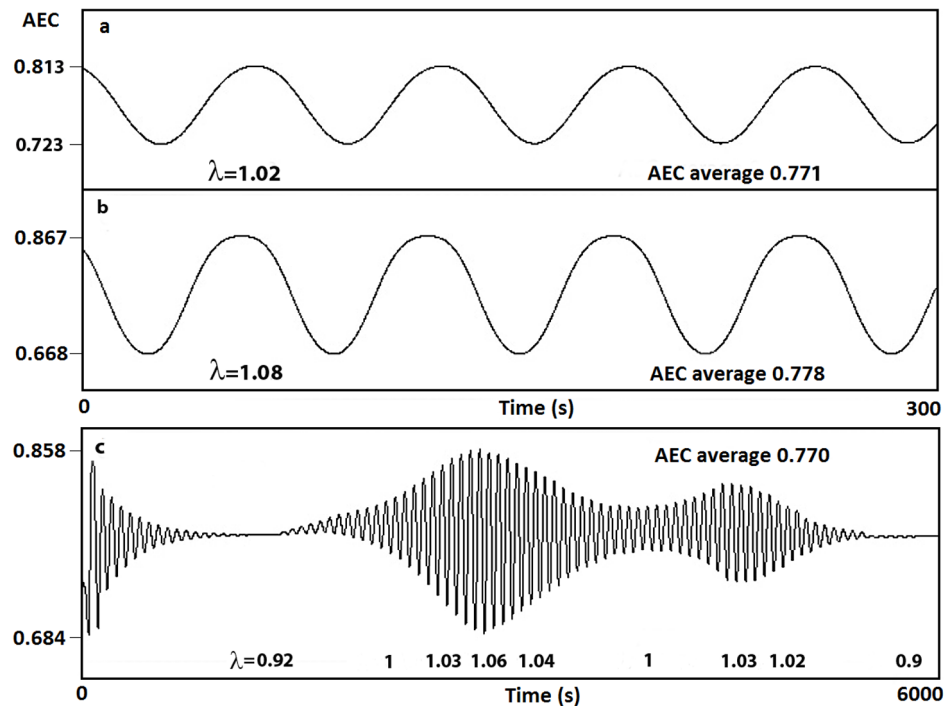


Figure 5. Emergence of oscillations in the AEC (Scenario I). Different oscillatory behavior appears when varying λ , the modifying factor for the ATP synthesis. (a) For $\lambda=1.02$ (normal activity of ATP synthesis) the AEC periodically oscillates with a very low relative amplitude of 0.090. (b) At higher values of ATP synthesis ($\lambda=1.08$) large oscillations emerge with an amplitude of 0.199. (c) AEC transitions between different periodic oscillations and steady state patterns for several values of λ (0.92, 1, 1.03, 1.06, 1.04, 1, 1.03, 1.02, 0.9). doi:10.1371/journal.pone.0108676.g005

Finally, when activity reaches a 10% increase ($\lambda=1.1$) the three dependent variables of the metabolic system oscillate with higher amplitude concentrations: 4.92 μmol , 4.84 μmol and 0.55 μmol , respectively.

Figure 4 shows three time series belonging to ATP, ADP and AMP (panels a, b and c, respectively), for $\lambda=1.02$. The largest oscillation values correspond to ATP (max = 9.79 μmol and min = 7.43 μmol) followed by ADP (max = 3.32 μmol and min = 1.01 μmol) and finally, AMP which oscillates with a low relative amplitude (max = 1.82 μmol and min = 1.58 μmol). We have also observed that ATP oscillates in anti-phase with ADP and consequently the maximum concentration of ATP corresponds to the minimum concentration of ADP.

In most metabolic processes, ATP (Mg-ATP) is the main energy source for biochemical reactions and its hydrolysis to ADP or AMP releases a large amount of energy. To this respect, we have estimated the Gibbs free energy change for ATP hydrolysis (to ADP) under an emergent oscillatory condition of the system, applying the known equation $\Delta G'_{\text{reaction}} = \Delta G^0_{\text{reaction}} + RT \ln(\beta/\alpha)$. The change of the standard Gibbs free energy for this reaction was previously evaluated by Alberty and co-workers [187] obtaining a value of -32 kJmol^{-1} under standard conditions of 298 K, 1 bar pressure, pH 7, 0.25 M ionic strength and the presence of 1 mM Mg^{2+} ions forming the ATP.Mg^{2+} complex, which has different thermodynamic properties than free ATP and, it is closer to physiological conditions.

Under these conditions, Figure 4d shows the values of Gibbs free energy change of ATP hydrolysis for $\lambda=1.02$ which corresponds to a normal activity for ATP synthesis. The resulting values for the oscillatory pattern were more negative than the standard value with a maximum and a minimum of $-37.64 \text{ kJmol}^{-1}$ and $-33.99 \text{ kJmol}^{-1}$, meaning that the hydrolysis

of ATP releases a large amount of free energy that can be captured and spontaneously used to drive other energetically unfavorable reactions in metabolism.

The total of adenine nucleotides is another relevant element in the study of cellular metabolic processes. Different experimental observations have shown that changes in the size levels of the adenine nucleotide pool occur under different physiological conditions [188]. We have estimated the total adenine nucleotide (TAN) pool as $[\text{ATP}] + [\text{ADP}] + [\text{AMP}]$, and Figure 4e shows for $\lambda=1.02$ an emergent oscillatory behavior for TAN with a maximum of 12.61 μmol and a minimum of 12.34 μmol , i.e., a little amplitude of only 0.27 μmol and a period of 65 sec.

Likewise, we have observed that the sum of ATP and ADP concentrations exhibits very small range. So, for $\lambda=1.02$, the amplitude is 93 nmol and for $\lambda=1.1$ it is 202 nmol (data not shown in the Figure).

Figure 4f illustrates ATP transitions between different periodic oscillations and a steady state pattern for several values of λ (0.97, 1.08, 1.02, 0.97).

Next, to analyze the dynamics of the energetic status of the system we have calculated the energy charge level. Figure 5 shows different oscillatory patterns for AEC. For $\lambda=1.02$ the AEC periodically oscillates with a low relative amplitude of 0.09 (max = 0.813 and min = 0.723) (Figure 5a). At higher values of ATP synthesis (an increment of 8%) larger oscillations emerge (max = 0.867 and min = 0.668) (Figure 5b).

Finally, Figure 5c illustrates AEC transitions between different periodic oscillations and steady state patterns for several arbitrary values of λ (0.92, 1, 1.03, 1.06, 1.04, 1, 1.03, 1.02, 0.9) and arbitrary integration times. All the oscillatory patterns for the energy charge maintain the AEC average within narrow physiological values between 0.7 and 0.9.

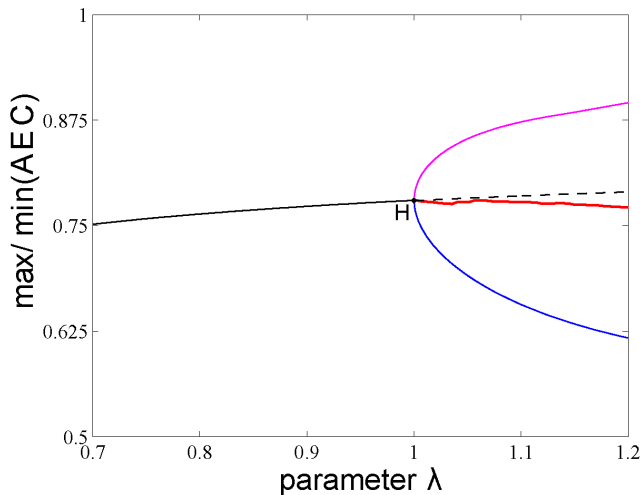


Figure 6. Robustness analysis for the adenylate energy charge (AEC) across different modeling conditions. In y-axis we have plotted the max and the min of the AEC. For situations with no oscillations stable fixed are point colored in solid black lines. In x-axis we have plotted the λ control parameter, which models the energy level stored in the proton gradient generated by the enzymatic oxidation of input nutrients. From left to right, we can see that the system has a fixed point solution which is stable for $\lambda < 1$ (black solid line) and becomes unstable for $\lambda > 1$ (black dashed line), i.e., there is a Hopf bifurcation (H) at $\lambda \sim 1$. For $\lambda > 1$, the limit cycle solution becomes stable, in magenta (blue) we have colored the max (min) of the oscillations. In red, we are coloring the average AEC value of the oscillations. For $\lambda < 1$, the AEC values range from 0.752 and 0.779, and for $\lambda > 1$, the AEC average value between the maximum and minimum per period range from 0.768 to 0.756. At very small λ values, for $\lambda \leq 0.45$ the AEC exhibits values below 0.6 (Figure 7). The AEC does not substantially change during the simulations indicating that it is strongly buffered against the changes of the main control parameter of the system.
doi:10.1371/journal.pone.0108676.g006

Figure 6 shows a robustness analysis of the system in which the values of the adenylate energy charge (AEC) do not substantially change when λ , the main control parameter, is heavily modified (a 50% of its value) indicating that AEC is strongly buffered.

Thus, at small λ values, for $0.7 \leq \lambda \leq 0.99$, the AEC displays a family of stable steady states and the AEC values range from 0.752 to 0.779 (notice that $\lambda = 0.7$ represents a 30% reduction of the ATP synthesis). These steady states lose stability at a Hopf bifurcation for $\lambda \sim 1$ and the AEC exhibits oscillatory behaviors of period 1, being the average between the maximum and minimum of $\overline{\text{AEC}} = 0.769$. Notice that $\lambda = 1$ corresponds to an optimal activity of ATP synthase with a maximum rate of $7.14 \mu\text{mol s}^{-1}$ [166].

As expected, the maximum and minimum per period get bigger as λ increases, and for $\lambda = 1.2$ the AEC maximum per oscillation reaches 0.896 and the $\overline{\text{AEC}}$ decreases to 0.769 ($\lambda = 1.2$ represents a 20% increase in activity of the optimal ATP synthesis).

This robustness analysis of the system for a perturbation of 50% in the λ values show that in the stable steady states the AEC values range from 0.752 to 0.779 and in the stable periodic behaviors the AEC average between the maximum and minimum per period ranges from 0.768 to 0.756.

At very small λ values, for $\lambda \leq 0.45$, the AEC exhibits values below 0.6, which are gradually descending up to reach very small energy values, when the system finally collapsed (Figure 7) [97–101].

During decades, experimental studies have shown that when yeast cells are harvested, starved and then supplemented they exhibit significant metabolic oscillations.

Following these observations, we have compared our results with a classical study for oscillations of the intracellular adenine nucleotides in a population of intact cells belonging to the yeast *Saccharomyces cerevisiae* [30]. These cells were quenched 5 min after adding 3 mM-KCN and 20 mM-glucose at time intervals of 5 s. Figure 8a shows the dynamics of adenine nucleotide concentrations experimentally obtained, exhibiting AEC rhythms between 0.6 and 0.9 values (in the first and second oscillation) and a period of around 50 s. In addition, Richard and colleagues attempted to fit a sinusoidal curve through the experimental points [30]. Figure 8b shows an AEC oscillatory pattern at high values of ATP synthesis ($\lambda = 1.1$), $\text{max} = 0.873$, $\text{min} = 0.656$ and a period of 65 s.

Scenario II

In this second Scenario we have considered r_2 as the control parameter, modeling time delays for ADP.

The numerical bifurcation analysis reveals that the temporal structure of the system (12) is complex, with several Hopf bifurcations emerging as well as secondary bifurcations of Neimark-Sacker type (torus), along two branches of limit cycles (Figure 3 d–f).

Concretely, using the numerical continuation package DDE-Biftools [189], we find 5 Hopf bifurcations. Two pairs of Hopf bifurcations are connected in parameter space, that is, the branch of limit cycles born at one, ends at the other, and the fifth Hopf bifurcation gives a branch that extends up to the upper limit of the interval considered, that is, $r_2 = 120$ s.

Gradually increasing r_2 from 1 s, we find that the branch of stable steady states that exists at $r_2 = 1$ s destabilizes at a first Hopf bifurcation occurring at $r_2 \sim 20.12$ s. This Hopf bifurcation is supercritical, which means that the emanating family of limit cycles is stable; this family remains stable until it disappears through the second (also supercritical) Hopf bifurcation at $r_2 \sim 37.04$ s, which allows the family of steady states to re-stabilize; it remains stable until a third supercritical Hopf bifurcation occurs at $r_2 \sim 71.94$ s, rapidly followed by another Hopf bifurcation, subcritical, at $r_2 \sim 72.83$ s.

This marks the beginning of the region where the system is multi-stable, with one stable steady state and (at least) one stable limit cycle. The last Hopf bifurcation, terminating the branch of limit cycle born at $r_2 \sim 72.83$ s, is subcritical.

The reason for the complex integral solutions in the Scenario II is the presence of several torus bifurcations detected along both branches of limit cycles in the region of r_2 between 71 s and 110 s. We recall that a Torus bifurcation occurs on a branch of limit cycles when a pair of complex-conjugated Floquet multipliers, leave the unit circle (in the complex plane). This corresponds to the fact that this branch of limit cycles becomes unstable and the stable solution starts winding on an invariant torus, periodic or quasi-periodic. We detect four Torus bifurcations, corresponding to the appearance and disappearance of multi-frequency oscillations, at the following values of r_2 : $r_2 \sim 73.22$ s, $r_2 \sim 74.66$ s, $r_2 \sim 85.66$ s, and $r_2 \sim 106.06$ s. Note the following additional details about the Figure 3 d–f we made: branches of stable (resp. unstable) steady states are represented by solid (resp. dashed) black lines; branches of stable (resp. unstable) limit cycles are represented by the max of the oscillation in blue and the minimum in red and by solid (resp. dashed). Hopf bifurcation points are represented with black dots labeled H; Torus bifurcation points with blue dots labeled TR. The horizontal axis corresponds to the

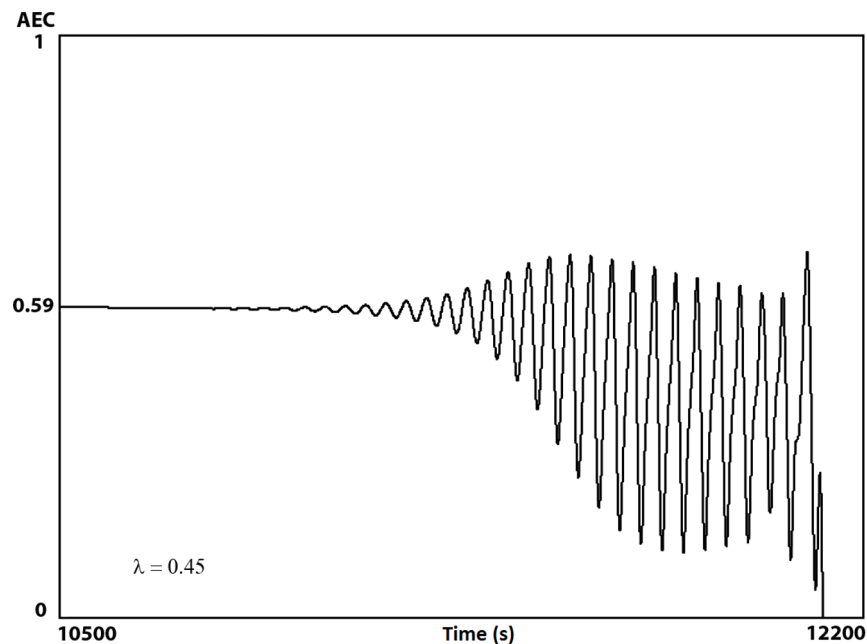


Figure 7. AEC dynamics under low production of ATP. AEC values as a function of time. At very small λ values ($\lambda \approx 0.45$), which represents a strong reduction of the ATP synthesis due to low substrate intake, the dynamic of the adenylate energy system shows a steady state behavior that slowly starts to descend, in a monotone way, up to reach the lowest energy values (AEC ~ 0.59) at which the steady state loses stability and oscillatory patterns emerge with a decreasing trend. Finally, when the maximum of the energy charge oscillations reaches a very small value (AEC ~ 0.28) the adenylate system suddenly collapses after 12,000 seconds of temporal evolution. doi:10.1371/journal.pone.0108676.g007

bifurcation parameters: λ (Scenario I) and r_2 (Scenario II). The vertical axis corresponds to the dependent variable's maxima along various computed branches.

Figure 9 illustrates several examples of oscillatory patterns for the adenylate energy charge under different delay times. For $r_2 = 37$ s the AEC periodically oscillates (Figure 9a). Increasing r_2 up to 72 s (Fig. 9b) and up to 94 s (Fig. 9c) there exist complex AEC oscillatory patterns. Finally, AEC transitions between different oscillatory behavior and steady state patterns are observed for several r_2 values (Figure 9d–e): (d) 50 s, 27 s, 30 s, 32 s, 33 s, 72 s, 52 s, (e) 50 s, 27 s, 30 s, 32 s, 34 s, 36 s, 33 s, 36 s, 38 s, 40 s. These r_2 values and the respective integration times have been arbitrarily taken.

Discussion

Energy is the fundamental element to maintain the turnover of the bio-molecular structures and the functional metabolic viability of all unicellular organisms.

The concentration levels of ATP, ADP and AMP reflect roughly the energetic status of cells, and a determined ratio between them was proposed by Atkinson as the adenylate energy charge (AEC) [96]. Under growth conditions, organisms seem to maintain their AEC within narrow physiological values, despite of extremely large fluctuations in the adenine nucleotide concentrations [96–101]. Intensive experimental studies have shown that the AEC ratio is preserved in a wide variety of organisms, both eukaryotes and prokaryotes (for details see Introduction section).

In order to understand some elements that determine the cellular energy status of cells we have analyzed a biochemical model conformed by some key essential parts of the adenylate energy system using a system of delay differential equations (12) in which the enzymatic rate equations of the main processes and all the corresponding physiological kinetic parameters have been

explicitly considered and tested experimentally *in vitro* by other groups. We have used delay-differential equations to model the asynchronous metabolite supplies to the enzymes (substrates and regulatory molecules).

From the model results, the main conclusions are the following:

I. The adenylate energy system exhibits complex dynamics, with steady states and oscillations including multi-stability and multi-frequency oscillations. The integral solutions are stable, and therefore the adenine nucleotide concentrations (dependent variables of the system) can perform transitions between different kinds of oscillatory behavior and steady state patterns in a stabilized way, which is similar to that in the prevailing conditions inside the cell [15,16].

II. The model is in agreement with previous experimental observations [15,16,30], showing oscillatory solutions for adenine nucleotides under different ATP synthesis conditions, at standard enzymatic concentrations, and for different ADP delay times.

III. In all the numerical results, the order of concentration ratios between the adenine nucleotides is maintained in a way that the highest concentration values correspond to ATP, followed by ADP and AMP which displays the lowest values, in agreement with the experimental data obtained by other authors [30,109].

IV. During the oscillatory patterns, ATP and ADP exhibit anti-phase oscillations (the maxima of ATP correspond with the minima of ADP) also experimentally observed in [30].

V. As a consequence of the rhythmic metabolic behavior, the total adenine nucleotide pool exhibits oscillatory patterns (see experimental examples of this phenomenon in [188,190], as well as the Gibbs free energy change for ATP hydrolysis (see [30]). In agreement with these results, we have found that the oscillation for the Gibbs free energy has a maximum and minimum values per period of $-37.64 \text{ kJmol}^{-1}$ and $-33.99 \text{ kJmol}^{-1}$, the same order

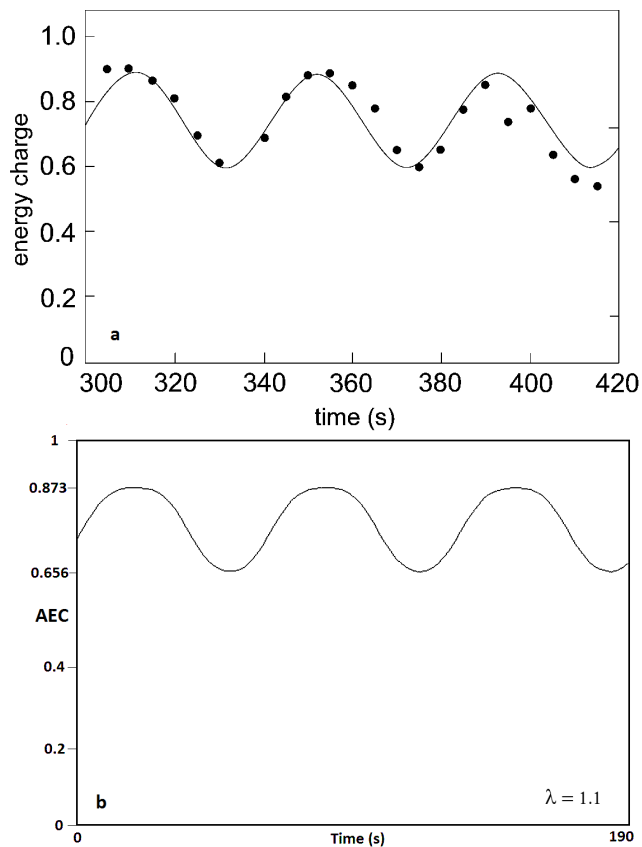


Figure 8. Experimental vs numerical results of AEC oscillations.

Figure 7a illustrates a classical study of the intracellular adenine nucleotides in a population of intact cells belonging to the yeast *Saccharomyces cerevisiae* [30] which exhibits AEC rhythms, with $\max=0.9$, $\min=0.6$ and a period around 50 s. The authors fitted the experimental points to a sinusoidal curve. Figure 7b shows AEC oscillations belonging to our model at high values of ATP synthesis ($\lambda=1.1$), with $\max=0.873$, $\min=0.656$ and a period of 65 s. doi:10.1371/journal.pone.0108676.g008

of magnitude as in experimental observations (about -50 kJmol^{-1} in rat hepatocytes) [191].

VI. The adenylate energy charge shows transitions between oscillatory behaviors and steady state patterns in a stabilized way. We have compared an integral solution of our model with a classical study of intracellular concentrations for adenine nucleotides in a population of intact cells belonging to the yeast *Saccharomyces cerevisiae* and the model fits well with these data [30].

VII. The adenylate energy charge (AEC) does not substantially change during the simulations, indicating that is strongly buffered against the perturbations, in agreement with experimental data [97–101].

We want to remark that we have observed oscillatory patterns in the AEC, in the sum of ATP plus ADP and in the total adenine nucleotide pool but with very low amplitude, what might make difficult the experimental observation with traditional methods.

In fact, it is not clear yet what methodologies are the most appropriate to monitor the values of adenine nucleotides [192]. Although bioluminescence assays and high-performance liquid chromatography are the ones most commonly used for most of the studies [151,192], these procedures are discontinuous and do not allow to observe real-time variations at short temporal periods. Moreover, adenosine nucleoside levels are critically dependent on

sample manipulation and extraction by traditional methods. It has been demonstrated that even short lapses in sample preparation (2 min) can dramatically affect results [193].

It has been assumed for a long time that the temporal evolution of ATP, ADP and AMP concentrations present permanent steady state solutions and that, consequently, cells maintain the AEC as a constant magnitude (*homeostasis*). But this conservation is hard to be fulfilled for open systems.

Recently, the use of nanobiosensors has shown to be able to perform real-time-resolved measurements of intracellular ATP in intact cells; the ATP concentration is indeed oscillating, either showing a rhythmic behavior or more complex dynamics with variations over time, but importantly, the ATP concentration is never constant [15,16].

As a consequence of our analysis we suggest that the appropriate notion to describe the temporal behavior of ATP, ADP and AMP concentrations is *homeorhesis* i.e. the non-linear dynamics of the adenylate energy system shape in the phase space permanent transitions between different kinds of attractors including steady states (in cellular conditions correspond to quasi-steady states) and oscillating attractors, which represent the sets of the asymptotic solutions followed by the adenine nucleotide variables.

Homeorhesis is substantially different to *homeostasis*, which basically implies the ability of the system to maintain the adenine nucleotide concentrations in a constant state.

The concept of *homeostasis* was first suggested by the physiologist Walter Cannon [194] in 1932, but its roots are found back to the French physiologist Claude Bernard who argued that an alleged constancy of the internal medium for any organism results from regulatory processes in biological systems [195,196]. For a long time, the notable idea by Claude Bernard of constancy in the internal medium has paved the route of how cellular processes behaved. However, this constancy seems to be apparent.

In mid-twentieth century, the term *homeorhesis* was suggested to be a substitute of *homeostasis* by the prominent biologist Conrad Waddington [197,198] to describe those systems which return back to a specific dynamics after being perturbed by the external environment, thus opposite to *homeostasis*, in which the system returns back to a fixed state. Later, that concept of *homeorhesis* was mathematically applied in distinct biological studies [199–204].

Rather than a permanent physiological stable state (*homeostasis*), living systems seem to be characterized by changing energy dynamics (*homeorhesis*).

In our numerical study, the temporal dynamics for the concentrations of ATP, ADP and AMP are determined by the adenylate energy system, and these adenosine nucleotide dynamics present complex transitions across time evolution suggesting the existence of homeorhesis.

In addition, we have observed that the values of the AEC do not substantially change during the simulations indicating that is strongly buffered against the perturbations. Recall that the AEC represents a particular functional relationship between the concentrations of adenine nucleotides.

As indicated in the introduction section, intensive experimental measurements under growth cellular conditions have shown that AEC values between 0.7 and 0.95 are invariantly maintained in practically all classes of cells which seems to represent a common key feature to all cellular organisms.

Hence, there appear to be two essential elements in determining the cellular energy level: first, the adenylate energy system originates complex transitions over time in the adenosine nucleotide concentrations so that there is no homeostasis for energy; second, it emerges a permanent relationship among the

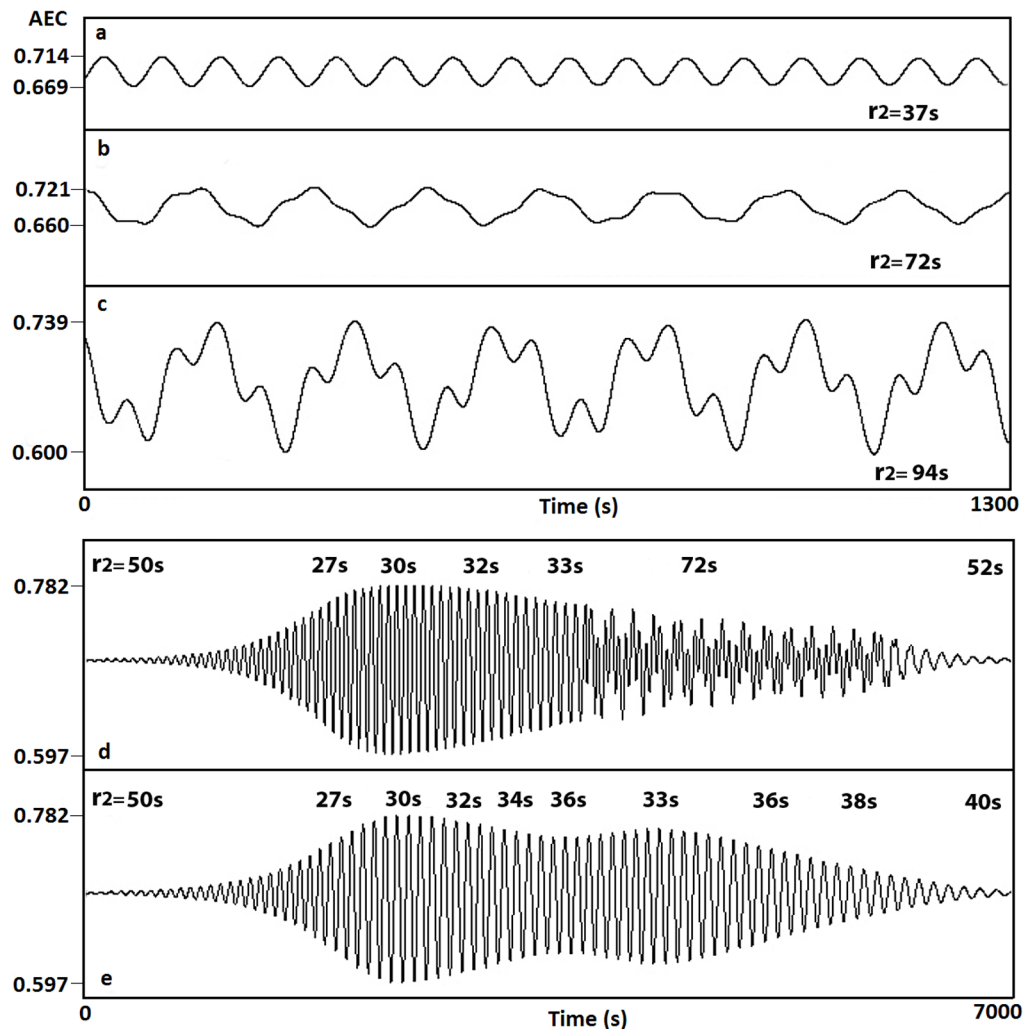


Figure 9. Emergence of oscillations in the AEC (Scenario II). Different oscillatory behavior appears when varying r_2 , controlling the ADP time delays. (a) For $r_2 = 37$ s the AEC periodically oscillates with a very low relative amplitude of 0.045. (b–c) Existence of complex AEC oscillatory patterns for: (b) $r_2 = 72$ s and (c) $r_2 = 94$ s. (d–e) AEC transitions between different oscillatory behavior and steady state patterns for several r_2 values. (d) 50 s, 27 s, 30 s, 32 s, 33 s, 72 s, 52 s. (e) 50 s, 27 s, 30 s, 32 s, 34 s, 36 s, 33 s, 36 s, 38 s, 40 s. doi:10.1371/journal.pone.0108676.g009

dynamics of adenine nucleotide concentrations (AEC values between 0.7 and 0.95), which seems to be strictly fulfilled during all the metabolic transformations that occur during the cell cycle.

These facts make possible to suppose that the cell is an open system where a given magnitude for energy is not conserved but there exists a functional restriction on the possible values that can adopt the adenine nucleotide concentrations.

At least, there seems to be a determinate function relating the adenine nucleotide values which appears to be invariant to all metabolic transformations occurring along the cell cycle. This invariant function, which it would define the real cellular energy state, might possibly have a complex attractor in the phase space since complex dynamic transitions in the adenine nucleotide concentrations have been observed *in vivo* [16], but these hypotheses need deserve further investigation.

Our interpretation to explain the essential elements of the cellular energy charge is that, in addition to the dynamical system which originates the complex transitions in the adenosine nucleotides, there exists an invariant of the energy function which restricts the values that adenylate pool dynamics can take, and the

equation of Atkinson is the manifestation of that invariant function.

The main biological significance of the invariant energy function would be that under growth cellular conditions, the adenylate pool must be highly phosphorylated keeping the rate of adenylate energy production similar to the rate of adenylate energy expenditure.

Cell is a complex non-linearly open system where there is not a specific energy value which is conserved, but rather dynamic forms of change for energy. Unicellular organisms need energy to accomplish the fundamental tasks of the cell metabolism: today, in the post-genomic era, the understanding of the elemental principles and quantitative laws that govern the adenylate energy system is crucial to elucidate some of the fundamental dynamics of cellular life.

Acknowledgments

We acknowledge fruitful conversations and wise advice from Edison Lagares, Ricardo Grande, Josué Tonelli and Andoni Arteagoitia.

Author Contributions

Conceived and designed the experiments: IMDF. Performed the experiments: IMDF JMC EV MD SR IM LM. Analyzed the data: IMDF JMC EV MD SR IM LM. Contributed reagents/materials/analysis tools:

References

- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* 407: 651–654.
- Sear RP (2005) The cytoplasm of living cells: a functional mixture of thousands of components. *J Phys Condens Matter* 17: S3587–S3595.
- Lane N, Martin W (2010) The energetics of genome complexity. *Nature* 467: 928–934.
- Harold FM (1986) *The vital force: A study of bioenergetics*. New York: WH Freeman.
- Xu Z, Spring DR, Yoon J (2011) Fluorescent sensing and discrimination of ATP and ADP based on a unique sandwich assembly of pyrene-adenine-pyrene. *Chem Asian J* 6: 2114–2122.
- Knowles JR (1980) Enzyme-catalyzed phosphoryl transfer reactions. *Annu Rev Biochem* 49: 877–919.
- Nelson DL, Cox MM (2008) *Lehninger Principles of biochemistry*. New York: WH Freeman.
- Hardie DG (2011) Signal transduction: How cells sense energy. *Nature* 472: 176–177.
- Khakh BS (2001) Molecular physiology of P2X receptors and ATP signalling at synapses. *Nat Rev Neurosci* 2: 165–174.
- Cohen PF, Colman RF (1972) Diphosphopyridine nucleotide dependent isocitrate dehydrogenase from pig heart. Characterization of the active substrate and modes of regulation. *Biochemistry* 11: 1501–1508.
- Ercan N, Gannon MC, Nuttall FQ (1996) Allosteric regulation of liver phosphorylase a: revisited under approximated physiological conditions. *Arch Biochem Biophys* 328: 255–264.
- Ercan-Fang N, Gannon MC, Rath VL, Treadway JL, Taylor MR, et al. (2002) Integrated effects of multiple modulators on human liver glycogen phosphorylase. *Am J Physiol Endocrinol Metab* 283: E29–E37.
- Nelson SW, Honzatko RB, Fromm HJ (2002) Hybrid tetramers of porcine liver fructose-1,6-bisphosphatase reveal multiple pathways of allosteric inhibition. *J Biol Chem* 277: 15539–15545.
- Ataullakhanov FI, Vitvitsky VM (2002) What determines the intracellular ATP concentration. *Biosci Rep* 22: 501–511.
- Ozalp VC, Pedersen TR, Nielsen LJ, Olsen LF (2010) Time-resolved measurements of intracellular ATP in the yeast *Saccharomyces cerevisiae* using a new type of nanobiosensor. *J Biol Chem* 285: 37579–37588.
- Ytting CK, Fuglsang AT, Hiltunen JK, Kastaniotis AJ, Özalp VC, et al. (2012) Measurements of intracellular ATP provide new insight into the regulation of glycolysis in the yeast *Saccharomyces cerevisiae*. *Integr Biol (Camb)* 4: 99–107.
- Yoshimoto Y, Sakai T, Kamiya N (1981) ATP oscillation in *Physarum plasmodium*. *Protoplasma* 109: 159–168.
- Akitaya T, Ohsaka S, Ueda T, Kobatake Y (1985) Oscillations in intracellular ATP, cAMP and cGMP concentration in relation to rhythmic sporulation under continuous light in the myxomycete *Physarum polycephalum*. *J Gen Microbiol* 131: 195–200.
- Ainscow EK, Mirshamsi S, Tang T, Ashford ML, Rutter GA (2002) Dynamic imaging of free cytosolic ATP concentration during fuel sensing by rat hypothalamic neurons: evidence for ATP-independent control of ATP-sensitive K(+) channels. *J Physiol* 544: 429–445.
- Kwon HJ (2013) ATP oscillations mediate inductive action of FGF and Shh signalling on prechondrogenic condensation. *Cell Biochem Funct* 31: 75–81.
- Kwon HJ, Ohmiya Y, Honma KI, Honma S, Nagai T, et al. (2012) Synchronized ATP oscillations have a critical role in prechondrogenic condensation during chondrogenesis. *Cell Death Dis* 3: e278.
- Yang JH, Yang L, Qu Z, Weiss JN (2008) Glycolytic oscillations in isolated rabbit ventricular myocytes. *J Biol Chem* 283: 36321–36327.
- Ainscow EK, Rutter GA (2002) Glucose-stimulated oscillations in free cytosolic ATP concentration imaged in single islet beta-cells: evidence for a Ca2+-dependent mechanism. *Diabetes* 51: S162–S170.
- Kennedy RT, Kauri LM, Dahlgren GM, Jung SK (2002) Metabolic oscillations in beta-cells. *Diabetes* 51: S152–S161.
- Dong K, Pelle E, Yarosh DB, Pernodet N (2012) Sirtuin 4 identification in normal human epidermal keratinocytes and its relation to sirtuin 3 and energy metabolism under normal conditions and UVB-induced stress. *Exp Dermatol* 21: 231–233.
- MacDonald MJ, Fahien LA, Buss JD, Hasan NM, Fallon MJ, et al. (2003) Citrate oscillates in liver and pancreatic beta cell mitochondria and in INS-1 insulinoma cells. *J Biol Chem* 278: 51894–51900.
- O'Neill JS, Reddy AB (2011) Circadian clocks in human red blood cells. *Nature* 469: 498–503.
- Gilbert DA, Hammond KD (2008) Phosphorylation dynamics in mammalian cells. In: Lloyd D, Rossi E, editors. *Ultradian rhythms from molecules to mind*. Springer, pp.105–128.
- Steven WE, Lloyd D (1978) Oscillations of respiration and adenine nucleotides in synchronous cultures of *Acanthamoeba castellanii*: mitochondrial respiratory control *in vivo*. *Journal of General Microbiology* 108: 197–204.
- Richard P, Teusink B, Hemker MB, Van Dam K, Westerhoff HV (1996) Sustained oscillations in free-energy state and hexose phosphates in yeast. *Yeast* 12: 731–740.
- Xu Z, Yaquchi S, Tsurugi K (2004) Gts1p stabilizes oscillations in energy metabolism by activating the transcription of TPS1 encoding trehalose-6-phosphate synthase 1 in the yeast *Saccharomyces cerevisiae*. *Biochem J* 383: 171–178.
- Womac AD, Burkeen JF, Neundorff N, Earnest DJ, Zoran MJ (2009) Circadian rhythms of extracellular ATP accumulation in suprachiasmatic nucleus cells and cultured astrocytes. *Eur J Neurosci* 30: 869–876.
- Burkeen JF, Womac AD, Earnest DJ, Zoran MJ (2011) Mitochondrial calcium signaling mediates rhythmic extracellular ATP accumulation in suprachiasmatic nucleus astrocytes. *J Neurosci* 31: 8432–8440.
- Getty-Kaushik L, Richard AM, Corkey BE (2005) Free fatty acid regulation of glucose-dependent intrinsic oscillatory lipolysis in perfused isolated rat adipocytes. *Diabetes* 54: 629–637.
- Rosenspire AJ, Kindzelskii AL, Petty HR (2001) Pulsed DC electric fields couple to natural NAD(P)H oscillations in HT-1080 fibrosarcoma cells. *J Cell Sci* 114: 1515–1520.
- Marquez S, Crespo P, Carlini V, Garbarino-Pico E, Baler R, et al. (2004) The metabolism of phospholipids oscillates rhythmically in cultures of fibroblasts and is regulated by the clock protein PERIOD 1. *FASEB J* 18: 519–521.
- Holz GG, Heart E, Leech CA (2008) Synchronizing Ca2+ and cAMP oscillations in pancreatic beta cells: a role for glucose metabolism and GLP-1 receptors? *Am J Physiol Cell Physiol* 294: c4–c6.
- Rengan R, Omann GM (1999) Regulation of oscillations in filamentous actin content in polymorphonuclear leukocytes stimulated with leukotriene B4 and platelet-activating factor. *Biochem Biophys Res Commun* 262: 479–486.
- Shankaran H, Ippolito DL, Chrisler WB, Resat H, Bollinger N, et al. (2009) Rapid and sustained nuclear-cytoplasmic ERK oscillations induced by epidermal growth factor. *Mol Syst Biol* 5: 332.
- Hans MA, Heinzele E, Wittmann C (2003) Free intracellular amino acid pools during autonomous oscillations in *Saccharomyces cerevisiae*. *Biotechnol Bioeng* 82: 143–151.
- Hartig K, Beck E (2005) Endogenous cytokinin oscillations control cell cycle progression of tobacco BY-2 cells. *Plant Biol* 7: 33–40.
- Hungerbuehler AK, Philippsen P, Gladfelder AS (2007) Limited functional redundancy and oscillation of cyclins in multinucleated *Ashbya gossypii* fungal cells. *Eukaryot Cell* 6: 473–486.
- Shaul O, Mironov V, Bursens S, Van Montagu M, Inze D (1996) Two Arabidopsis cyclin promoters mediate distinctive transcriptional oscillation in synchronized tobacco BY-2 cells. *Proc Natl Acad Sci USA* 93: 4868–4872.
- Chabot JR, Pedraza JM, Luitel P, van Oudenaarden A (2007) Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock. *Nature* 450: 1249–1252.
- Tian B, Nowak DE, Brasier AR (2005) A TNF-induced gene expression program under oscillatory NF-κB control. *BMC Genomics* 6: 137.
- Tonozuka H, Wang J, Mitsui K, Saito T, Hamada Y, et al. (2001) Analysis of the upstream regulatory region of the GTS1 gene required for its oscillatory expression. *J Biochem* 130: 589–595.
- Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genome-wide oscillation in transcription gates DNA replication and cell cycle. *Proc Natl Acad Sci USA* 101: 1200–1205.
- Lange G, Mandelkow EM, Jagla A, Mandelkow E (1988) Tubulin oligomers and microtubule oscillations. Antagonistic role of microtubule stabilizers and destabilizers. *Eur J Biochem* 178: 61–69.
- Placantonakis D, Welsh J (2001) Two distinct oscillatory states determined by the NMDA receptor in rat inferior olive. *J Physiol* 534: 123–140.
- Mellon D Jr, Wheeler CJ (1999) Coherent oscillations in membrane potential synchronize impulse bursts in central olfactory neurons of the crayfish. *J Neurophysiol* 81: 1231–1241.
- García-Muñoz A, Barrio LC, Buño W (1993) Membrane potential oscillations in CA1 hippocampal pyramidal neurons in vitro: intrinsic rhythms and fluctuations entrained by sinusoidal injected current. *Exp Brain Res* 97: 325–333.
- Sánchez-Armás S, Sennoune SR, Maiti D, Ortega F, Martínez-Zaguilán R (2006) Spectral imaging microscopy demonstrates cytoplasmic pH oscillations in glial cells. *Am J Physiol Cell Physiol* 290: C524–C538.
- Lloyd D, Eshantha L, Salgado J, Turner MP, Murray DB (2002) Respiratory oscillations in yeast: clock-driven mitochondrial cycles of energization. *FEBS Lett* 519: 41–44.

54. Danù S, Sørensen PG, Hynne F (1999) Sustained oscillations in living cells. *Nature* 402: 320–322.
55. Ishii K, Hirose K, Iino M (2006) Ca²⁺ shuttling between endoplasmic reticulum and mitochondria underlying Ca²⁺ oscillations. *EMBO Rep* 7: 390–396.
56. Jules M, Francois J, Parrou JL (2005) Autonomous oscillations in *Saccharomyces cerevisiae* during batch cultures on trehalose. *FEBS J* 272: 1490–1500.
57. Getty L, Panteleon AE, Mittelman SD, Dea MK, Bergman RN (2000) Rapid oscillations in omental lipolysis are independent of changing insulin levels *in vivo*. *J Clin Invest* 106: 421–430.
58. Klevecz RR, Murray DB (2001) Genome wide oscillations in expression. Wavelet analysis of time series data from yeast expression arrays uncovers the dynamic architecture of phenotype. *Mol Biol Rep* 28: 73–82.
59. Brodsky VY, Boikov PY, Nechaeva NV, Yurovitsky YG, Novikova TE, et al. (1992) The rhythm of protein synthesis does not depend on oscillations of ATP level. *J Cell Sci* 103: 363–370.
60. Kindzelskii AL, Zhou MJ, Haugland RP, Boxer LA, Petty HR (1998) Oscillatory pericellular proteolysis and oxidant deposition during neutrophil locomotion. *Biophys J* 74: 90–97.
61. Fuentes JM, Pascual MR, Salido G, Soler G, Madrid JA (1994) Oscillations in rat liver cytosolic enzyme activities of the urea cycle. *Arch Int Physiol Biochim Biophys* 102: 237–241.
62. Wittmann C, Hans M, Van Winden WA, Ras C, Heijnen JJ (2005) Dynamics of intracellular metabolites of glycolysis and TCA cycle during cell-cycle-related oscillation in *Saccharomyces cerevisiae*. *Biotechnol Bioeng* 89: 839–847.
63. Aon MA, Roussel MR, Cortassa S, O'Rourke B, Murray DB, et al. (2008) The scale-free dynamics of eukaryotic cells. *PLoS ONE* 3: e3624.
64. Garmendia-Torres C, Goldbeter A, Jacquet M (2007) Nucleocytoplasmic oscillations of the yeast transcription factor Msn2: evidence for periodic PKA activation. *Curr Biol* 17: 1044–1049.
65. Baril EF, Potter VR (1968) Systematic oscillations of amino acid transport in liver from rats adapted to controlled feeding schedules. *J Nutrition* 95: 228–237.
66. Møller AC, Hauser MJ, Olsen LF (1998) Oscillations in peroxidase-catalyzed reactions and their potential function in vivo. *Biophys Chem* 72: 63–72.
67. Chiam KH, Rajagopal G (2007) Oscillations in intracellular signaling cascades. *Phys Rev E* 75: 061901.
68. Smrcinová M, Sørensen PG, Krempaský J, Ballo P (1998) Chaotic oscillations in a chloroplast system under constant illumination. *Int J Bifurcation Chaos* 8: 2467–2470.
69. Murray DB, Beckmann M, Kitano H (2007) Regulation of yeast oscillatory dynamics. *Proc Natl Acad Sci USA* 104: 2241–2246.
70. Allegrini P, Buiatti M, Grigolini P, West BJ (1988) Fractional Brownian motion as a non stationary process: An alternative paradigm for DNA sequences. *Phys Rev E* 57: 4558–4562.
71. Haimovich AD, Byrne B, Ramaswamy R, Welsh WJ (2006) Wavelet analysis of DNA walks. *J Comput Biol* 13: 1289–1298.
72. Ramanujan VK, Biener G, Herman B (2006) Scaling behavior in mitochondrial redox fluctuations. *Biophys J* 90: L70–L72.
73. Kazachenko VN, Astashev ME, Grinevitch AA (2007) Multifractal analysis of K⁺ channel activity. *Biol Membrany* 24: 175–182.
74. De la Fuente IM, Martínez L, Benítez N, Veguillas J, Aguirregabiria JM (1998) Persistent behavior in a phase-shift sequence of periodical biochemical oscillations. *Bull Math Biol* 60: 689–702.
75. De la Fuente IM, Martínez L, Aguirregabiria JM, Veguillas J (1998) R/S analysis in strange attractors. *Fractals* 6: 95–100.
76. Eke A, Herman P, Kocsis L, Kozak LR (2002) Fractal characterization of complexity in temporal physiological signals. *Physiol Meas* 23: R1–R38.
77. De la Fuente IM, Martínez L, Aguirregabiria JM, Veguillas J, Iriarte M (1999) Long-range correlations in the phase-shifts of numerical simulations of biochemical oscillations and in experimental cardiac rhythms. *J Biol Syst* 7: 113–130.
78. Mahasweta D, Gebber GL, Barman SM, Lewis CD (2003) Fractal properties of sympathetic nerve discharge. *J Neurophysiol* 89: 833–840.
79. De la Fuente IM, Pérez-Samartín AL, Martínez L, García MA, Vera-López A (2006) Long-range correlations in rabbit brain neural activity. *Ann Biomed Eng* 34: 295–299.
80. Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Cell Biology: Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science* 310: 1152–1158.
81. Lloyd D, Murray DB (2006) The temporal architecture of eukaryotic growth. *FEBS Lett* 580: 2830–2835.
82. Oliva A, Rosebrock A, Ferrezuelo F, Pyne S, Chen H, et al. (2005) The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *PLoS Biol* 3: e225.
83. Lloyd D, Murray DB (2005) Ultradian metronome: timekeeper for orchestration of cellular coherence. *Trends Biochem Sci* 30: 373–377.
84. De la Fuente IM, Benítez N, Santamaría A, Aguirregabiria JM, Veguillas J (1999) Persistence in metabolic nets. *Bull Math Biol* 61: 573–595.
85. De la Fuente IM, Martínez L, Pérez-Samartín AL, Ormaetxea L, Amezcaga C, et al. (2008) Global self-organization of the cellular metabolic structure. *PLoS ONE* 3: e3100.
86. De la Fuente IM, Vadillo F, Pérez-Pinilla M-B, Vera-López A, Veguillas J (2009) The number of catalytic elements is crucial for the emergence of metabolic cores. *PLoS ONE* 4: e7510.
87. De la Fuente IM, Vadillo F, Pérez-Samartín AL, Pérez-Pinilla M-B, Bidaurrezaga J, et al. (2010) Global self-regulations of the cellular metabolic structure. *PLoS ONE* 5: e9484.
88. De la Fuente IM, Cortes JM, Pérez-Pinilla MB, Ruiz-Rodríguez V, Veguillas J (2011) The metabolic core and catalytic switches are fundamental elements in the self-regulation of the systemic metabolic structure of Cells. *PLoS ONE* 6: e27224.
89. De la Fuente IM, Cortes JM, Pelta DA, Veguillas J (2013) Attractor metabolic networks. *PLoS ONE* 8: e58284.
90. De la Fuente IM (2010) Quantitative analysis of cellular metabolic dissipative, self-organized structures. *Int J Mol Sci* 11: 3540–3599.
91. De la Fuente IM (2014) Metabolic dissipative structures. In: Aon MA, Saks V, Schlattner U, editors. *Systems biology of metabolic and signaling networks: energy, mass and information transfer*. Springer Berlin Heidelberg. pp. 179–212.
92. Edwards JM, Roberts TH, Atwell BJ (2012) Quantifying ATP turnover in anoxic coleoptiles of rice (*Oryza sativa*) demonstrates preferential allocation of energy to protein synthesis. *J Exp Bot* 63: 4389–4402.
93. Boender LGM, Almering MJH, Dijk M, van Maris AJA, de Winder JH, et al. (2011) Extreme calorie restriction and energy source starvation in *Saccharomyces cerevisiae* represent distinct physiological states. *Biochim Biophys Acta* 1813: 2133–2144.
94. Lim EL, Hollingsworth KG, Thelwall PE, Taylor R (2010) Measuring the acute effect of insulin infusion on ATP turnover rate in human skeletal muscle using phosphorus-31 magnetic resonance saturation transfer spectroscopy. *NMR Biomed* 23: 952–957.
95. Hochachka PW, McClelland GB (1997) Cellular metabolic homeostasis during large-scale change in ATP turnover rates in muscles. *J Exp Biol* 200: 381–386.
96. Atkinson DE, Walton GM (1967) Adenosine triphosphate conservation in metabolic regulation. Rat liver citrate cleavage enzyme. *J Biol Chem* 242: 3239–3241.
97. Chapman AG, Fall L, Atkinson DE (1971) Adenylate energy charge in *Escherichia coli* during growth and starvation. *J Bacteriol* 108: 1072–1086.
98. Ball WJ Jr, Atkinson DE (1975) Adenylate energy charge in *Saccharomyces cerevisiae* during starvation. *J Bacteriol* 121: 975–982.
99. Swedes JS, Sedo RJ, Atkinson DE (1975) Relation of growth and protein synthesis to the adenylate energy charge in an adenine-requiring mutant of *Escherichia coli*. *J Biol Chem* 250: 6930–6938.
100. Chapman AG, Atkinson DE (1977) Adenine nucleotide concentrations and turnover rates. Their correlation with biological activity in bacteria and yeast. *Adv Microb Physiol* 15: 253–306.
101. Walker-Simmons M, Atkinson DE (1977) Functional capacities and the adenylate energy charge in *Escherichia coli* under conditions of nutritional stress. *J Bacteriol* 130: 676–683.
102. Privalle LS, Burris RH (1983) Adenine nucleotide levels in and nitrogen fixation by the cyanobacterium *Anabaena* sp. strain 7120. *J Bacteriol* 154: 351–355.
103. Dai R, Liu H, Qu J, Zhao X, Ru J, et al. (2007) Relationship of energy charge and toxin content of *Microcystis aeruginosa* in nitrogen-limited or phosphorus-limited cultures. *Toxicon* 51: 649–658.
104. Beaman KD, Pollack JD (1981) Adenylate energy charge in *Acholeplasma laidlawii*. *J Bacteriol* 146: 1055–1058.
105. Holert J, Hahnke S, Cypionka H (2011) Influence of light and anoxia on chemiosmotic energy conservation in *Dinoroseobacter shibae*. *Environ Microbiol Rep* 3: 136–141.
106. Barrette WC Jr, Hannum DM, Wheeler WD, Hurst JK (1988) Viability and metabolic capability are maintained by *Escherichia coli*, *Pseudomonas aeruginosa*, and *Streptococcus lactis* at very low adenylate energy charge. *J Bacteriol* 170: 3655–3659.
107. Bulthuis BA, Koningsstein GM, Stouthamer AH, van Verseveld HW (1993) The relation of proton motive force, adenylate energy charge and phosphorylation potential to the specific growth rate and efficiency of energy transduction in *Bacillus licheniformis* under aerobic growth conditions. *Anton Leeuw Int J G* 63: 1–16.
108. Kahru A, Liiders M, Vanatalu K, Vilu R (1982) Adenylate energy charge during batch culture of *Thermoactinomyces vulgaris* 42. *Arch Microbiol* 133: 142–144.
109. Weber J, Kayser A, Rinas U (2005) Metabolic flux analysis of *Escherichia coli* in glucose-limited continuous culture. II. Dynamic response to famine and feast, activation of the methylglyoxal pathway and oscillatory behaviour. *Microbiology* 151: 707–716.
110. Smith BA, Dworkin M (1980) Adenylate energy charge during fruiting body formation by *Myxococcus xanthus*. *J Bacteriol* 142: 1007–1009.
111. González F, Fernández-Vivas A, Muñoz J, Arias JM, Montoya E (1989) Adenylate energy charge during the life cycle of *Myxococcus coralloides* D. *FEMS Microbiol Lett* 58: 21–24.
112. Skjoldal HR (1981) ATP concentration and adenylate energy charge of tropical zooplankton from waters inside the great barrier reef. *Mar Biol* 62: 119–123.
113. Hünken M, Karsten U, Wiencke C (2005) Determination of the adenylate energy charge (AEC) as a tool to determine the physiological status of macroalgal tissues after UV exposure. *Phycologia* 44: 249–253.
114. Guimarães PMR, Londesborough J (2008) The adenylate energy charge and specific fermentation rate of brewer's yeasts fermenting high- and very high-gravity worts. *Yeast* 25: 47–58.

115. Chen Y, Xing D, Wang W, Ding Y, Du L (2007) Development of an ion-pair HPLC method for investigation of energy charge changes in cerebral ischemia of mice and hypoxia of Neuro-2a cell line. *Biomed Chromatogr* 21: 628–634.
116. Derr RF, Zieve L (1972) Adenylate energy charge: relation to guanylate energy charge and the adenylate kinase equilibrium constant. *Biochem Biophys Res Commun* 49: 1385–1390.
117. Suska M, Skotnicka E (2010) Changes in adenylate nucleotides concentration and Na⁺, K⁺ - ATPase activities in erythrocytes of horses in function of breed and sex. *Veterinary Medicine International* 2010: ID 987309.
118. Bhatt DP, Chen X, Geiger JD, Rosenberger TA (2012) A sensitive HPLC-based method to quantify adenine nucleotides in primary astrocyte cell cultures. *J Chromatogr B* 889–890: 110–115.
119. Mills DCB, Thomas DP (1969) Blood platelet nucleotides in man and other species. *Nature* 222: 991–992.
120. Biegniewska A, Zietara MS, Rurangwa E, Ollevier F, Swierczynski J, et al. (2010) Some differences between carp (*Cyprinus carpio*) and African catfish (*Clarias gariepinus*) spermatozoa motility. *J Appl Ichthyol* 26: 674–677.
121. Plaideau C, Liu J, Hartleib-Geschwindner J, Bastin-Coyette L, Bontemps F, et al. (2012) Overexpression of AMP-metabolizing enzymes controls adenine nucleotide levels and AMPK activation in HEK293T cells. *FASEB J* 26: 2685–2694.
122. Rajab P, Fox J, Riaz S, Tomlinson D, Ball D, et al. (2000) Skeletal muscle myosin heavy chain isoforms and energy metabolism after clenbuterol treatment in the rat. *Am J Physiol Regul Integr Comp Physiol* 279: R1076–R1081.
123. Zubatkina IS, Emelyanova LV, Savina MV (2008) Adenine nucleotides and Atkinson energetic charge in liver tissue of cyclostomes and amphibians in ontogenesis. *J Evol Biochem Phys* 44: 763–765.
124. Rakotonirainy MS, Arnold S (2008) Development of a new procedure based on the energy charge measurement using ATP bioluminescence assay for the detection of living mould from graphic documents. *Luminescence* 23: 182–186.
125. Dinesh R, Chaudhuri SG, Sheeja TE (2006) ATP levels and adenylate energy charge in soils of mangroves in the Andamans. *Curr Sci India* 90: 1258–1263.
126. Pradet A, Raymond P (1983) Adenine nucleotide ratios and adenylate energy charge in energy metabolism. *Annu Rev Plant Physiol* 34: 199–224.
127. Singh J (1981) Isolation and freezing tolerances of mesophyll cells from cold hardened and nonhardened winter rye. *Plant Physiol* 67: 906–909.
128. Hanhijarvi AM, Fagerstedt KV (1995) Comparison of carbohydrate utilization and energy charge in the yellow flag iris (*Iris pseudacorus*) and garden iris (*Iris germanica*) under anoxia. *Physiol Plantarum* 93: 493–497.
129. McKee KL, Mendelsohn IA (1984) The influence of season on adenine nucleotide concentrations and energy charge in four marsh plant species. *Physiol Plantarum* 62: 1–7.
130. Pomeroy MK, Andrews CJ (1986) Changes in adenine nucleotides and energy charge in isolated winter wheat cells during low temperature stress. *Plant Physiol* 81: 361–366.
131. Sel'kov EE (1968) Self-oscillations in glycolysis. I. A simple kinetic model. *Eur J Biochem* 4: 79–86.
132. Goldbeter A (1974) Modulation of the adenylate energy charge by sustained metabolic oscillations. *FEBS Lett* 43: 327–330.
133. Rapoport TA, Heinrich R, Rapoport SM (1976) The regulatory principles of glycolysis in erythrocytes in vivo and in vitro. A minimal comprehensive model describing steady states, quasi-steady states and time-dependent processes. *Biochem J* 154: 449–469.
134. Sel'kov EE (1975) Stabilization of energy charge, generation of oscillations and multiple steady states in energy metabolism as a result of purely stoichiometric regulation. *Eur J Biochem* 59: 151–157.
135. Reich JG, Sel'kov EE (1974) Mathematical analysis of metabolic networks. *FEBS Lett* 40: Suppl S119–S127.
136. Chen JQ, Cammarata PR, Baines CP, Yager JD (2009) Regulation of mitochondrial respiratory chain biogenesis by estrogens/estrogen receptors and physiological, pathological and pharmacological implications. *Biochim Biophys Acta* 1793: 1540–1570.
137. Weber J (2006) ATP synthase: subunit-subunit interactions in the stator stalk. *Biochim Biophys Acta* 1757: 1162–1170.
138. Steigmüller S, Turina P, Gräber P (2008) The thermodynamic H⁺/ATP ratios of the H⁺-ATP synthases from chloroplasts and *Escherichia coli*. *Proc Natl Acad Sci USA* 105: 3745–3750.
139. Ådén J, Weise CF, Brännström K, Olafsson A, Wolf-Watz M (2013) Structural topology and activation of an initial adenylate kinase-substrate complex. *Biochemistry* 52: 1055–1061.
140. Lange PR, Geserick C, Tschendorf G, Zrenner R (2008) Functions of chloroplastic adenylate kinases in Arabidopsis. *Plant Physiol* 146: 492–504.
141. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298: 1912–1934.
142. Gottlieb A, Frenkel-Morgenstern M, Safo M, Horn D (2011) Common peptides study of aminoacyl-tRNA synthetases. *PLoS ONE* 6: e20361.
143. Bönsdorff T, Gautier M, Farstad W, Ronningen K, Lingaas F, et al. (2004) Mapping of the bovine genes of the de novo AMP synthesis pathway. *Anim Genet* 35: 438–444.
144. Versées W, Steyaert J (2003) Catalysis by nucleoside hydrolases. *Curr Opin Struct Biol* 13: 731–738.
145. Erlinge D (2010) Purinergic and pyriminergic activation of the endothelium in regulation of tissue perfusion. In: Gerasimovskaya EV, Kaczmarek E, editors. *Extracellular ATP and adenosine as regulators of endothelial cell function: Implications for health and disease*. Springer Netherlands, pp. 1–13.
146. Tanaka K, Gilroy S, Jones AM, Stacey G (2010) Extracellular ATP signaling in plants. *Trends Cell Biol* 20: 601–608.
147. Falzoni S, Donvito G, Di Virgilio F (2013) Detecting adenosine triphosphate in the pericellular space. *Interface Focus* 3: 20120101.
148. Forsyth AM, Wan J, Owrusky PD, Abkarian M, Stone HA (2011) Multiscale approach to link red blood cell dynamics, shear viscosity and ATP release. *Proc Natl Acad Sci USA* 108: 10986–10991.
149. Burnstock G (2012) Discovery of purinergic signalling, the initial resistance and current explosion of interest. *Brit J Pharmacol* 167: 238–255.
150. Nath S, Jain S (2000) Kinetic modeling of ATP synthesis by ATP synthase and its mechanistic implications. *Biochem Biophys Res Commun* 272: 629–633.
151. Valero E, Varón R, García-Carmona F (2006) A kinetic study of a ternary cycle between adenine nucleotides. *FEBS J* 273: 3598–3613.
152. Sheng XR, Li X, Pan XM (1999) An iso-random Bi Bi mechanism for adenylate kinase. *J Biol Chem* 274: 22238–22242.
153. Goldbeter A, Lefebvre R (1972) Dissipative structures for an allosteric model. Application to glycolytic oscillations. *Biophys J* 12: 1302–1315.
154. Goldbeter A (1990) *Rythmes et chaos dans les systèmes biochimiques et cellulaires*. Paris: Masson.
155. Curien G, Bastien O, Robert-Genthon M, Cornish-Bowden A, Cárdenas ML, et al. (2009) Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters. *Mol Syst Biol* 5: 271.
156. Ellis RJ (2001) Macromolecular crowding: an important but neglected aspect of intracellular environment. *Curr Opin Struct Biol* 11: 114–119.
157. Nenninger A, Mastroianni G, Mullineaux CW (2010) Size dependence of protein diffusion in the cytoplasm of *Escherichia coli*. *J Bacteriol* 192: 4535–4540.
158. Peiyang Z (2007) Modeling the airway surface liquid regulation in human lungs. The University of North Carolina at Chapel Hill. Ed. ProQuest.
159. Mori Y, Matsumoto K, Ueda T, Kobatake Y (1986) Spatio-temporal organization of intracellular ATP content and oscillation patterns in response to blue light by *Physarum polycephalum*. *Protoplasma* 135: 31–37.
160. Ueda T, Mori Y, Kobatake Y (1987) Patterns in the distribution of intracellular ATP concentration in relation to coordination of amoeboid cell behavior in *Physarum polycephalum*. *Exp Cell Res* 169: 191–201.
161. De la Fuente IM (1999) Diversity of temporal self-organized behaviors in a biochemical system. *BioSystems* 50: 83–97.
162. De la Fuente IM, Martínez L, Veguillas J, Aguirregabiria JM (1996) Quasiperiodicity route to chaos in a biochemical system. *Biophys J* 71: 2375–2379.
163. De la Fuente IM, Martínez L, Veguillas J (1996). Intermittency route to chaos in a biochemical system. *BioSystems* 39: 87–92.
164. De la Fuente IM, Martínez L, Aguirregabiria JM, Veguillas J (1998) Coexistence of multiple periodic and chaotic regimes in biochemical oscillations. *Acta Biotheor* 46: 37–51.
165. De la Fuente IM, Cortes JM (2012) Quantitative analysis of the effective functional structure in yeast glycolysis. *PLoS ONE* 7: e30162.
166. Soga N, Kinoshita K Jr, Yoshida M, Suzuki T (2011) Efficient ATP synthesis by the thermophilic *Bacillus* F₀F₁-ATP synthase. *FEBS J* 278: 2647–2654.
167. Abruscì P, Chiarelli LR, Galizzi A, Fermo E, Bianchi P, et al. (2007) Erythrocyte adenylate kinase deficiency: characterization of recombinant mutant forms and relationship with nonspherocytic hemolytic anemia. *Exp Hematol* 35: 1182–1189.
168. Thuma E, Schirmer RH, Schirmer I (1972) Preparation and characterization of a crystalline human ATP:AMP phosphotransferase. *Biochim Biophys Acta* 268: 81–91.
169. Blangy D, Buc H, Monod J (1968) Kinetics of the allosteric interactions of phosphofructokinase from *Escherichia coli*. *J Mol Biol* 31: 13–35.
170. Hagen J (2006) *Industrial catalysis: A practical approach*. Weinheim, Germany: Wiley-VCH.
171. Jin J, Dong W, Guarino LA (1998) The LEF-4 subunit of Baculovirus RNA polymerase has RNA 5'-triphosphatase and ATPase activities. *J Virol* 72: 10011–10019.
172. Petty HR, Kindzelskii AL (2001) Dissipative metabolic patterns respond during neutrophil transmembrane signaling. *Proc Natl Acad Sci USA* 98: 3145–3149.
173. Burnstock G (1999) Current status of purinergic signalling in the nervous system. *Prog Brain Res* 120: 3–10.
174. Virginio C, MacKenzie A, Rassendren FA, North RA, Surprenant A (1999) Pore dilation of neuronal P2X receptor channels. *Nat Neurosci* 2: 315–321.
175. North RA (2002) The molecular physiology of P2X receptors. *Physiol Rev* 82: 1013–1067.
176. Blackwell KT (2013) Approaches and tools for modeling signaling pathways and calcium dynamics in neurons. *J Neurosci Methods* 220: 131–140.
177. Jacob T, Ascher E, Alapat D, Olevskaia Y, Hingorani A (2005) Activation of P38MAPK signaling cascade in a VSMC injury model: Role of P38MAPK inhibitors in limiting VSMC proliferation. *Eur J Vasc Endovasc Surg* 29: 470–478.
178. dos Passos JB, Vanhalewyn M, Brandao RL, Castro IM, Nicolli JR, et al. (1992) Glucose-induced activation of plasma-membrane H⁺-ATPase in mutants of the yeast *Saccharomyces cerevisiae* affected in cAMP metabolism, cAMP-dependent protein-phosphorylation and the initiation of glycolysis. *Biochim Biophys Acta* 1136: 57–67.

179. Srividhya J, Gopinathan MS, Schnell S (2007) The effects of time delays in a phosphorylation-dephosphorylation pathway. *Biophys Chem* 125: 286–297.
180. Li J, Kuang Y, Mason CC (2006) Modeling the glucose-insulin regulatory system and ultradian insulin secretory oscillations with two explicit time delays. *J Theor Biol* 242: 722–735.
181. Yildirim N, Mackey MC (2003) Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophys J* 84: 2841–2851.
182. Locasale JW (2008) Signal duration and the time scale dependence of signal integration in biochemical pathways. *BMC Syst Biol* 2: 108.
183. Sung MH, Hager GL (2012) Nonlinear dependencies of biochemical reactions for context-specific signaling dynamics. *Sci Rep* 2: 616.
184. Chen BS, Chen PW (2009) On the estimation of robustness and filtering ability of dynamic biochemical networks under process delays, internal parametric perturbations and external disturbances. *Math Biosci* 222: 92–108.
185. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297: 1551–1555.
186. Geryk J, Slanina F (2013) Modules in the metabolic network of *E. coli* with regulatory interactions. *Int J Data Min Bioinform* 8: 188–202.
187. Alberty RA, Goldberg RN (1992) Standard thermodynamic formation properties of adenosine 5′-triphosphate series. *Biochemistry* 31: 10610–10615.
188. Bonzon M, Hug M, Wagner E, Greppin H (1981) Adenine nucleotides and energy charge evolution during the induction of flowering in spinach leaves. *Planta* 152: 189–194.
189. Engelborghs K, Luzyanina T, Samaey G (2000) DDE-BIFTOOL: a Matlab package for bifurcation analysis of delay differential equations. *TW Report* 305.
190. Ching TM, Ching KK (1972) Content of adenosine phosphates and adenylate energy charge in germinating ponderosa pine seeds. *Plant Physiol.* 50: 536–540.
191. Moran LA, Horton RA, Scrimgeour G, Perry M (2011) *Principles of biochemistry* (5th Edition). New Jersey: Prentice Hall.
192. Manfredi G, Yang L, Gajewski CD, Mattiazi M (2002) Measurements of ATP in mammalian cells. *Methods* 26: 317–326.
193. Buckstein MH, He J, Rubin H (2008) Characterization of nucleotide pools as a function of physiological state in *Escherichia coli*. *J Bacteriol* 190: 718–726.
194. Cannon WB (1932) *The wisdom of the body*. New York: WW Norton & Co.
195. Bernard C (1865) *Introduction à l'étude de la médecine expérimentale*. Paris: Flammarion.
196. Bernard C (1957) *An introduction to the study of experimental medicine*. New York: Dover.
197. Waddington CH (1957) *The strategy of the genes. A discussion of some aspects of theoretical biology*. London: George Allen and Unwin Ltd.
198. Waddington CH (1968) Towards a theoretical biology. *Nature* 218: 525–527.
199. Mamontov E. (2007). Modelling homeorhesis with ordinary differential equations. *Math Comput Model* 45: 694–707.
200. Mamontov E, Psiuk-Maksymowicz K, Koptioug A (2006) Stochastic mechanics in the context of the properties of living systems. *Math Comput Model* 44: 595–607.
201. Mamontov E, Koptioug A, Psiuk-Maksymowicz K (2006) The minimal, phase-transition model for the cell-number maintenance by the hyperplasia-extended homeorhesis. *Acta Biotheor* 54: 61–101.
202. Piotrowska MJ, Mamontov E, Peterson A, Koptioug A (2008) A model and simulation for homeorhesis in the motion of a single individual. *Math Comput Model* 48: 1122–1143.
203. Psiuk-Maksymowicz K, Mamontov E (2008) Homeorhesis-based modelling and fast numerical analysis for oncogenic hyperplasia under radiotherapy. *Math Comput Model* 47: 580–596.
204. Mamontov E (2011) In search for theoretical physiology—a mathematical theory of living systems: comment on “Toward a mathematical theory of living systems focusing on developmental biology and evolution: a review and perspectives” by N. Bellomo and B. Carbonaro. *Phys Life Rev* 8: 24–27.

A combinatorial approach to the design of vaccines

Luis Martínez · Martin Milanič · Leire Legarreta ·
Paul Medvedev · Iker Malaina · Ildefonso M. de la Fuente

Received: 23 October 2013 / Revised: 28 April 2014 / Published online: 25 May 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract We present two new problems of combinatorial optimization and discuss their applications to the computational design of vaccines. In the shortest λ -superstring problem, given a family S_1, \dots, S_k of strings over a finite alphabet, a set \mathcal{T} of “target” strings over that alphabet, and an integer λ , the task is to find a string of minimum length containing, for each i , at least λ target strings as substrings of S_i . In the shortest

Luis Martínez and Leire Legarreta were supported by the Spanish Government, grant MTM2011-28229-C02-02, partly with FEDER funds, by the Basque Government, grant IT753-13. Luis Martínez, Leire Legarreta and Ildefonso M. de la Fuente were supported by the University-Society grant US11/13 of the UPV/EHU. Martin Milanič was supported in part by the Slovenian Research Agency (research program P1-0285 and research projects J1-4010, J1-4021, J15433, MU-PROM/2012-023 and N1-0011: GReGAS, supported in part by the European Science Foundation). Technical and human support provided by IZO-SGIker (UPV/EHU, MICINN, GV/EJ, ESF) is gratefully acknowledged.

L. Martínez (✉) · L. Legarreta · I. M. de la Fuente
Department of Mathematics, University of the Basque Country UPV/EHU, 48080 Bilbao, Spain
e-mail: luis.martinez@ehu.es

L. Martínez · L. Legarreta · I. Malaina · I. M. de la Fuente
Biocruces Health Research Institute I.I.S. Biocruces, Barakaldo, Basque Country, Spain

L. Legarreta
e-mail: leire.legarreta@ehu.es

M. Milanič
University of Primorska, UP IAM, Muzejski trg 2, 6000 Koper, Slovenia

M. Milanič
University of Primorska, UP FAMNIT, Glagoljaška 8, 6000 Koper, Slovenia
e-mail: martin.milanic@upr.si

P. Medvedev
Department of Computer Science and Engineering, The Pennsylvania State University,
State College, USA
e-mail: pashadag@cse.psu.edu

λ -cover superstring problem, given a collection X_1, \dots, X_n of finite sets of strings over a finite alphabet and an integer λ , the task is to find a string of minimum length containing, for each i , at least λ elements of X_i as substrings. The two problems are polynomially equivalent, and the shortest λ -cover superstring problem is a common generalization of two well known combinatorial optimization problems, the shortest common superstring problem and the set cover problem. We present two approaches to obtain exact or approximate solutions to the shortest λ -superstring and λ -cover superstring problems: one based on integer programming, and a hill-climbing algorithm. An application is given to the computational design of vaccines and the algorithms are applied to experimental data taken from patients infected by H5N1 and HIV-1.

Keywords Vaccine design · Combinatorial Optimization · Integer programming · Hill-climbing · Shortest common superstring problem · Set cover problem

Mathematics Subject Classification 68Q25 · 68W32 · 90C90 · 90C59 · 90C90 · 92C40 · 92C50 · 92D20

1 Introduction

Cellular organisms are complex metabolic structures shaped by sophisticated biochemical networks with hundreds of thousands of enzymatic reactions (Jeong et al. 2000) in which chaotic patterns (Goldbeter 1997), persistent behaviors (Audit et al. 2004; De la Fuente 1998; Kazachenko et al. 2007) and other dynamic properties emerge (Allegrini et al. 1998; De la Fuente et al. 2009). In particular, important combinatorial problems arise in the analysis of sequences of nucleotides and amino acids in which computational complexity impose limitations in the effectiveness of the algorithms and the techniques that can be used (Jones and Pevzner 2004; Medvedev et al. 2007).

The computational design of vaccines remains an important open problem with immediate applications to human health. In response to the presence of a virus, the

P. Medvedev
Department of Biochemistry and Molecular Biology,
The Pennsylvania State University, State College, USA

P. Medvedev
Genomic Sciences Institute of the Huck, The Pennsylvania State University, State College, USA

I. Malaina
Department of Physiology, University of the Basque Country UPV/EHU, 48080 Bilbao, Spain
e-mail: iamalaina001@ikasle.ehu.es

I. M. de la Fuente
Institute of Parasitology and Biomedicine López-Neyra, CSIC, Granada, Spain
e-mail: mtpmadei@ehu.es

I. M. de la Fuente
Unit of Biophysics (CSIC, UPV/EHU), and Department of Biochemistry and Molecular Biology,
University of the Basque Country, Bilbao, Spain

immune system develops proteins called antibodies which bind to parts of the virus called antigens. The antibody binds to one or more surface amino acid sequences of the antigen, called epitopes. Epitopes can be linear (consisting of consecutive amino acids in the primary structure of the antigen) or conformational (the amino acids are not consecutive in the sequence but are co-located on the folded structure). Once the immune system develops an antibody for a virus, it is “memorized” and used to neutralize any future viruses with the same epitopes. This forms the idea behind vaccines, which are developed to mimic the epitopes of viruses. Unfortunately, viruses can mutate and the sequence of the epitopes can change, avoiding antibody detection.

The problem of designing a vaccine can thus be formulated combinatorially as choosing an amino acid sequence that would contain epitopes that would maximize the efficiency of the antibodies against actual viruses. Some epitopes occur more frequently than others in natural viral populations; therefore, a common approach is to maximize the coverage of the epitopes appearing in the vaccine given a limit on its length, in the sense that the more frequent epitopes are more likely to be included. Different techniques are applied to solve the problem: For instance, [Nickle et al. \(2007\)](#) based their method in the search of the sequence at the center of tree followed by the addition of a set of epitopes (COT⁺), [Fischer et al. \(2006\)](#) used genetic algorithms, [Toussaint et al. \(2008\)](#) used integer linear programming, [Kirovski et al. \(2007\)](#) used a probabilistic least-constraining most-constrained algorithm, [Jojic et al. \(2005\)](#) used a probabilistic model for maximizing coverage of a vaccine construct and [Giles and Ross \(2011\)](#) used a three-round consensus.

However, current optimization problem formulations do not capture several biological constraints. A synthetic peptide (i.e. vaccine) needs to be biologically viable: it has to be cleaved, transported, and presented all in the correct manner. Unfortunately, not enough is understood to be able to predict this viability as a function of the peptide sequence. Delivery methods (i.e. via vector) further impose constraints on the length. Recently, [Kulkarni et al. \(2013\)](#) argued that the notion of coverage that is optimized may not be the correct one and that including certain immunodominant epitopes may actually diminish the development of more protective antibodies. Without extensive in vivo validation, it therefore remains unclear what needs to be optimized.

The main motivation of this paper is to introduce a new criterion in the design of vaccines which complements the criterion of getting high coverage. We establish a combinatorial condition imposing a determined level of balance by each viral sequence, which guarantees that all viral sequences cover at least a minimum number of epitopes. We will show that this restriction has an interesting consequence that the frequencies of the epitopes in the vaccine are high. This leads to the desirable condition that frequent epitopes are more likely to be covered in the vaccine. Additionally, we show that this combinatorial condition guarantees a better distribution of the covered epitopes among the target strings, helping thus to fight the ability of viruses to escape the immunological diversity.

We introduce two new combinatorial optimization problems that integrate this new criterion. In the shortest λ -superstring problem, we are given a family S_1, \dots, S_k of strings over a finite alphabet, a set \mathcal{T} of “target” strings over the same alphabet, an integer λ , and the task is to find a λ -superstring of minimum length, where a λ -superstring is a string containing, for each i , at least λ target strings as substrings

of S_i . In biological terms, the $\{S_i\}$ are the set of known viral amino acid sequences, \mathcal{T} is the set of epitopes, and the λ -superstring is the desired vaccine. The parameter λ specifies a lower bound on the number of different epitopes that the vaccine must cover in each viral sequence. A second formulation is the shortest λ -cover superstring problem, where we are given a collection X_1, \dots, X_n of finite sets of strings over a finite alphabet and an integer λ . The task is to find a λ -cover superstring of minimum length, where a λ -cover superstring is a string containing, for each i , at least λ elements of X_i as substrings. Here, each X_i represent the set of epitopes that are present in a given viral sequence. While we show that the two problems are polynomially equivalent, each formulation lends itself to different types of algorithms for solving it.

In Sect. 2, we give a formal definition of the shortest λ -superstring problem, together with some bounds for its optimal value, and a discussion about coverage. In Sect. 3, we turn our attention to the shortest λ -cover superstring problem. We establish polynomial equivalence of the two problems, as well as hardness results: the two problems are NP-hard, and also hard to approximate. We then present two approaches to obtain exact or approximate solutions to the shortest λ -superstring and λ -cover superstring problems. In Sect. 4 we describe an exact algorithm for the shortest λ -cover superstring problem based on integer linear programming. In Sect. 5, we present a randomized hill-climbing algorithm for the shortest λ -superstring problem—a heuristic local optimization approach that produces approximate solutions to the problem. In Sect. 6, we apply our techniques to a practical setting: obtaining vaccines for different values of λ for a set of 123 patients infected with H5N1 and for two sets of 169 and 166 patients, respectively, infected with HIV-1. In this setting, we model a vaccine as a λ -superstring, with the set \mathcal{T} of target strings being the set of epitopes (see Sect. 2 for more details). We consider two scenarios: a simple one in which we take as the set of epitopes all sequences of amino acids of a given length, for which we use the hill-climbing algorithm, and a more complex and realistic scenario in which we select the set of epitopes from the HIV Molecular Immunology Database, for which we use the integer programming algorithm.

2 The shortest λ -superstring problem

In this paper A will denote a finite alphabet. We denote by A^* the set $A^* = \bigcup_{n=1}^{\infty} A^n \cup \{\epsilon\}$ of all finite strings over A , where ϵ denotes the empty string. The set A^* is a semigroup with the operation $+$ of concatenation, where $(s_1, \dots, s_n) + (t_1, \dots, t_m) = (s_1, \dots, s_n, t_1, \dots, t_m)$. A string $\mathbf{s} = (s_1, \dots, s_n)$ is said to be of length n , and we will denote by $l(\mathbf{s})$ the length of \mathbf{s} . A string $\mathbf{s} = (s_1, \dots, s_m)$ is said to be a *substring* of another string $\mathbf{t} = (t_1, \dots, t_n)$ if there exists an index u in $\{1, \dots, n - m + 1\}$ such that $t_{u+i-1} = s_i$ for every i in $\{1, \dots, m\}$. In other words, \mathbf{s} is a substring of \mathbf{t} if \mathbf{t} can be written as $\mathbf{t} = \mathbf{u} + \mathbf{s} + \mathbf{w}$ for some strings \mathbf{u} and \mathbf{w} over A . The words string and sequence will be used interchangeably.

The following is the central definition of our paper.

Definition 2.1 Let S_1, \dots, S_k be in A^* , let $\mathcal{T} \subset A^*$ be a set of *target strings*, and let $\lambda \in \mathbb{N}$. A λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$ is a string $\mathbf{v} \in A^*$ such that for every

$i \in \{1, \dots, k\}$, at least λ different target strings are common substrings of both S_i and \mathbf{v} .

More formally, denoting by $\mathcal{CS}(\mathbf{s}, \mathbf{t})$ the set of all common substrings of two strings \mathbf{s} and \mathbf{t} , a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$ is a string $\mathbf{v} \in A^*$ such that

$$|\mathcal{CS}(S_i, \mathbf{v}) \cap \mathcal{T}| \geq \lambda \text{ for all } i = 1, \dots, k.$$

Definition 2.2 If $\mathbf{s} = (s_1, \dots, s_n)$, $\mathbf{t} = (t_1, \dots, t_m)$ are in A^* , the degree of overlapping of \mathbf{s} and \mathbf{t} is

$$ov(\mathbf{s}, \mathbf{t}) = \max\{i \in \{0, 1, \dots, \min\{m, n\}\} \mid s_{n-i+j} = t_j \text{ for } j = 1, \dots, i\}.$$

We can define an operation of overlapping sum $+'$ in A^* by

$$(s_1, \dots, s_n) +' (t_1, \dots, t_m) = (s_1, \dots, s_{n-ov(\mathbf{s}, \mathbf{t})}) + (t_1, \dots, t_m).$$

Remark 1 Unlike the case of concatenation, the set A^* with the overlapping sum is not a semigroup, because associativity does not hold. For instance, if $S_1 = S_3 = a$ and $S_2 = S_4 = b$, then $((S_1 +' S_2) +' S_3) +' S_4 = abab$, while $(S_1 +' S_2) +' (S_3 +' S_4) = ab$. Therefore, for k strings S_1, \dots, S_k with $k \geq 3$, we define their overlapping sum inductively as $S_1 +' \dots +' S_k = (S_1 +' \dots +' S_{k-1}) +' S_k$.

Example 1 1. If $S_1 = \dots = S_k$ and $\mathcal{T} = A^\ell$ (for some $\ell \in \mathbb{N}$), then S_1 is a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$, where λ is the number of different substrings of S_1 of length ℓ .
 2. Again, let $\mathcal{T} = A^\ell$ for some $\ell \in \mathbb{N}$. Then, $S = S_1 +' \dots +' S_k$ is a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$, where $\lambda = \min\{\alpha_1, \dots, \alpha_k\}$, and α_i is the number of different substrings of S_i of length ℓ . The length of S is $n = \sum_{i=1}^k l(S_i) - \sum_{i=1}^{k-1} ov(S_i, S_{i+1})$.

The following example is a more interesting one.

Example 2 Let $A = \{0, 1\}$, $\mathcal{T} = A^3$, and

$$S_1 = 0110101111, S_2 = 0010111100, S_3 = 1001001000, S_4 = 1101000000, S_5 = 1000011011.$$

Then:

1. 1010 is a 1-superstring for $(S_1, \dots, S_5, \mathcal{T})$ of length 4.
2. 00101 is a 2-superstring for $(S_1, \dots, S_5, \mathcal{T})$ of length 5.
3. 0001011 is a 3-superstring for $(S_1, \dots, S_5, \mathcal{T})$ of length 7.
4. 110001011 is a 4-superstring for $(S_1, \dots, S_5, \mathcal{T})$ of length 9.

Obviously, for the given S_i 's and \mathcal{T} , it is not possible to find examples with $\lambda \geq 5$, because S_3 has only four different substrings of length 3.

Consider the following combinatorial optimization problem:

SHORTEST λ -SUPERSTRING.

Instance: Strings S_1, \dots, S_k over a finite alphabet A , a finite set $\mathcal{T} \subset A^*$ of target strings, a coverage requirement $\lambda \in \mathbb{N}$.

Task: Find a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$ of minimum length.

A similar problem for $\lambda = 1$ and for a set \mathcal{T} of strings of the same length was considered by Holley et al. (1991). They were interested in finding a smallest set of strings in \mathcal{T} that contain at least one substring from each S_i .

The SHORTEST λ -SUPERSTRING problem is a minimization problem, with the set of feasible solutions given by all λ -superstrings for $(S_1, \dots, S_k, \mathcal{T})$. Clearly, the problem defined with an instance $(S_1, \dots, S_k, \mathcal{T}, \lambda)$ is feasible if and only if at least λ different substrings of each S_i belong to \mathcal{T} . Since this condition can be efficiently tested, we will assume in the rest of the paper that the input instances are always feasible, that is, they are such that they admit a λ -superstring.

Clearly, if $\lambda_1 \leq \lambda_2$ then every λ_2 -superstring for $(S_1, \dots, S_k, \mathcal{T})$ is also a λ_1 -superstring. Consequently, denoting by $\alpha(S_1, \dots, S_k, \mathcal{T}; \lambda)$ the minimum length of a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$, it holds that

$$\alpha(S_1, \dots, S_k, \mathcal{T}; \lambda_1) \leq \alpha(S_1, \dots, S_k, \mathcal{T}; \lambda_2),$$

that is, the optimal solution value to the problem is a non-decreasing function of the coverage requirement λ .

Before continuing with our mathematical treatment of the problem, let us pause for a moment to mention an application of the SHORTEST λ -SUPERSTRING problem to vaccine design. In such applications, the alphabet A is the set of 20 amino acids, the input strings S_1, \dots, S_k represent the relevant protein sequences, and the set \mathcal{T} of target strings is the set of *epitopes*. Every feasible solution to the problem (that is, a λ -superstring) represents a possible *vaccine*, where λ specifies a lower bound on the number of different epitopes that the vaccine must cover in each sequence. An optimal solution \mathbf{v} to the problem represents a shortest vaccine for a given λ .

There is a tradeoff between the optimal solution value and λ . On the one hand, higher value of λ corresponds to a better vaccine since it covers a larger number of epitopes in each sequence. On the other hand, the vaccine can only be effective if it is not too large. (A too large vaccine would develop an autoimmune response.) Hence, in the vaccine design applications, the shortest λ -superstring problem will typically be solved several times, for different values of λ . Among all obtained (optimal or approximate) solutions, the ones achieving better epitope coverage will generally be preferred (typically, this will correspond to larger values of λ). If λ is high enough, then there is a good chance that other (non-tested) sequences will also have a good percentage of epitopes covered by the same vaccine. Ideally, the value of λ should be set to the minimum value required to develop immunogenicity. However, due to the fact that the set of tested sequences is a subset of a bigger population, such a value of λ is not uniquely defined but should be determined experimentally. At the same time, of course, other biological considerations have to be taken into account

when determining the feasibility of a particular candidate vaccine represented by a λ -superstring.

Let us now return to the mathematical treatment of the problem. We have the following trivial upper bound for $\alpha(S_1, \dots, S_k, \mathcal{T}; \lambda)$ (of course, implicitly assuming, as we previously said, that the instance admits a λ -superstring):

Proposition 2.3 $\alpha(S_1, \dots, S_k, \mathcal{T}; \lambda) \leq k\lambda\tau$, where τ denotes the maximum length of a target string.

In the particular case when $\mathcal{T} = A^\ell$ and no target string appears more than once in any S_i we have the following improved upper bound for $\alpha(S_1, \dots, S_k, \mathcal{T}; \lambda)$:

Proposition 2.4 $\alpha(S_1, \dots, S_k, A^\ell; \lambda) \leq k(\ell + \lambda - 1)$.

Definition 2.5 For given strings S_1, \dots, S_k and a target string $\mathbf{t} \in \mathcal{T}$, we define the frequency of \mathbf{t} in $\{S_1, \dots, S_k\}$ to be $f(\mathbf{t}) = |\{i \mid \mathbf{t} \text{ is a substring of } S_i\}|$.

Remark 2 Observe that, in the definition of $f(\mathbf{t})$ we count only the number of strings S_i covering \mathbf{t} , independently of the number of times that \mathbf{t} can be expressed as a substring of a given S_i .

Definition 2.6 If \mathbf{v} is a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$, we define the *coverage* of \mathbf{v} to be

$$c(\mathbf{v}) = \frac{\sum_{\mathbf{t} \in \mathcal{T} : \mathbf{t} \text{ substring of } \mathbf{v}} f(\mathbf{t})}{\sum_{\mathbf{t} \in \mathcal{T} : \mathbf{t} \text{ substring of some } S_i} f(\mathbf{t})}.$$

Remark 3 The notion of coverage which is usually given in the literature is equivalent to the one just introduced. If we consider, for each epitope \mathbf{t} , the relative frequency $rf(\mathbf{t})$ of \mathbf{t} as the quotient $f(\mathbf{t})/k$, then $rf(\mathbf{t})$ measures how well conserved \mathbf{t} is among the strings S_1, \dots, S_k , and the coverage is usually defined as the quotient of the sum of the $rf(\mathbf{t})$ for the epitopes \mathbf{t} in the vaccine over the sum of the $rf(\mathbf{t})$ for the epitopes in the union of the S_i . This quotient is the same as in the previous definition once we cancel out k .

Remark 4 Obviously, $0 \leq c(\mathbf{v}) \leq 1$ holds. Sometimes it is interesting to express $c(\mathbf{v})$ as a percentage, multiplying it by 100. This is done, in particular, in the design of vaccines, where the substrings in \mathcal{T} are the epitopes, and it is usual in the literature to express the coverage as the percentage of epitopes covered by the vaccine.

In the particular case when the set of target strings is A^ℓ and the strings S_1, \dots, S_k are all of the same length, say m , good properties of λ -superstrings with respect to the coverage $c(\mathbf{v})$ can be proved, in the sense that, when λ goes to $m - \ell + 1$, the coverage $c(\mathbf{v})$ goes to 1.

Proposition 2.7 Let $\mathcal{T} = A^\ell$ and $S_1, \dots, S_k \in A^m$ for some positive integers ℓ, m . Then, the coverage of every λ -superstring \mathbf{v} satisfies $c(\mathbf{v}) \geq \frac{\lambda}{m-\ell+1}$.

Proof On the one hand,

$$\sum_{\mathbf{t} \in \mathcal{T} : \mathbf{t} \text{ substring of } \mathbf{v}} f(\mathbf{t}) = \sum_{i=1}^k |\{\mathbf{t} \in \mathcal{T} : \mathbf{t} \text{ common substring of } \mathbf{v} \text{ and } S_i\}| \geq \lambda k.$$

On the other hand,

$$\sum_{\mathbf{t} \in \mathcal{T} : \mathbf{t} \text{ substring of some } S_i} f(\mathbf{t}) = \sum_{i=1}^k |\{\mathbf{t} \in \mathcal{T} : \mathbf{t} \text{ substring of } S_i\}| \leq k(m - \ell + 1).$$

□

Remark 5 The lower bound for $c(\mathbf{v})$ showed in the previous proposition is a very conservative one (although it can be tight for some examples), and making a good choice of a λ -superstring, one can obtain coverages better than the ones predicted by that bound, as one can see, for example, in Table 1 (in Sect. 5).

Remark 6 Similar reasoning as the one above can also be used to obtain lower bounds for $c(\mathbf{v})$ in the case when $\mathcal{T} = A^\ell$ but S_1, \dots, S_k are of different lengths, say m_1, \dots, m_k .

The bound in this case is:

$$c(\mathbf{v}) \geq \frac{\lambda k}{\sum_i (m_i - \ell + 1)}.$$

This bound is, again, very conservative, as it can be seen in Table 3 (in Sect. 6), where we obtain values of $c(\mathbf{v})$ much bigger than the ones given by the bound.

Remark 7 As we have shown in Proposition 2.7 and in the previous remark after that, λ -superstrings have high levels of coverage as λ increases. Observe that, for a fixed value of λ , there may be two substrings of the same length and the same level of coverage, one of them being a λ -superstring and the other not. Let us give an instance of this situation. In Example 2 we gave λ -superstrings for λ from 1 to 4 for $A = \{0, 1\}$, $\mathcal{T} = A^3$ and 5 strings S_1, \dots, S_5 . In particular, we presented the 2-superstring $s = 00101$. The distribution of the number of target strings which are substrings of both of s and S_i for $i = 1, \dots, 5$ is $(2, 3, 2, 2, 2)$, and the coverage of s is $11/27$. (As a matter of fact, $11/27$ is the maximum coverage attainable by a string of length 5 for this choice of A, \mathcal{T} and S_1, \dots, S_5 .) The string $s' = 01001$ has also length 5 and coverage $11/27$, but it is not a 2-superstring; in fact, the distribution of the number of target strings is $(1, 3, 3, 2, 2)$ in this case, which is not as balanced as the previous one. This is precisely the advantage of obtaining λ -superstrings for bigger λ : a more balanced distribution of target strings.

3 Computational complexity aspects

In this section we will derive hardness results for the SHORTEST λ - SUPERSTRING and SHORTEST λ - COVER SUPERSTRING problems. The two problems are computationally difficult: not only are they NP-hard, they are also hard to approximate. We will first show that the two problems are polynomially equivalent. Next, we will show that the SHORTEST λ - COVER SUPERSTRING problem generalizes two well known combinatorial optimization problems: the shortest common superstring problem and the set cover problem. For background on computational complexity, see, e.g., Ausiello et al. (1999), and Garey and Johnson (1979).

We now formally introduce the SHORTEST λ - COVER SUPERSTRING problem, using the following extension of the notion of a λ -superstring.

Definition 3.1 Let $X_1, \dots, X_n \subseteq A^*$ be a collection of finite sets of strings over a finite alphabet A , and let $\lambda \in \mathbb{N}$. A λ -cover superstring for (X_1, \dots, X_n) is a string $\mathbf{v} \in A^*$ such that for every i , at least λ elements of X_i are substrings of \mathbf{v} .

SHORTEST λ - COVER SUPERSTRING.	
Instance:	A collection $X_1, \dots, X_n \subseteq A^*$ of finite sets of strings over a finite alphabet A , a coverage requirement $\lambda \in \mathbb{N}$.
Task:	Find a λ -cover superstring for (X_1, \dots, X_n) of minimum length.

More formally, the requirement of the SHORTEST λ - COVER SUPERSTRING problem is to find $\mathbf{v} \in A^*$ minimizing $l(\mathbf{v})$ such that for all $i \in \{1, \dots, n\}$, it holds that

$$|\{s \in X_i : s \text{ is a substring of } \mathbf{v}\}| \geq \lambda.$$

Before we prove the equivalence of the SHORTEST λ - SUPERSTRING and the SHORTEST λ - COVER SUPERSTRING problems, let us define formally what kind of equivalence we have in mind. Given an optimization problem Π and an instance I , let us denote by $\mathcal{F}_\Pi(I)$ the set of feasible solutions of Π given I .

Definition 3.2 Given two minimization problems Π_1 and Π_2 , we say that Π_1 is *polynomially reducible* to Π_2 if every instance I_1 to Π_1 can be mapped in polynomial time to an instance I_2 to Π_2 such that the following two conditions hold:

1. $\mathcal{F}_{\Pi_1}(I_1) = \mathcal{F}_{\Pi_2}(I_2)$,
2. $f_1(x) = f_2(x)$ for all $x \in \mathcal{F}_{\Pi_1}(I_1)$, where f_i is the objective function of Π_i , for $i = 1, 2$.

Moreover, two minimization problems Π_1 and Π_2 are *polynomially equivalent* if each of them is polynomially reducible to the other one.

We now prove that the SHORTEST λ - SUPERSTRING and the SHORTEST λ - COVER SUPERSTRING problems are polynomially equivalent, in this strong sense defined above. We split the proof into two propositions, each proving polynomial reducibility in one direction.

Proposition 3.3 *The SHORTEST λ - SUPERSTRING problem is polynomially reducible to the SHORTEST λ - COVER SUPERSTRING problem.*

Proof Let $I = (A, S_1, \dots, S_k, \mathcal{T}, \lambda)$ be an instance to the SHORTEST λ - SUPERSTRING problem. We describe a polynomial time transformation of I to an equivalent instance $I' = (A', X_1, \dots, X_n, \lambda')$ of the SHORTEST λ - COVER SUPERSTRING problem:

- Set $n = k$, $\lambda' = \lambda$ and $A' = A$.
- For each $i \in \{1, \dots, n\}$, define X_i as the set of all target strings $\mathbf{t} \in \mathcal{T}$ that are substrings of S_i .

Clearly, I' can be computed from I in polynomial time. Now let us argue that the sets of feasible solutions of both problems (given the corresponding input instances) are the same. First, suppose that $\mathbf{v} \in A^*$ is a feasible solution to the SHORTEST λ - SUPERSTRING problem given I . Then, for each $i \in \{1, \dots, n\}$, at least λ different target strings are common substrings of both S_i and \mathbf{v} . Hence, for each i , there exists a subset $T_i \subseteq \mathcal{T}$ of cardinality at least λ such that every member of T_i is a substring of both S_i and \mathbf{v} . In particular, $T_i \subseteq X_i$ and every member of T_i is a substring of \mathbf{v} . Hence, \mathbf{v} is a feasible solution to the SHORTEST λ - COVER SUPERSTRING problem given I' .

Conversely, suppose that $\mathbf{v} \in A^*$ is a feasible solution to the SHORTEST λ - COVER SUPERSTRING problem given I' . Then, for each $i \in \{1, \dots, n\}$, there exists a subset $T_i \subseteq X_i$ of cardinality at least λ all the members of which are substrings of \mathbf{v} . Every member of T_i is, by the definition of X_i , a member of \mathcal{T} and a substring of S_i . Hence, at least λ different target strings are common substrings of both S_i and \mathbf{v} , and \mathbf{v} is a feasible solution to the SHORTEST λ - SUPERSTRING problem given I . Since condition (2) from the definition of polynomial reducibility follows directly from the definitions of the two problems, the proof is complete. \square

Proposition 3.4 *The SHORTEST λ - COVER SUPERSTRING problem is polynomially reducible to the SHORTEST λ - SUPERSTRING problem.*

Proof Let $I = (A, X_1, \dots, X_n, \lambda)$ be an instance to the SHORTEST λ - COVER SUPERSTRING problem. We describe a polynomial time transformation of I to an equivalent instance $I' = (A', S_1, \dots, S_k, \mathcal{T}, \lambda')$ of the SHORTEST λ - SUPERSTRING problem:

- Set $k = n$, $A' = A \cup \{*\}$ where $* \notin A$, and $\lambda' = \lambda$.
- For each $i \in \{1, \dots, n\}$, let $X_i = \{x_1^i, \dots, x_{n_i}^i\}$. Construct a string S_i as the concatenation of all strings in X_i separated by $*$:

$$S_i = x_1^i + * + x_2^i + * + \dots + * + x_{n_i}^i.$$

- Set $\mathcal{T} = \cup_{i=1}^n X_i$.

Clearly, I' can be computed from I in polynomial time. Now let us argue that the sets of feasible solutions of both problems (given the corresponding input instances) are the same. Suppose that $\mathbf{v} \in A^*$ is a feasible solution to the SHORTEST λ - COVER SUPERSTRING problem given I . Consider an arbitrary index $i \in \{1, \dots, k\}$. By the assumption on \mathbf{v} , there exists a subset $T_i \subseteq X_i$ of cardinality at least λ all the members of which are substrings of \mathbf{v} . Let $\mathbf{t} \in T_i$. Since $T_i \subseteq X_i$, we have $\mathbf{t} = x_j^i$ for some $j \in \{1, \dots, n_i\}$. Consequently \mathbf{t} is a substring of S_i . Moreover, by construction of \mathcal{T} , we also have $\mathbf{t} \in \mathcal{T}$. In particular, T_i is a set of λ strings from \mathcal{T} all of which are common

substrings of S_i and \mathbf{v} . Since $A \subseteq A'$, we have $\mathbf{v} \in (A')^*$, and hence we conclude that \mathbf{v} is a feasible solution to the SHORTEST λ - SUPERSTRING problem given I' .

Conversely, suppose that $\mathbf{v} \in (A')^*$ is a feasible solution to the SHORTEST λ - SUPERSTRING problem given I' . Notice that symbol $*$ does not appear in any string from \mathcal{T} . Consider an arbitrary index $i \in \{1, \dots, n\}$. Then, there exists a set $T_i \subseteq \mathcal{T}$ consisting of at least λ common substrings of both S_i and \mathbf{v} . By the above observation, no member of T_i contains symbol $*$, and hence $T_i \subseteq A^*$. In particular, due to the structure of S_i , for every string $\mathbf{t} \in T_i$ there exists an index $j \in \{1, \dots, n_i\}$ such that $\mathbf{t} = x_j^i \in X_i$. Thus, T_i is a subset of X_i of cardinality at least λ all the members of which are substrings of \mathbf{v} , which means that \mathbf{v} is a feasible solution to the SHORTEST λ - COVER SUPERSTRING problem given I .

Again, condition (2) from the definition of polynomial reducibility follows directly from the definitions of the two problems, and the proof is complete. \square

Propositions 3.3 and 3.4 imply the following.

Theorem 3.5 *The SHORTEST λ SUPERSTRING and SHORTEST λ - COVER SUPERSTRING problems are polynomially equivalent.*

Hence, every hardness result for the SHORTEST λ - COVER SUPERSTRING problem will immediately imply the analogous hardness result for the SHORTEST λ SUPERSTRING problem.

We now relate the SHORTEST λ - COVER SUPERSTRING to the well known SHORTEST COMMON SUPERSTRING problem:

SHORTEST COMMON SUPERSTRING (SCS).

Instance: A finite set $S \subseteq A^*$ of strings over an alphabet A .

Task: Find a shortest string $\mathbf{t} \in A^*$ that contains each of the input strings $s \in S$ as a substring.

The SCS problem is NP-hard (Gallant et al. 1980; Garey and Johnson 1979), and also APX-hard (Blum et al. 1994), which implies that a polynomial-time approximation scheme for this problem is unlikely. We now show that these hardness results for the SCS problem carry over to the SHORTEST λ - COVER SUPERSTRING problem. Even though we will strengthen this result in Theorem 3.9, we keep the short proof of Proposition 3.6, as it shows that the SCS problem is a special case of the SHORTEST λ - COVER SUPERSTRING problem.

Proposition 3.6 *The SHORTEST λ - COVER SUPERSTRING problem is NP-hard, and also APX-hard.*

Proof Given an instance $I = (A, S)$ to the SCS problem, consider the instance $I' = (A, X_1, \lambda)$ where $X_1 = S$, and $\lambda = |S|$ to the shortest λ -cover superstring problem. Then, a string $\mathbf{v} \in A^*$ is a feasible solution to the shortest λ -cover superstring problem given I' if and only if it is a common superstring of all strings in S .

The result follows. \square

Corollary 3.7 *The SHORTEST λ - SUPERSTRING problem is NP-hard, and also APX-hard.*

Proposition 3.6 and its corollary imply that the SHORTEST λ - SUPERSTRING problem does not admit a PTAS (polynomial time approximation scheme) unless $P = NP$. We now strengthen Proposition 3.6, by showing that for some absolute constant $c > 0$, there is no polynomial time algorithm approximating the SHORTEST λ - SUPERSTRING and SHORTEST λ - COVER SUPERSTRING problems within a factor of $c \ln n$, unless $P = NP$. To do this, we make a reduction from the SET COVER problem (Garey and Johnson 1979).

SET COVER.
Instance: A set-system $\mathcal{C} = (U, \mathcal{F})$, where U is a finite ground set and \mathcal{F} is a collection of subsets of U .
Task: Find a minimum size subcollection $\mathcal{F}' \subseteq \mathcal{F}$ such that every element $u \in U$ appears in some set in \mathcal{F}' .

The decision version of the SET COVER problem is NP-complete (Garey and Johnson 1979). Moreover, Alon et al. obtained the following inapproximability result:

Theorem 3.8 (Alon et al. 2006) *There exists a constant $c > 0.2267$ such that there is no polynomial time algorithm approximating the set cover problem within a factor of $c \ln |U|$, unless $P = NP$.*

Using this result, we now derive an analogous result for the SHORTEST λ - COVER SUPERSTRING problem, even for the case of the binary alphabet $A = \{0, 1\}$ and $\lambda = 1$.

Theorem 3.9 *There exists a constant $c > 0.2267$ such that there is no polynomial time algorithm approximating the SHORTEST λ - COVER SUPERSTRING problem within a factor of $c \ln n$ unless $P = NP$, even for the case of the binary alphabet $A = \{0, 1\}$ and $\lambda = 1$.*

Proof Let c be the constant from Theorem 3.8. Suppose that there exists a polynomial time algorithm \mathcal{A} approximating the shortest λ -cover superstring problem over the binary alphabet and $\lambda = 1$ within a factor of $c \ln n$.

We will construct a polynomial time algorithm \mathcal{A}' approximating the set cover problem within a factor of $c \ln |U|$. The conclusion will then follow from Theorem 3.8.

Let $\mathcal{C} = (U, \mathcal{F})$ be an instance to the set cover problem with $U = \{u_1, \dots, u_n\}$ and $\mathcal{F} = \{F_1, \dots, F_k\}$. To every set $F_j \in \mathcal{F}$ we associate a binary string S_j , as follows:

$$S_j = 0^j + 1 + 0^{k-j} + 1^{k-j} + 0 + 1^j \text{ for all } j = 1, \dots, k,$$

where a^j for $a \in \{0, 1\}$ and $j \in \mathbb{N}$ denotes the string s of length j with $s_i = a$ for all $i = 1, \dots, j$. Notice that each of the strings S_j is of length $2k + 2$, and for every two strings S_i and S_j with $i \neq j$, we have $ov(S_i, S_j) = 0$.

We set $I = (A, X_1, \dots, X_n, \lambda)$ where $A = \{0, 1\}$, $\lambda = 1$ and

$$X_i = \{S_j \mid u_i \in F_j\}$$

for each $i \in \{1, \dots, n\}$.

The algorithm \mathcal{A}' proceeds in three steps:

1. Compute $I = (\{0, 1\}, X_1, \dots, X_n, 1)$ as specified above.
2. Run the approximation algorithm \mathcal{A} for the shortest λ -cover superstring problem on instance I . Let \mathbf{v} denote the obtained λ -cover superstring for $(X_1, \dots, X_n, 1)$.
3. For each $i = 1, \dots, n$, find a substring S_j of \mathbf{v} such that $S_j \in X_i$, and let S denote the set of all these strings. Output $\mathcal{F}' = \{F_j \mid S_j \in S\}$.

It is clear that \mathcal{A}' runs in polynomial time. Moreover, every element $u_i \in U$ appears in some set in \mathcal{F}' . (Indeed, if $u_i \in U$ then there exists some $S_j \in S$ such that $S_j \in X_i$. Hence $u_i \in F_j \in \mathcal{F}'$.) Thus, \mathcal{F}' is a feasible solution to the set cover problem given \mathcal{C} .

Let us first observe that, since no two strings S_j have a nontrivial overlap, we have

$$l(\mathbf{v}) \geq \sum_{s \in S} l(s) = (2k + 2)|S| = (2k + 2)|\mathcal{F}'|. \tag{1}$$

Next, if \mathbf{v}^{opt} is an optimal solution to the shortest λ -cover superstring problem given I , then the assumption on \mathcal{A} implies that

$$l(\mathbf{v}) \leq c \ln n \cdot l(\mathbf{v}^{\text{opt}}). \tag{2}$$

Consider an optimal solution $\mathcal{F}^{\text{opt}} = \{F_{i_1}, \dots, F_{i_p}\}$ to the set cover problem given \mathcal{C} . Let \mathbf{v}^* be the string defined by concatenating all the strings corresponding to sets in \mathcal{F}^{opt} , that is, $\mathbf{v}^* = S_{i_1} + S_{i_2} + \dots + S_{i_p}$. We have $l(\mathbf{v}^*) = (2k + 2)|\mathcal{F}^{\text{opt}}|$, and consequently

$$l(\mathbf{v}^{\text{opt}}) \leq (2k + 2)|\mathcal{F}^{\text{opt}}|. \tag{3}$$

Finally, putting all these observations together, we can bound the size of \mathcal{F}' from above as follows:

$$\begin{aligned} |\mathcal{F}'| &\leq \frac{1}{(2k + 2)} \cdot l(\mathbf{v}) \quad (\text{by (1)}) \\ &\leq \frac{1}{(2k + 2)} \cdot c \ln n \cdot l(\mathbf{v}^{\text{opt}}) \quad (\text{by (2)}) \\ &\leq \frac{1}{(2k + 2)} \cdot c \ln n \cdot (2k + 2)|\mathcal{F}^{\text{opt}}| \quad (\text{by (3)}) \\ &= c \ln n \cdot |\mathcal{F}^{\text{opt}}| \quad (\text{by (3)}). \end{aligned}$$

Hence, algorithm \mathcal{A}' approximates the set cover problem within a factor of $c \ln |U|$. By Theorem 3.8, this is only possible if $\mathbf{P} = \mathbf{NP}$. This completes the proof. \square

Corollary 3.10 *There exists a constant $c > 0.2267$ such that there is no polynomial time algorithm approximating the SHORTEST λ -SUPERSTRING problem within a factor of $c \ln k$, unless $\mathbf{P} = \mathbf{NP}$.*

4 An integer programming approach

In this section, we describe how to solve the SHORTEST λ -COVER SUPERSTRING problem using integer programming (IP). (For background on integer programming, see, e.g., Schrijver 1986.) Our approach is to model the problem as a generalization of the *generalized Traveling Salesman Problem* introduced in Henry-Labordere (1969), Saksena (1970), and Srivastava et al. (1969), in which the set of vertices of a given complete directed edge-weighted graph is divided into clusters and the objective is to find a minimum-cost tour passing through one node from each cluster.

Let $(A, X_1, \dots, X_n, \lambda)$ be an instance of the SHORTEST λ -COVER SUPERSTRING problem. We construct a complete directed edge-weighted graph $D = (V, E, w)$, called the *distance graph*, as follows:

- The vertex set V is the set of all input strings, together with a new vertex s^* :

$$V = \cup_{i=1}^n X_i \cup \{s^*\},$$

- For every two distinct vertices $s, t \in V \setminus \{s^*\}$, add the arc (s, t) to E and assign to it the weight $w_{s,t} = l(s) - ov(s, t)$. (This quantity will also be denoted by $dist(s, t)$.) Clearly, the weights are well defined and non-negative.
- For every vertex $s \in V \setminus \{s^*\}$, add the arc (s, s^*) to E and assign to it weight $w_{s,s^*} = l(s)$.
- For every vertex $s \in V \setminus \{s^*\}$, add the arc (s^*, s) to E and assign to it weight $w_{s^*,s} = 0$.

As the following proposition shows, the SHORTEST λ -COVER SUPERSTRING problem is equivalent to that of finding in G a directed cycle C through s^* of minimum total length subject to the constraint that for every set X_i , at least λ strings from X_i appear as vertices of C .

Proposition 4.1 *Suppose that \mathbf{v} is an optimal solution to the SHORTEST λ -COVER SUPERSTRING problem on the instance $(A, X_1, \dots, X_n, \lambda)$, and let S be the set of strings from $\cup_{i=1}^n X_i$ that are substrings of \mathbf{v} . Let (s_1, \dots, s_k) be the order of the strings from S as they appear in \mathbf{v} for the first time. Then, $C := (s^*, s_1, \dots, s_k)$ is a directed cycle in G of total length at most $l(\mathbf{v})$ and such that for every set X_i , at least λ strings from X_i appear as vertices of C .*

Conversely, suppose that $C = (s^, s_1, \dots, s_k)$ is a directed cycle through s^* in G such that for every set X_i , at least λ strings from X_i appear as vertices of C . Then, $\mathbf{v} = s_1 +' \dots +' s_k$ is a feasible solution to the SHORTEST λ -COVER SUPERSTRING problem such that $l(\mathbf{v}) = w(C)$.*

Proof Let $C := (s^*, s_1, \dots, s_k)$ where (s_1, \dots, s_k) is the order of the strings from S as they appear in \mathbf{v} for the first time. Then, the length of C is equal to

$$\begin{aligned} w(C) &= w_{s^*,s_1} + \sum_{j=1}^{k-1} w_{s_j,s_{j+1}} + w_{s_k,s^*} = 0 + \sum_{j=1}^{k-1} dist(s_j, s_{j+1}) + l(s_k) \\ &= l(s_1 +' \dots +' s_k) \leq l(\mathbf{v}). \end{aligned}$$

Since \mathbf{v} is a feasible solution to the SHORTEST λ -COVER SUPERSTRING problem, for every set X_i , at least λ strings from X_i appear as substrings of \mathbf{v} , and hence they also appear as vertices of C .

The other direction can be verified similarly. □

Hence, we seek a directed cycle C in G through vertex s^* that contains at least λ vertices from each set X_i , of minimal total length.

Define the variables

$$x_{ij} = \begin{cases} 1, & \text{if arc } (i, j) \text{ is in } C; \\ 0, & \text{otherwise.} \end{cases}$$

where (i, j) ranges over all ordered pairs of distinct elements of V , and

$$y_i = \begin{cases} 1, & \text{if vertex } i \text{ is in } C; \\ 0, & \text{otherwise.} \end{cases}$$

where i ranges over all elements of V .

Consider the following integer program

$$\begin{aligned} \min \quad & \sum_{i,j} w_{ij} x_{ij} \\ \text{s.t.} \quad & y_{s^*} = 1 \\ & \sum_{i \in V: i \neq j} x_{ij} = y_j \quad \forall j \in V \\ & \sum_{j \in V: j \neq i} x_{ij} = y_i \quad \forall i \in V \\ & \sum_{i \in X_j} y_i \geq \lambda \quad \forall j \in \{1, \dots, n\} \\ & 0 \leq x_{ij} \leq 1, \quad x_{ij} \text{ integer} \\ & 0 \leq y_i \leq 1, \quad y_i \text{ integer} \end{aligned} \tag{4}$$

There is a bijective correspondence between the set of feasible solutions of this integer program and the set of subgraphs of G that consist of one or more vertex-disjoint directed cycles, called *subtours*, such that s^* is contained in one of them. Due to Proposition 4.1, we are only interested in solutions that consist of a single directed cycle. There are several ways to eliminate the subtours. One possibility is to use the so-called Miller–Tucker–Zemlin (MTZ) formulation (Miller et al. 1960) (see also Pataki 2003), adding extra variables u_i ($i \in V$) and the constraints

$$\begin{aligned} & u_{s^*} = 1 \\ & 2 \leq u_i \leq |V| \quad \forall i \neq s^* \\ & u_i - u_j + 1 \leq (|V| - 1)(1 - x_{ij}) \quad \forall i \neq s^*, \forall j \neq s^*, j \neq i \end{aligned} \tag{5}$$

It indeed excludes subtours, as: (1) the last constraint for (i, j) forces $u_j \geq u_i + 1$, when $x_{ij} = 1$, and (2) if a feasible solution of (4)–(5) contained more than one subtour,

then at least one of these would not contain node s^* , and along this subtour the u_i values would have to increase to infinity.

Another way to exclude subtours is by adding to the original set of constraints the family of subtour (or subtour elimination) constraints

$$\sum_{i \in S, j \in S} x_{ij} \leq |S| - 1 \quad (S \subsetneq \text{supp}(y), \quad |S| > 1), \quad (6)$$

where $\text{supp}(y) = \{i \in V \mid y_i = 1\}$ is the support of y . (As the subtour inequality for $\text{supp}(y) \setminus S$ is a linear combination of the inequality for S and of the constraints $\sum_i x_{ij} = y_j$ and $\sum_j x_{ij} = y_i$, it is enough to use the subtour inequalities with S having size at most $|\text{supp}(y)|/2$.) These constraints are not linear since they depend on the values of y -variables. However, given an optimal solution (x, y) to the current IP formulation, we may add one or more constraints of the form (6) and solve the so obtained IP.

Following the approach of Pataki (2003), one can combine the MTZ and subtour formulations to obtain the ease of use of the first and some of the strength of the second. Denoting by **maxrounds** a non-negative integer parameter describing how much we want to strengthen the MTZ formulation and by **maxconstraints** an upper bound for the number of constraints we will add in each round, we obtain a “cutting-plane algorithm”, Algorithm 1. The pseudocode is shown below.

5 A hill-climbing algorithm

We have developed a hill-climbing algorithm to find short λ -superstrings for given strings S_1, \dots, S_k , a given set \mathcal{T} of target strings, and a given parameter λ . As in the formulation of the SHORTEST λ -SUPERSTRING problem, we have set the length of the λ -superstring as a function to minimize. We first select randomly an initial λ -superstring by taking the overlapping sum $\mathbf{v} = \mathbf{v}_1 +' \dots +' \mathbf{v}_k$, where each \mathbf{v}_i is likewise an overlapping sum of λ consecutive different substrings of S_i from \mathcal{T} , where the search for these strings begins at a randomly chosen initial point of string S_i (and continues at the beginning of the string, if necessary; here, if one target string appears more than once in an S_i we consider only one of them, randomly chosen, and then consecutive means consecutive with respect to the linear ordering of the target strings appearing in one S_i). This, of course, will result in a λ -superstring. Next, several transformations of two kinds are made to this initial candidate to λ -superstring. In the transformations of the first kind, a substring $\mathbf{v}_{i,j}$ is deleted from \mathbf{v} . In the transformations of the second kind, each substring $\mathbf{v}_{i,j}$ is changed for every possible substring of $S_1 +' \dots +' S_k$ from \mathcal{T} that is not already a substring of \mathbf{v} . Changes of the first kind and of the second kind are applied consecutively to \mathbf{v} (but each one of them only to \mathbf{v} , that is, they are not composed). If for one of them we continue having a λ -superstring and the length of the new λ -superstring \mathbf{v}' diminishes, then we replace \mathbf{v} with \mathbf{v}' and repeat again the sequence of substitutions. If, on the other hand, none of the changes diminishes the length of the λ -superstring, then we record the λ -superstring obtained and we choose again randomly a λ -superstring \mathbf{v} and repeat the process from the beginning. We do this for a prefixed number n of times and, finally, we take the shortest of the n λ -superstrings obtained.

Algorithm 1: Integer programming approach to SHORTEST λ -COVER SUPERSTRING.

Input: An instance $(A, X_1, \dots, X_n, \lambda)$ to the SHORTEST λ -COVER SUPERSTRING problem.

Output: A shortest λ -cover superstring for (X_1, \dots, X_n) .

- 1 Compute the distance graph $D = (V, E, w)$.
- 2 Set $k = 1$, and let the current IP formulation be as follows:

$$\begin{aligned}
 \min \quad & \sum_{i,j} w_{ij}x_{ij} \\
 \text{s.t.} \quad & y_{s^*} = 1 \\
 & \sum_{i \in V: i \neq j} x_{ij} = y_j \quad \forall j \in V \\
 & \sum_{j \in V: j \neq i} x_{ij} = y_i \quad \forall i \in V \\
 & \sum_{i \in X_j} y_i \geq \lambda \quad \forall j \in \{1, \dots, n\} \\
 & 0 \leq x_{ij} \leq 1, \quad x_{ij} \text{ integer} \\
 & 0 \leq y_i \leq 1, \quad y_i \text{ integer}
 \end{aligned}$$

- 3 **while** $k \leq \text{maxrounds}$ **do**
- 4 Solve the IP over the current formulation. Assume that the optimal solution consists of r subtours S_1, \dots, S_r .
- 5 **if** $r = 1$ **then**
- 6 The current (optimal) solution (x^*, y^*) is optimal. Let $C^* = (s^*, s_1, \dots, s_\ell)$ denote the corresponding directed cycle.
- 7 **return** $\mathbf{v} = s_1 +' \dots +' s_\ell$
- 8 **else**
- 9 Add to the formulation at most **maxconstraints** subtour constraints, in which S is the union of several S_i sets and $|S| \leq |\text{supp}(y)|/2$.
- 10 Set $k = k + 1$.
- 11 Solve to optimality the mixed integer program obtained from the current IP formulation by adding to it the following constraints:

$$\begin{aligned}
 & u_{s^*} = 1 \\
 & 2 \leq u_i \leq |V| \quad \forall i \neq s^* \\
 & u_i - u_j + 1 \leq (|V| - 1)(1 - x_{ij}) \quad \forall i \neq s^*, \forall j \neq s^*
 \end{aligned}$$

- 12 Let (x^*, y^*) denote the obtained optimal solution, and let $C^* = (s^*, s_1, \dots, s_\ell)$ denote the corresponding directed cycle.
- 13 **return** $\mathbf{v} = s_1 +' \dots +' s_\ell$

Below (Algorithm 2) is the pseudocode of the algorithm just described.

We have made numerical simulations to test our hill-climbing algorithm. We have generated several sets of 50 sequences of length 50 each with symbols on an alphabet of cardinality 20. For the set \mathcal{T} of target strings we took the set of all strings of length $\ell = 5$. We have used the hill-climbing algorithm to produce λ -superstrings for all possible values of λ , that is, for $\lambda = 1, \dots, 46$. The sets of sequences were generated of the following way: first, we generated a random *root sequence* of length 50 and, after that, we generated for each $\alpha = 1, \dots, 9$ a set of 50 sequences of length 50 by constructing first three variations of the root sequence; each variation was constructed from the root sequence by taking mutations in some position with probability of mutation in each position of $\alpha/100$. When a mutation was made in a

Algorithm 2: A randomized hill-climbing algorithm

Input: $S_1, \dots, S_k, \mathcal{T}, \lambda$.
Output: A λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$.

- 1 Set $C = 1, A = \emptyset$;
- 2 Choose randomly initial points m_1, \dots, m_k such that $m_i \in \{1, \dots, l(S_i)\}$;
- 3 Take $\mathbf{v} = \mathbf{v}_1 +^l \dots +^l \mathbf{v}_k$, with $\mathbf{v}_i = \mathbf{v}_{i,1} +^l \dots +^l \mathbf{v}_{i,\lambda}$ where $\mathbf{v}_{i,1}, \dots, \mathbf{v}_{i,\lambda}$ are λ different consecutive substrings of S_i belonging to \mathcal{T} , where the search for the $\mathbf{v}_{i,j}$'s begins in position m_i and, if necessary, continues at the beginning of S_i ;
- 4 **repeat**
- 5 Set *improvedSolutionFound* = *false*;
- 6 Set *listEp* = list of elements of \mathcal{T} in \mathbf{v} ;
- 7 **while** *listEp* \neq *empty list* **do**
- 8 $\mathbf{v}' = \mathbf{v}$ with the first element of *listEp* deleted;
- 9 **if** \mathbf{v}' is a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$ **then**
- 10 $\mathbf{v} = \mathbf{v}'$;
- 11 *improvedSolutionFound* = *true*;
- 12 Set *listEp* = empty list;
- 13 **if** *listEp* \neq *empty list* **then**
- 14 Remove the first element from *listEp*;
- 15 **if** *improvedSolutionFound* == *false* **then**
- 16 Set *listEp* = list of elements of \mathcal{T} in \mathbf{v}
- 17 **while** *listEp* \neq *empty list* **do**
- 18 *listNewEp* = the set of strings $T \in \mathcal{T}$ that are substrings of $S_1 +^l \dots +^l S_k$
- 19 **while** *listNewEp* \neq *empty list* **do**
- 20 $\mathbf{v}' = \mathbf{v}$ with the first element of *listEp* substituted by the first element of *listNewEp*;
- 21 **if** \mathbf{v}' is a λ -superstring for $(S_1, \dots, S_k, \mathcal{T})$ and $l(\mathbf{v}') < l(\mathbf{v})$ **then**
- 22 $\mathbf{v} = \mathbf{v}'$;
- 23 *improvedSolutionFound* = *true*;
- 24 Set *listEp* = empty list;
- 25 Set *listNewEp* = empty list;
- 26 **if** *listNewEp* \neq *empty list* **then**
- 27 Remove the first element from *listNewEp*;
- 28 **if** *listEp* \neq *empty list* **then**
- 29 Remove the first element from *listEp*;
- 30 **if** *improvedSolutionFound* == *false* **then**
- 31 Add \mathbf{v} to A ;
- 32 $C = C + 1$;
- 33 **until** $C > n$;
- 34 **return** the vaccine in A of minimum length.

position, a different symbol was selected with a uniform distribution of probability. Then, to construct each one of the 50 sequences, first one of the three variations was randomly selected and, after that, a new process of mutations was developed again in the way just described. In Table 1 the lengths of the obtained λ -superstrings and their coverages (in parentheses) are shown for λ in the range from 1 to 46 and for α in the range from 1 to 9. These lengths and coverages are shown in Fig. 1a and b, respectively. For every α there is first a slow increase in the length of the λ -superstrings and a rapid increase in the coverage. As λ grows, the increase in the length is accelerated, and the increase in the coverage is decelerated, obtaining a good tradeoff between small lengths and high coverage for medium values of λ around $\lambda = 30$. As we will see in Sect. 6, this phenomenon of good performance for intermediate values of λ seems to happen also when experimental data are considered.

Table 1 Length and coverage for 50 chains of length 50, epitopes of length 5, n (number of iterations) = 10,000

	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %
$\lambda = 1$	5 (0.022)	6 (0.042)	9 (0.037)	8 (0.04)	9 (0.038)	10 (0.037)	7 (0.037)	11 (0.05)	11 (0.034)
$\lambda = 2$	9 (0.085)	8 (0.084)	11 (0.076)	10 (0.072)	11 (0.069)	12 (0.071)	12 (0.07)	14 (0.1)	16 (0.077)
$\lambda = 3$	11 (0.127)	10 (0.126)	13 (0.114)	12 (0.122)	13 (0.102)	14 (0.125)	14 (0.133)	17 (0.144)	20 (0.112)
$\lambda = 4$	13 (0.17)	12 (0.166)	15 (0.151)	14 (0.18)	16 (0.141)	17 (0.157)	18 (0.147)	20 (0.168)	23 (0.148)
$\lambda = 5$	14 (0.203)	14 (0.186)	18 (0.179)	16 (0.211)	18 (0.18)	19 (0.198)	20 (0.189)	23 (0.184)	27 (0.195)
$\lambda = 6$	15 (0.232)	15 (0.186)	20 (0.218)	19 (0.2)	20 (0.217)	22 (0.221)	23 (0.237)	24 (0.213)	31 (0.208)
$\lambda = 7$	16 (0.246)	16 (0.247)	22 (0.264)	21 (0.233)	22 (0.26)	24 (0.241)	25 (0.247)	29 (0.275)	36 (0.24)
$\lambda = 8$	17 (0.275)	17 (0.263)	24 (0.301)	23 (0.294)	24 (0.276)	27 (0.28)	28 (0.302)	31 (0.315)	38 (0.281)
$\lambda = 9$	18 (0.287)	18 (0.284)	26 (0.277)	25 (0.352)	26 (0.284)	30 (0.316)	31 (0.301)	33 (0.329)	42 (0.307)
$\lambda = 10$	19 (0.308)	20 (0.325)	27 (0.315)	26 (0.365)	28 (0.313)	33 (0.364)	33 (0.324)	36 (0.361)	48 (0.32)
$\lambda = 11$	22 (0.347)	22 (0.352)	29 (0.331)	29 (0.365)	30 (0.352)	33 (0.371)	35 (0.349)	41 (0.389)	54 (0.327)
$\lambda = 12$	23 (0.367)	24 (0.345)	30 (0.353)	31 (0.406)	33 (0.396)	40 (0.458)	38 (0.39)	44 (0.408)	62 (0.369)
$\lambda = 13$	24 (0.387)	25 (0.366)	34 (0.384)	32 (0.406)	35 (0.377)	43 (0.447)	40 (0.421)	45 (0.436)	66 (0.397)
$\lambda = 14$	25 (0.408)	26 (0.386)	35 (0.41)	34 (0.444)	37 (0.399)	48 (0.442)	47 (0.446)	53 (0.47)	76 (0.433)
$\lambda = 15$	27 (0.417)	28 (0.426)	37 (0.462)	36 (0.473)	40 (0.474)	51 (0.482)	50 (0.459)	62 (0.496)	81 (0.445)
$\lambda = 16$	28 (0.439)	30 (0.461)	38 (0.461)	38 (0.486)	41 (0.459)	54 (0.518)	53 (0.503)	68 (0.516)	89 (0.469)
$\lambda = 17$	29 (0.46)	31 (0.472)	42 (0.493)	41 (0.48)	47 (0.545)	59 (0.528)	59 (0.509)	75 (0.545)	94 (0.487)
$\lambda = 18$	31 (0.478)	33 (0.503)	45 (0.501)	43 (0.5)	52 (0.531)	70 (0.55)	67 (0.534)	90 (0.561)	108 (0.519)
$\lambda = 19$	33 (0.557)	34 (0.523)	52 (0.587)	47 (0.533)	56 (0.557)	77 (0.578)	71 (0.548)	101 (0.582)	121 (0.547)
$\lambda = 20$	34 (0.589)	36 (0.556)	55 (0.633)	48 (0.564)	59 (0.61)	86 (0.611)	77 (0.563)	116 (0.593)	127 (0.566)
$\lambda = 21$	35 (0.599)	37 (0.562)	57 (0.579)	50 (0.583)	66 (0.593)	87 (0.633)	83 (0.579)	131 (0.614)	139 (0.577)
$\lambda = 22$	36 (0.62)	39 (0.593)	62 (0.7)	56 (0.602)	74 (0.65)	98 (0.637)	88 (0.613)	143 (0.63)	151 (0.605)
$\lambda = 23$	37 (0.64)	40 (0.613)	65 (0.682)	60 (0.64)	83 (0.657)	109 (0.658)	99 (0.647)	161 (0.655)	163 (0.619)

Table 1 continued

	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %
$\lambda = 24$	38 (0.676)	41 (0.633)	69 (0.737)	63 (0.666)	91 (0.693)	118 (0.684)	108 (0.663)	173 (0.648)	183 (0.652)
$\lambda = 25$	39 (0.68)	44 (0.678)	75 (0.74)	68 (0.677)	98 (0.715)	131 (0.696)	124 (0.673)	188 (0.656)	188 (0.666)
$\lambda = 26$	40 (0.717)	45 (0.71)	87 (0.796)	74 (0.684)	107 (0.737)	139 (0.723)	136 (0.704)	202 (0.675)	221 (0.677)
$\lambda = 27$	41 (0.738)	50 (0.745)	91 (0.802)	83 (0.712)	119 (0.773)	161 (0.739)	167 (0.701)	237 (0.697)	244 (0.692)
$\lambda = 28$	42 (0.758)	54 (0.77)	97 (0.82)	97 (0.72)	133 (0.772)	166 (0.766)	193 (0.719)	263 (0.71)	271 (0.699)
$\lambda = 29$	43 (0.778)	55 (0.796)	106 (0.826)	102 (0.747)	146 (0.788)	198 (0.761)	219 (0.727)	280 (0.723)	299 (0.712)
$\lambda = 30$	44 (0.798)	56 (0.815)	128 (0.834)	112 (0.786)	161 (0.783)	215 (0.766)	256 (0.741)	314 (0.737)	335 (0.72)
$\lambda = 31$	45 (0.818)	57 (0.834)	131 (0.805)	120 (0.8)	178 (0.801)	236 (0.779)	272 (0.761)	352 (0.745)	370 (0.737)
$\lambda = 32$	46 (0.84)	62 (0.89)	141 (0.843)	131 (0.818)	200 (0.814)	263 (0.782)	307 (0.768)	403 (0.762)	429 (0.746)
$\lambda = 33$	47 (0.86)	68 (0.903)	152 (0.867)	149 (0.808)	228 (0.823)	280 (0.797)	345 (0.782)	436 (0.778)	474 (0.759)
$\lambda = 34$	48 (0.88)	70 (0.917)	160 (0.872)	164 (0.827)	252 (0.833)	310 (0.807)	386 (0.792)	473 (0.79)	515 (0.779)
$\lambda = 35$	49 (0.902)	82 (0.919)	175 (0.888)	177 (0.855)	281 (0.837)	343 (0.82)	429 (0.805)	530 (0.802)	561 (0.797)
$\lambda = 36$	50 (0.923)	91 (0.925)	186 (0.897)	195 (0.864)	307 (0.849)	400 (0.83)	467 (0.82)	579 (0.82)	622 (0.81)
$\lambda = 37$	68 (0.93)	115 (0.925)	219 (0.903)	242 (0.869)	345 (0.853)	479 (0.847)	567 (0.837)	650 (0.833)	699 (0.829)
$\lambda = 38$	77 (0.938)	129 (0.929)	242 (0.909)	275 (0.884)	380 (0.867)	522 (0.858)	614 (0.854)	701 (0.85)	760 (0.844)
$\lambda = 39$	84 (0.946)	142 (0.933)	264 (0.915)	307 (0.897)	423 (0.88)	564 (0.873)	660 (0.87)	749 (0.864)	816 (0.863)
$\lambda = 40$	95 (0.953)	157 (0.941)	286 (0.922)	329 (0.904)	485 (0.892)	626 (0.891)	737 (0.887)	812 (0.882)	878 (0.882)
$\lambda = 41$	103 (0.961)	179 (0.95)	308 (0.93)	382 (0.916)	540 (0.908)	685 (0.905)	792 (0.904)	869 (0.901)	934 (0.9)
$\lambda = 42$	178 (0.969)	247 (0.96)	409 (0.943)	495 (0.932)	636 (0.927)	795 (0.923)	894 (0.923)	971 (0.919)	1031 (0.919)
$\lambda = 43$	197 (0.976)	280 (0.968)	441 (0.956)	563 (0.948)	691 (0.945)	861 (0.944)	961 (0.944)	1037 (0.94)	1083 (0.938)
$\lambda = 44$	214 (0.983)	321 (0.978)	498 (0.971)	608 (0.965)	748 (0.963)	904 (0.963)	1015 (0.962)	1089 (0.959)	1149 (0.958)
$\lambda = 45$	238 (0.992)	349 (0.989)	534 (0.985)	648 (0.981)	810 (0.981)	955 (0.981)	1080 (0.981)	1139 (0.979)	1220 (0.98)
$\lambda = 46$	257 (1)	391 (1)	577 (1)	715 (1)	873 (1)	1007 (1)	1130 (1)	1203 (1)	1292 (1)

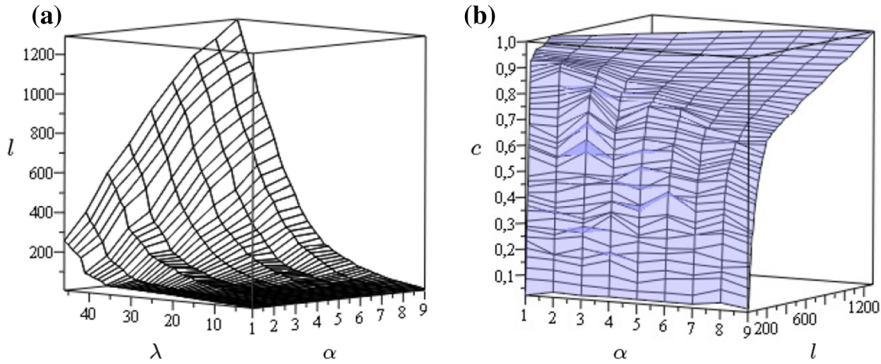


Fig. 1 Length l and coverage c of the λ -superstrings for the numerical simulations

6 Numerical computations on experimental data

In this section we describe an application of the integer programming algorithm and of the hill-climbing algorithm introduced in the previous two sections to obtain optimal and feasible λ -superstrings, respectively, for sets of strings S_1, \dots, S_k obtained experimentally and taken from biological sequence databases.

6.1 Hill-climbing algorithm for hemagglutinin

Giles and Ross (2011) succeeded in designing and elaborating a vaccine which protected mice and ferrets against clade 2 H5N1 by using their computationally optimized broadly reactive antigen (COBRA) system. In the design of their vaccine they used 129 input sequences from human clade 2 infections. We used 123 of such sequences to test our algorithms. The reason for not using all of them is that, although of course not all of them have the same length (in fact, it would be very unusual if all had the same length), six of them had significantly smaller length than the rest of the sequences, and hence we didn't include them in our calculations so that we can work with high values of the parameter λ . We include in Table 2 the GenBank (2013) IDs of the sequences corresponding to the hemagglutinin (HA) genes. We ran our hill-climbing for that set of sequences with 10,000 iterations for values of the parameter λ taken in steps of 10 from $\lambda = 10$ to $\lambda = 500$, by taking an alphabet A of cardinality 20 representing the amino acids and taking $\mathcal{T} = A^{10}$ as the set of target strings. In Table 3 the lengths of the λ -superstrings obtained and their coverages are shown. These lengths and coverages are plotted in Fig. 2a and b, respectively. As shown in the figures and in the table, the performance of the λ -superstrings is better for small and medium values of λ , in the sense that for relatively small λ , say of about 360, the length of the λ -superstrings (the candidate vaccine) is relatively small and it keeps below the average length of the hemagglutinin, which is 559.9 for the set of 123 sequences, and at the same time the coverage is high, being over the 73 % of the epitopes. As λ increases, the performance of the λ -superstring is not so good, because although the coverage increases, which is desirable, also the length increases considerably, and this can be problematic, as we

Table 2 GenBank IDs of the sequences for the hemagglutinin

EU146737	EU146688	CY014203	CY014457	EU015407	FJ492886	FJ492881
EU146753	EU146713	CY014205	CY014518	EU015408	EF200512	HM114537
EU146745	EU146697	CY014206	CY014510	EU015409	EF200513	FJ492880
EU146793	EU146705	CY014207	CY014481	EU015410	DQ464377	
EU146755	EU146729	CY014209	CY014489	EU015411	EU095023	
EU146785	EF541394	CY014210	CY014497	EU015412	EU095024	
CY014272	EU146777	CY014211	CY014177	EU015413	EU146867	
CY014280	EU146801	CY014212	CY014529	EU015414	EU146868	
CY014288	EU146809	CY014213	CY014543	EU015416	DQ371928	
CY014296	EU146817	CY014311	CY017662	CY062439	DQ371929	
CY014303	EU146825	CY014368	CY017670	DQ435202	EF624256	
CY014477	EU146632	CY014376	CY017678	EU146870	DQ835313	
CY014433	CY014197	CY014384	CY017688	EU146876	FJ492882	
CY014465	CY014160	CY014393	CY017638	EU146877	FJ492884	
EU146648	CY014198	CY014401	CY017646	EU146869	DQ371930	
EU146640	CY014199	CY014409	CY017654	EU146878	EU263981	
EU146656	CY014200	CY014417	EU015403	EF619982	FJ492879	
EU146664	CY014201	CY014425	EU015404	EF619989	FJ492885	
EU146672	CY014204	CY014441	EU015405	EF619990	AB462295	
EU146681	CY014202	CY014449	EU015406	EF619998	EF137706	

said in Sect. 2. Nonetheless, even for a value of $\lambda = 490$ the length is less than twice the average length of the protein and the coverage is over 95 %.

6.2 Hill-climbing algorithm with addition of frequent epitopes for Nef and Gag

Now we will do a comparative study with the results obtained by [Nickle et al. \(2007\)](#). In that paper they considered, for the Nef and Gag proteins in HIV-1, all 9mer peptides in a 169-sequence dataset taken from GenBank (2013) as basic pieces to test the obtained coverages. Their method was based in calculating first a center of tree sequence (COT) derived from a phylogenetic analysis of different strains followed by a second stage when they add several frequent 9-amino acid sequences (9mers). For the addition of frequent 9mers and for the subsequent calculation of the coverages they considered, as we said before, the set of all sequences of length 9. They constructed sequences of different lengths with relatively high coverages, and they emphasized the case when their sequences had three-genes length, because beyond that value the increase in coverage was lower with respect to increase in length. They obtained sequences of three-genes length with a coverage of 62 % in the Nef protein and of 82 % in the Gag protein. By using our λ -superstrings we have obtained, for the same set of 169 sequences (shown in Table 4) in the case of Nef protein and for a subset of 166 sequences (shown in Table 5) in the case of Gag protein, the same level of coverage after

Table 3 Length and coverage of the λ -superstrings for hemagglutinin with 10,000 repetitions

$\lambda = 10 : 19$ (0.018)	$\lambda = 210 : 263$ (0.417)	$\lambda = 410 : 729$ (0.835)
$\lambda = 20 : 29$ (0.036)	$\lambda = 220 : 273$ (0.434)	$\lambda = 420 : 773$ (0.837)
$\lambda = 30 : 39$ (0.054)	$\lambda = 230 : 298$ (0.453)	$\lambda = 430 : 790$ (0.854)
$\lambda = 40 : 49$ (0.073)	$\lambda = 240 : 319$ (0.493)	$\lambda = 440 : 820$ (0.872)
$\lambda = 50 : 64$ (0.099)	$\lambda = 250 : 337$ (0.514)	$\lambda = 450 : 861$ (0.884)
$\lambda = 60 : 79$ (0.125)	$\lambda = 260 : 356$ (0.534)	$\lambda = 460 : 914$ (0.906)
$\lambda = 70 : 89$ (0.143)	$\lambda = 270 : 370$ (0.558)	$\lambda = 470 : 973$ (0.929)
$\lambda = 80 : 99$ (0.161)	$\lambda = 280 : 390$ (0.585)	$\lambda = 480 : 1,031$ (0.937)
$\lambda = 90 : 109$ (0.179)	$\lambda = 290 : 406$ (0.609)	$\lambda = 490 : 1,095$ (0.953)
$\lambda = 100 : 119$ (0.198)	$\lambda = 300 : 422$ (0.62)	$\lambda = 500 : 1,153$ (0.958)
$\lambda = 110 : 129$ (0.215)	$\lambda = 310 : 436$ (0.64)	
$\lambda = 120 : 139$ (0.234)	$\lambda = 320 : 461$ (0.679)	
$\lambda = 130 : 156$ (0.252)	$\lambda = 330 : 483$ (0.698)	
$\lambda = 140 : 166$ (0.27)	$\lambda = 340 : 504$ (0.717)	
$\lambda = 150 : 176$ (0.288)	$\lambda = 350 : 523$ (0.71)	
$\lambda = 160 : 190$ (0.311)	$\lambda = 360 : 545$ (0.735)	
$\lambda = 170 : 210$ (0.338)	$\lambda = 370 : 574$ (0.774)	
$\lambda = 180 : 226$ (0.353)	$\lambda = 380 : 600$ (0.785)	
$\lambda = 190 : 242$ (0.381)	$\lambda = 390 : 650$ (0.807)	
$\lambda = 200 : 252$ (0.399)	$\lambda = 400 : 691$ (0.826)	

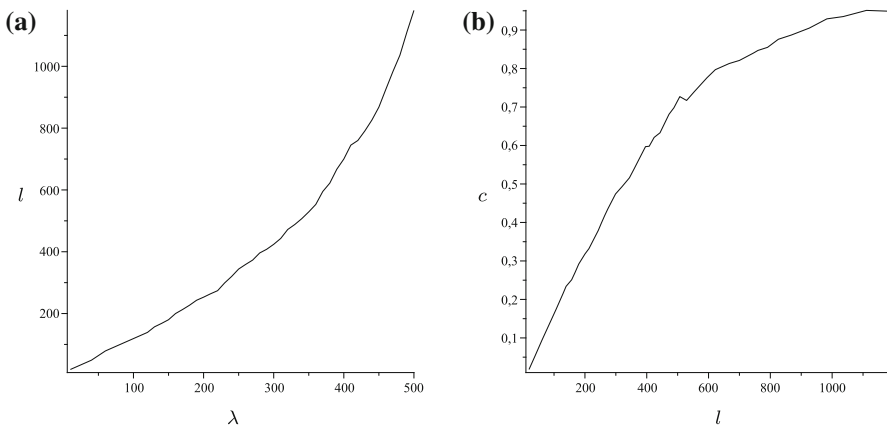


Fig. 2 Length l and coverage c of the λ -superstrings for hemagglutinin

rounding to an integer value (61.75 % for Nef and 81.59 % for Gag). The procedure that Nickle et al. followed was to obtain first a “center of the tree” sequence and, after that, adding a set of frequent epitopes until the desired length is obtained.

We used a similar two stages method in which we first computed a λ -superstring that played the role of the center of the tree in Nickle et al.’s method, followed by a second stage in which we added the most frequent epitopes in a way that they overlap well with the epitopes present in the λ -superstring.

Table 4 GenBank IDs of the sequences for the Nef protein

AB012824	AF120887	AF129375	AF203172	AF238268	AY121441	AY835772	L15515	U34603
AB034257	AF120898	AF129376	AF203180	AF252897	AY173951	AY835776	L15518	U43106
AB034272	AF120909	AF129377	AF203188	AF252910	AY308762	AY835779	M17451	U44444
AB078005	AF129334	AF129378	AF203192	AF462708	AY314063	AY835780	M21098	U44450
AB221005	AF129335	AF129379	AF203194	AF462753	AY3331285	AY857022	M26727	U44462
AF004394	AF129342	AF129382	AF203198	AF538302	AY3331290	AY857144	M58173	U44468
AF011471	AF129343	AF129388	AF219672	AF538304	AY3331293	AY899356	M93259	U66543
AF011474	AF129346	AF129389	AF219685	AF538305	AY352275	AY899382	U03295	U69584
AF011481	AF129347	AF129390	AF219691	AF538306	AY444311	DQ007902	U03338	U71182
AF011487	AF129350	AF129392	AF219729	AJ271445	AY713408	DQ085869	U03343	
AF011493	AF129351	AF129394	AF219755	AJ430664	AY739040	DQ121815	U12055	
AF042101	AF129352	AF203108	AF219760	AY037269	AY779550	DQ121883	U16863	
AF047082	AF129354	AF203111	AF219765	AY037282	AY786630	DQ127537	U16875	
AF063926	AF129355	AF203116	AF219771	AY116676	AY786750	DQ127548	U16934	
AF069139	AF129362	AF203126	AF219782	AY116713	AY835748	DQ487191	U23487	
AF120745	AF129364	AF203137	AF219792	AY116714	AY835751	DQ659737	U24455	
AF120772	AF129369	AF203141	AF219800	AY116727	AY835753	L07422	U26087	
AF120840	AF129370	AF203153	AF219812	AY116781	AY835762	L15482	U26110	
AF120851	AF129372	AF203161	AF219819	AY116805	AY835765	L15489	U26119	
AF120867	AF129373	AF203165	AF219845	AY116830	AY835770	L15500	U26138	

Table 5 GenBank IDs of the sequences for the Gag protein

AB078005	AF146728	AY173951	AY206663	AY560108	AY786962	AY835776	DQ295192	U34604
AB078703	AF224507	AY173952	AY206664	AY560110	AY818644	AY835777	DQ295193	U39362
AB078704	AF256204	AY173954	AY247251	AY679786	AY819715	AY835778	DQ295195	U43096
AB078709	AF286365	AY173955	AY275555	AY682547	AY835748	AY835779	DQ487188	U43141
AB078711	AF538302	AY173956	AY275556	AY751406	AY835751	AY835780	DQ487189	U69584
AB097870	AF538303	AY180905	AY275557	AY751407	AY835753	AY839827	DQ487190	U71182
AB221005	AF538304	AY206647	AY308760	AY779550	AY835754	AY857022	DQ487191	
AF003887	AF538305	AY206648	AY308762	AY779553	AY835755	AY857144	K02007	
AF004394	AF538306	AY206649	AY314044	AY779556	AY835757	AY970946	L02317	
AF042100	AF538307	AY206651	AY314063	AY779557	AY835758	AY970950	M13136	
AF042101	AJ271445	AY206652	AY331283	AY779563	AY835759	CQ958304	M17451	
AF042102	AJ437030	AY206653	AY331285	AY779564	AY835761	D10112	M19921	
AF042103	AJ437033	AY206654	AY331287	AY786790	AY835762	DQ085869	M26727	
AF042104	AJ437038	AY206656	AY331290	AY786830	AY835764	DQ097739	M38429	
AF042105	AJ437039	AY206657	AY331292	AY786870	AY835765	DQ097744	M38431	
AF049494	AJ437044	AY206658	AY331297	AY786910	AY835766	DQ097747	M93258	
AF049495	AJ437047	AY206659	AY332236	AY786919	AY835769	DQ127536	U21135	
AF069140	AJ437050	AY206660	AY352275	AY786920	AY835770	DQ127539	U23487	
AF075719	AJ437051	AY206661	AY423387	AY786949	AY835772	DQ127542	U26546	
AF086817	AJ437058	AY206662	AY560107	AY786952	AY835774	DQ127548	U34603	

We first calculated, for a given λ , and considering as the set of epitopes all the subsequences of length 9, a λ -superstring following the hill-climbing algorithm described in Sect. 5. Then, we ranked the set of epitopes attending at both the frequency of the epitopes and the level of overlapping of the epitopes in the set \mathcal{T}_v of epitopes in the λ -superstring. More specifically, we assigned to each epitope e not in \mathcal{T}_v the fitness

$$\phi(e) = p \cdot rf(e) + \frac{1 - p}{18|\mathcal{T}_v|} \cdot \sum_{e' \in \mathcal{T}_v} (ov(e, e') + ov(e', e)),$$

where $rf(e)$ is the relative frequency of the epitope and p is a parameter which determines the ponderation in the fitness of the high frequency with respect to the level of overlapping with the epitopes in \mathcal{T}_v . The constant 18 is introduced in the equation defining the fitness to normalize each sum of two bilateral overlappings so that the quotient is between 0 and 1. After doing the ranking, we sorted the epitopes not in \mathcal{T}_v in descending order with respect to $\phi(e)$ and we did, for different values of n , the following procedure until we reached the desired length of three-genes length: We added to the initial λ -superstring the first n epitopes with higher fitness and then we applied the usual greedy algorithm to find an approximation to a shortest common superstring of the obtained set of epitopes (see, e.g., [Tarhio and Ukkonen 1988](#)). A heuristic study suggested that, for the problem of the Nef protein studied by Nickle et al., a value of $p = 0.99$ was appropriate for $\lambda = 45$. We run the hill-climbing algorithm with 10,000 iterations and added frequent epitopes as described above 30 times and we took the solution with the biggest coverage, which was 61.75.

The found solution, of length 621, was:

```
YTPGPGTRFPLTFGWCFKLVDPVDPPEEVGFPVKPQVPLRPMTYKAAVDLSHFLQNYTPGPGTRYPLTFGWCFKLVPEVD
QNYTPGPGVRYPLTFGWPTVRERMRRAEPAEAGVGVAVSRDLERHGAITSSNTAATNADCAWLERPMTYKAAALDLSHFLR
EKGGLEGLIHSQKRQDILDWVIYHTQGYFPAADGVGAASRDLEKHGMDDPEREVLEWRFD SRLAFPHHVARELHPEYKYD
CFKLVPEPEKIEEANEGENNSLLHPMSLHGMEDEPEKVLVWKFDSRLVPEPEKVEEANEGENNCLLHPMSQHMGGKW
SKRSVEKANEGENNAACAWLEAQEDEVGPPVRPQVPLRPMTYKGAALDLSHFLKEAREKHPEYKQRQEILDWVYHTQG
YFPDWMGGKWSKSSITSSNTAANNADCAWLEAQEEEEVGFVPRPMTYKGAVDLSHFLKEGGLEGLVYSQRRQDILDW
VYHNSLLHPMSQHGMDDEPEKVLMWKFDSRLAFPHMARELHPEYKNCLLHPMSLHGMDDEPEKGGLEGLIYSQKRQDIL
DLWVYNTQGYFPDWQNYTPGPGIRYPLTFGWPAVRERMRRAEPAADGVGAASRDLEKHGAITSSNTAT
```

We analyzed how well this solution captures the well-conserved regions of the sequence population. O’Neill et al. (2006, Fig. 1) studied the frequency of the amino acids at each position in Nef protein of HIV-1. They noted that 63 residues were very well conserved at 99 %. Those 63 residues were scattered through the protein and the maximum number of consecutive ones is 5. To analyze longer series of consecutive residues, we studied the ones conserved at 90 %, and we found in O’Neill et al.’s table 144 such residues distributed in 12 groups of a single residue, 5 groups of two consecutive ones, 7 groups of length 3, 5 groups of length 4, one group of length 5, two groups of lengths 6, 7 and 8, respectively, and one group of lengths 9, 12 and 13, respectively. *All those 39 sequences appear as subsequences of our solution.*

One can wonder how much the value of 62 % for the coverage can be improved. The following general bound for the coverage is trivial to prove. In it, $S_1, \dots, S_k, f(\mathbf{t})$ and $c(\mathbf{v})$ are as in Definition 2.6, and $\mathcal{T} = A^\ell$.

Proposition 6.1 *If $(\mathbf{t}_i)_{1 \leq i \leq n}$ is a list of the elements in A^ℓ with $f(\mathbf{t}_i) \geq f(\mathbf{t}_{i+1}) \forall i$ and $v \in A^m$, then*

$$c(\mathbf{v}) \leq \frac{\sum_{i=1}^{m-l+1} f(\mathbf{t}_i)}{\sum_{i=1}^n f(\mathbf{t}_i)}.$$

In our calculations we took the length of a gene for the Nef protein to be 207, because the mean of the lengths of the 169 sequences is 207.11. Hence, the length for a three-genes length Nef protein is 621, and the previous proposition shows that the coverage for such a sequence of length 621 corresponding to $l = 9, k = 169$ and S_1, \dots, S_{169} obtained from the mentioned GenBank (2013) sequences is 67.8. This, of course, doesn't mean that a coverage of 67.8 can be found by using other methods, because to obtain that value it should occur that 613 sequences in $\mathcal{T} = A^\ell$ with the highest frequencies can be assembled in such a way that each one of them overlaps with the following one in 8 positions, and this situation is very unlikely in the general case when the number of sequences is big.

We did a similar analysis for the Gag protein. We analyzed 166 of the 169 sequences considered by Nickle et al. We did not use a nonfunctional gag protein gene and two very short sequences which we excluded because, as we told in the previous subsection, λ -superstrings are interesting only when the length of the sequence is big enough with respect to λ . For Gag, we applied the procedure described above with $p = 0.999, \lambda = 50$ for 1,000 iterations and repeated the whole process 30 times and we took the solution with the biggest coverage, which was 81.59, for a three-genes length. The solution that we found, of length 1,495, was:

```

MGARASVLSGGQLDRWEKIRLRPGGKKKYLKHIVWQEQIGWMTNPNPVPVGEVLYPLASRLSPGNDPLSQSSEELRSL
YNTVATLYCVHQRIEIKDTEKALEKIEEEQNKTLRAEQASQDVKNWMTETLLVQANANPDCKTILKALGPAATLEEMMTA
CQGVGGPSHKARVLAEAMSQATGSEELKSLFNTVATLYCVHQKIDVKDTKEALEWDRHLHPVQAGPVAPGQNYPIVQNIQ
GQMVHQAI SPRTLNAWVKVIEEKAFSPVPIPMFSALESEGATPNSATIMMQKGNFRNQRKTKVCFNCGKKGCWKCGKEGH
QMKDCLRAEQASQEVKNWMTMGARASILSGGKLDKWELRSLYNTIATLYCVHQRIEIKDTEKALEKIEEENKSKKKAQ
QAAAGTGNSSGCRQLGQLQPSLQGTGNNSQVSNYPIVQNMQGMVHQALSPRTLNAWVKVIEEKAFSPVPIPMFTAL
SEGATPQDLNLTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHAGPLHPVHAGPIAPGQIREPRGSDIAGTTSTLQEQ
IGWMTSNPPIVGEIYKRWIILGLNKIVRMYSPSILDIKQGGPKPPFRDYDRFYKTLRAEQATQEVKNWMTETLLVQN
SNPDCKTILKALGPGATLEEMMTFLQSRPEPSAPPEESFRPGGKKKYRLKHLVWASRELERFALNPGLETSEGCRQIL
EQLQPALQGMVHQAGPIAPGQMREPIKCFNCGKEGHIAKNCRAPRKRKGCWKCGKEGHIAKNCRAPRKKKYRLKHIVWA
SPTSILDIRGGPSHKARILAEAMSQVTNPACQGVGGPGHKARVLAEAMSQVTNSATVMMQGRGNFRNQRKIVKCFNCGK
EGHLAEAMSQMTSTLQEQIAWMTNPNPPIVGDYIKRWIIEVRDTEKALEDKIVRMYSPSILDIRGQPKPPFRASVLSGG
KLDREKIRLRPGGKKKYLKHIVWASRELERFAVNPGLLETSGGCRQILEQLQPSLQGTGSEELKSLYNTVATVNPGLL
ETAEGCRQILGQLPALQGTGSEELRSLFNTVATVLSGGELDKWEKIRLRPGGRKKAQQAADTGNSSQVSNYPIVQNL
QGMVHQAI SPRTLNAFLGKIWPSYKGRPGNFLQSRPEPTAPPEESFRFGEEATATPVEEENKSKKKAQQTAAASYNTI
AVLYCVHQKIEVKDTKEEAAEWDRHLHPVHAGPVAPGQMREPRGSDIAGTTSNLQEQIGWMTNPNPVPVGEIYKRWIIMG
LNKIVRMYSPSILDIMMQRGNFKNRKNCRAPRKRKGCWKCGREGHQMKDCTERQANFLGKIWPSHKGRPGNFLQNRPE
PTAPPAESFRFGEEITPPQKQEPIDKELYPLASLKSFLGNDPSFGEETTTSPQKQEPIDKELYPLASLKSFLGNDPSF
LYCVHQRIDVKDTKELYPLTSLRSLFGNDPSSQVNTSATIMMQRGNFRASVLSGGELDRWEKIRLRPGGKKRY
    
```

In this case, the obtained coverage is also close to the upper bound given by Proposition 6.1, which is 85.4.

6.3 Integer programming algorithm for Nef

An optimal solution to the integer programming problem derived in Sect. 4, extended with the MZT formulation, provides an optimal solution to the SHORTEST λ -COVER SUPERSTRING problem. Thus, the (IP) approach could in principle give better solutions than the hill-climbing method from Sect. 5. However, denoting $t = |\mathcal{T}|$, notice that the derived IP has $t^2 + 3t + 2$ variables (of which $t^2 + 2t + 1$ are integer-valued and $t + 1$ are real-valued), and $3t^2 + 7t + 4 + n$ linear constraints (recall that n is the number of collections of input strings). Therefore, this limits the applicability of the IP approach to our biological setting if the set of epitopes is given by $\mathcal{T} = A^\ell$, with $|A| = 20$ and $\ell \in \{9, 10\}$ (as was done in Subsects. 6.1 and 6.2), as it would amount to a number of variables and constraints exceeding 10^{23} . Nevertheless, the approach can be useful if the set of epitopes consists of several hundreds of epitopes.

We implemented in Java (2013) our integer programming model described in Sect. 4 (extended with the MZT formulation) using IBM® ILOG® CPLEX® Optimization Studio (2013), and applied it to a set of epitopes for the Nef protein in HIV-1 taken from the HIV Molecular Immunology Database (2013). We applied the algorithm to the 346 distinct epitopes found using that database for the set of 169 sequences mentioned in the previous section for λ ranging between 1 and 20. We thus have $t = 346$, $n = 169$, and the resulting IPs (one for each value of λ) had 120,756 variables and 361,743 constraints. For $\lambda = 20$ we obtained the following 20-superstring of length 131:

```
FLKEKGLDGLWLEAQEEVEVGFPPRPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIYSQKRQDILDLWVYHTQGYFPD
WQNYNTPGPGIRYTPGPGVRYPLTFGWCFKLVFVWKFDSRLAFHHVARELHPEY
```

A comparative study of the integer programming algorithm with the hill-climbing one is feasible only for relatively small values of λ : for big values of λ and big sets of epitopes the integer programming algorithm is not effective because of the required usage of memory and computation time. In order to compare the performance of the integer programming algorithm to the one of the hill-climbing algorithm, we have calculated the length of the λ -superstring for λ ranging from 1 to 20 for the following algorithms:

1. We took a random selection of epitopes until we obtained a λ -superstring, and then we did the overlapping sum of the epitopes. This process was repeated 10^6 times, and then the shortest one was selected.
2. The hill-climbing algorithm presented in Sect. 5 was applied 10^5 times.
3. The optimal integer programming algorithm presented in Sect. 4 and implemented as described above was used.

The lengths of the λ -superstrings are showed in Table 6 and in Fig. 3. As expected, the hill-climbing algorithm and the integer programming algorithm both outperformed notably the brute force algorithm consisting in a random concatenation of epitopes. Obviously, the optimal solution given by the integer programming algorithm was shorter than the suboptimal solution given by the hill-climbing algorithm. Although for small values of λ the lengths of the three solutions are practically the same, when λ

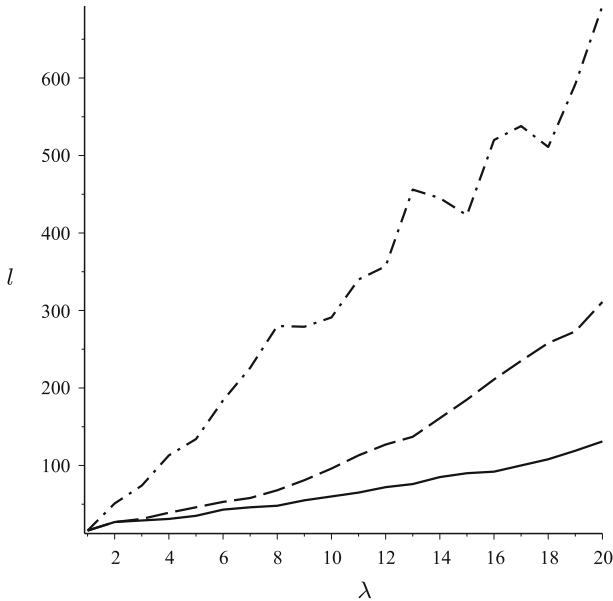


Fig. 3 Length l of the λ -superstrings as a function of λ for a random concatenation of epitopes (*dashed-dotted*), the hill-climbing algorithm (*dashed*) and the integer programming algorithm (*solid line*)

increases the lengths of the λ -superstrings obtained with the three algorithms diverge. Thus, for $\lambda = 20$, the length of the suboptimal solution found by the hill-climbing algorithm is 45 % of the length of the one found by the brute force algorithm, and the length of the optimal solution found by the integer programming algorithm is 19 % of the length of the one found by the brute force algorithm.

7 Conclusion and future work

In this paper we have introduced two new problems of combinatorial optimization, and presented an application of these problems to the computational design of vaccines. When applied to this biological problem, even suboptimal solutions have relatively small lengths and good levels of coverage and, at the same time, due to the combinatorial properties which define them, present an adequate balance of epitopes over the selected sample strings. We have presented two approaches to give optimal and suboptimal solutions to the problem: one based on a hill-climbing method, which produces suboptimal solutions, and one based on integer programming, which produces optimal solutions. Although this latter approach could in principle give better solutions, nonetheless it is not effective when the set of target strings is formed by all the substrings of a given length. Since it is sensible to restrict the set of epitopes attending to biological criteria, one future line of research that we will follow consists of studying more closely the case when we restrict our set of epitopes to a smaller

Table 6 Lengths of the λ -superstrings for λ between 1 and 20 for a random concatenation of epitopes, the hill-climbing algorithm and the integer programming algorithm

λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Random	16	51	74	113	134	184	226	280	279	291	340	357	456	445	423	520	538	511	592	693
Hill-climbing	16	27	31	39	46	53	58	68	81	96	113	127	137	161	185	211	235	258	273	311
Integer programming	16	27	29	31	35	43	46	48	55	60	65	72	76	85	90	92	100	108	119	131

number, biologically significant and at the same time affordable for using the integer programming approach.

Another future line of research lies in constructing a λ -superstring \mathbf{v} of about one-gene length with λ as big as possible, and where the set \mathcal{T} is taken to be A^ℓ for a set A of cardinality 20 and a small fixed value of ℓ . This strategy of taking for \mathcal{T} the complete set A^ℓ of small sequences could hopefully produce good candidates for immunogenic synthetic proteins. The λ -superstring \mathbf{v} should satisfy, of course, additional conditions such that, for instance, the relative order of elements in \mathcal{T} is, to the extent it is possible, the same in \mathbf{v} and in the sequences S_1, \dots, S_k , or that the fold of the associated protein is preserved. In this setting the elements of \mathcal{T} are not, strictly speaking, epitopes, but they are instead elementary pieces for a construction of the synthetic protein.

A third future line of research consists in formulating mathematical restrictions other than the level of coverage that could be biologically significant in the elicitation of an immune response. We recognize that the ultimate test of the efficacy of any algorithm is the *in vivo* design and validation of a vaccine, without which it is hard to predict the biological impact. We hope that, nevertheless, this work presents an alternate approach to vaccine design that can be evaluated more biologically in future projects.

References

- Allegrini P, Buiatti M, Grigolini P, West BJ (1998) Fractional brownian motion as a nonstationary process: an alternative paradigm for DNA sequences. *Phys Rev E* 57(4):4558
- Alon N, Moshkovitz D, Safra S (2006) Algorithmic construction of sets for k -restrictions. *ACM Trans Algorithms (TALG)* 2(2):153–177
- Audit B, Vaillant C, Arnéodo A, d'Aubenton-Carafa Y, Thermes C (2004) Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences. *J Biol Phys* 30(1):33–81
- Ausiello G, Protasi M, Marchetti-Spaccamela A, Gambosi G, Crescenzi P, Kann V (1999). *Complexity and approximation: combinatorial optimization problems and their approximability properties*, 1st edn. Springer, Secaucus
- Blum A, Jiang T, Li M, Tromp J, Yannakakis M (1994) Linear approximation of shortest superstrings. *J ACM (JACM)* 41(4):630–647
- De la Fuente IM, Martínez L, Benítez N, Veguillas J, Aguirregabiria J (1998) Persistent behavior in a phase-shift sequence of periodical biochemical oscillations. *Bull Math Biol* 60(4):689–702
- De la Fuente IM, Vadillo F, Pérez-Pinilla M-B, Vera-López A, Veguillas J (2009) The number of catalytic elements is crucial for the emergence of metabolic cores. *PLoS ONE* 4(10):e7510
- Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, Funkhouser R, Kuiken C, Haynes B, Letvin NL, Walker BD et al (2006) Polyvalent vaccines for optimal coverage of potential t-cell epitopes in global HIV-1 variants. *Nat Med* 13(1):100–106
- Gallant J, Maier D, Storer JA (1980) On finding minimal length superstrings. *J Comput Syst Sci* 20(1):50–58
- Garey MR, Johnson DS (1979) *Computers and intractability: a guide to the theory of NP-completeness*. W. H. Freeman & Co., New York
- GenBank (2013). <http://www.ncbi.nlm.nih.gov/genbank/> (Online; Accessed 21 Sept 2013)
- Giles BM, Ross TM (2011) A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. *Vaccine* 29(16):3043–3054
- Goldbeter A (1997) *Biochemical oscillations and cellular rhythms: the molecular bases of periodic and chaotic behaviour*. Cambridge University Press, Cambridge
- Henry-Labordere A (1969) The record balancing problem: A dynamic programming solution of a generalized traveling salesman problem. *Rev Franç Inform Rech Opér* 3(B-2):43–49
- HIV Molecular Immunology Database (2013). <http://www.hiv.lanl.gov/content/immunology/> (Online; Accessed 21 Sept 2013)

- Holley LH, Goudsmit J, Karplus M (1991) Prediction of optimal peptide mixtures to induce broadly neutralizing antibodies to human immunodeficiency virus type 1. *Proc Natl Acad Sci USA* 88(15):6800–6804
- Ibm, ILOG CPLEX Optimization Studio (2013). <http://www-03.ibm.com/software/products/us/en/ibmilogcplexoptistud/> (Online; Accessed 21 Sept 2013)
- Java (2013). <http://www.java.com> (Online; Accessed 21 Sept 2013)
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L (2000) The large-scale organization of metabolic networks. *Nature* 407(6804):651–654
- Jojic N, Jojic V, Frey B, Meek C, Heckerman D (2005) Using “epitomes” to model genetic diversity: rational design of HIV vaccine cocktails. NIPS2005
- Jones NC, Pevzner P (2004) An introduction to bioinformatics algorithms. MIT Press, Cambridge
- Kazachenko V, Astashev M, Grinevich A (2007) Multifractal analysis of K⁺ channel activity. *Biochem (Moscow) Suppl Ser A Membr Cell Biol* 1(2):169–175
- Kirovski D, Heckerman D, Jojic N (2007) Combinatorics of the vaccine design problem: definition and an algorithm. Microsoft Research Technical, Report MSR-TR-2007-2148
- Kulkarni V, Rosati M, Valentin A, Ganneru B, Singh AK, Yan J, Rolland M, Alicea C, Beach RK, Zhang G-M et al (2013) HIV-1 p24gag derived conserved element DNA vaccine increases the breadth of immune response in mice. *PLoS ONE* 8(3):e60245
- Medvedev P, Georgiou K, Myers G, Brudno M (2007) Computability of models for sequence assembly. In: Algorithms in bioinformatics. Springer, Berlin, pp 289–301
- Miller CE, Tucker AW, Zemlin RA (1960) Integer programming formulation of traveling salesman problems. *J ACM (JACM)* 7(4):326–329
- Nickle DC, Rolland M, Jensen MA, Pond SLK, Deng W, Seligman M, Heckerman D, Mullins JI, Jojic N (2007) Coping with viral diversity in HIV vaccine design. *PLoS Comput Biol* 3(4):e75
- O’Neill E, Kuo LS, Krisko JF, Tomchick DR, Garcia JV, Foster JL (2006) Dynamic evolution of the human immunodeficiency virus type 1 pathogenic factor, Nef. *J Virol* 80(3):1311–1320
- Pataki G (2003) Teaching integer programming formulations using the traveling salesman problem. *SIAM Rev* 45(1):116–123
- Saksena JP (1970) Mathematical model of scheduling clients through welfare agencies. *CORS J* 8:185–200
- Schrijver A (1986) Theory of linear and integer programming., Wiley-Interscience Series in Discrete Mathematics. Wiley, Chichester
- Srivastava SS, Kumar S, Garg RC, Sen P (1969) Generalized travelling salesman problem through n sets of nodes. *CORS J* 7:97–101
- Tarhio J, Ukkonen E (1988) A greedy approximation algorithm for constructing shortest common superstrings. *Theoret Comput Sci* 57(1):131–145
- Toussaint NC, Dönnies P, Kohlbacher O (2008) A mathematical framework for the selection of an optimal set of peptides for epitope-based vaccines. *PLoS Comput Biol* 4(12):e1000246

Montevideo Units Vs Autoregressive Models on Preterm Labor Detection

Iker Malaina¹, Roberto Matorras^{2,3}, Luis Martinez¹, Carlos Bringas⁴, Larraitz Aranburu⁵, Luis Fernández-Llebrez², Ildefonso Martínez de la Fuente^{1,6}.

¹ Department of Mathematics, University of the Basque Country UPV/EHU, Bilbao, Spain
{iker.malaina, luis.martinez, mtpmadei}@ehu.eus

² Department of Obstetrics and Gynecology, Cruces University Hospital, Baracaldo, Spain
{roberto.matorras, luis.fernandez-llebrezdelrey}@osakidetza.net

³ Department of Medical Surgery Specialties, University of the Basque Country UPV/EHU, Bilbao, Spain

⁴ Department of Cellular Biology and Histology, University of the Basque Country UPV/EHU, Bilbao, Spain

cbringas001@ikasle.ehu.eus

⁵ Department of Applied Mathematics, Statistics and Operation Research, University of the Basque Country UPV/EHU, Bilbao, Spain

larraitz.aranburu@ehu.eus

⁶ Institute of Parasitology and Biomedicine “López-Neyra”, CSIC, Granada, Spain

Abstract: Consequences derived from premature labor are the leading cause of child mortality under the age of five. Although the number of studies on uterine activity has increased, the current quantitative indicators related to preterm labor detection have been unable to solve this problem. In order to estimate the immediacy of premature birth, we have analyzed 417 cardiotocograms from women suspected of threatened premature delivery with gestational ages under 37.0 weeks, subdivided into: the ones whose labor occurred after more than seven days since their visit, and the ones who delivered during those seven days. We have calculated and compared the classical uterine activity measurement, that is, the Montevideo Units, with the autoregressive parameters obtained by modeling the tocograms. It can be observed that the autoregressive approach outperformed the classical approach in all the predictive parameters, suggesting that its use could improve the current methodology on preterm birth detection.

Keywords: Montevideo Units, autoregressive models, quantitative diagnosis, preterm labor, forecast.

Corresponding author: Iker Malaina, Department of Mathematics, University of the Basque Country UPV/EHU, 48080 Bilbao (Spain), email: iker.malaina@ehu.eus.

1 Introduction

Preterm labor is a very serious complication in pregnancy that has a substantial personal and economical impact. According to the World Health Organization, the difficulties resulting from preterm birth lead to the death of one million children every year, and, only in the United States, the annual cost associated with premature labor has been reported to be of at least 26.2 billion USD [1]. It is therefore of utmost importance to be able to forecast the cases of premature delivery as soon as possible.

In order to detect threatened preterm birth, obstetrical emergency units normally carry out a vaginal fibronectine determination, a vaginal ultrasound, a systemic and obstetric examination, a blood analysis and an external cardiotocography (CTG). Although premature delivery has been associated with positive results on fibronectine test [2], short cervix length [3] and a high value on Bishop score [4], many cases still go undetected and a new approach seems necessary.

Over the last decades, measuring uterine and cervical electrical activity has become the object of many research studies due to its capacity to give a better insight into the pregnant uterus and the process of labor [5,6,7], but neither the results based on external tocodynamometry nor the ones based on uterine electromyography [8,9] have been able to provide a solution. On one hand, external tocodynamometry, which is generally used to monitor uterine activity and to determine the response to tocolytic therapy, is limited by a lack of quantitative assessment of uterine contractions and by the provision of only a single measure of global (rather than topographic) uterine pressure [10]. Indeed, this technique has been shown to have a relatively low predictive accuracy for preterm delivery [11,12,13,14]. On the other hand, several studies have attempted to correlate between uterine electromyography and term or preterm labor, but their success has been limited [15,16,17,18,19,20,21]. Nonetheless, current efforts to forecast preterm delivery have not been successful enough to have positive predictive values or sensitivities, that make the predictions clinically useful [22].

In this paper, we have used autoregressive (AR) models to estimate the immediacy of labor based on the tocograms of women with suspected threatened premature delivery. Besides extending the use of this technique to women with gestational ages comprehended between 35 and 37 weeks, we have compared this approach with the classical method to quantify uterine activity, that is, the calculation of Montevideo Units [23,24]. We have observed that AR models outperform significantly the forecasting capacity of this technique, suggesting its use on tocograms to improve the current methodology of preterm birth detection.

2 Methodology

2.1 Sample acquisition and processing

2.1.1 General considerations. In Cruces University Hospital (Basque Country), suspected threatened premature delivery (STPD) was considered when a pregnant woman with a gestational age comprehended between 24.0 and 37.0 weeks was admitted to the obstetrical emergency unit, because of any of the following causes: a) self-reported regular uterine contractions, b) intermittent abdominal pain after excluding other pathological conditions, or c) self-reported expulsion of amniotic fluid. Gestational age was established based on last menstrual period (LMP) and vaginal ultrasound. Our STPD protocol included medical history, systemic and obstetric examination, blood analysis, vaginal fibronectine determination, vaginal ultrasound and external cardiotocogram (CTG). At least 30 minutes of CTG were recorded (indistinctively by Philips Avalon FM30, Philips Avalon FM20, Hewlett Packard Vidria 50XM and Hewlett Packard 50IP cardiotocographs).

2.1.2 Our study. From the 1643 women consulting because of STPD in the study period (2010-2013), 1617 ended in premature labor (before 37.0 weeks) where 36 had a term delivery. Of those ending in preterm delivery, 423 concluded by a cesarean section.

Two specific populations were considered for this study:

- a) The delayed group, constituted by all those women whose labor occurred more than seven days after the initial consultation (n=123).
- b) The anticipated group, constituted by a subset of those women whose delivery occurred in the following seven days since their visit. This group was composed by 480 women selected by means of a simple random sampling procedure without replacement (i.e., by randomly choosing a set of unrepeated individuals from a larger population).

34 cases in the delayed group (27.6%) and 152 cases in the anticipated group (31.6%) were excluded from the study because of missing data precluding the analysis (mainly unavailable CTG, or labor date). Thus, our population was reduced to 89 cases for the delayed group, and 328 cases for the anticipated group.

Contraction parameters including contraction frequency, duration, baseline uterine tone, and relaxation time were analyzed using standard definitions [23]. The amplitude of the contractions used to calculate Montevideo units (MVUs) was determined based on the peak of the contraction [24], and this score was calculated considering the fragment with highest uterine activity of the CTG.

Cardiotocograms were scanned and later processed by Engauge Digitizer 4.0. To maintain the original proportions, the Cartesian coordinate system origin was placed

on the first square, that is, the one on the south-west of the recording, and the length of a square (which corresponds to 30 seconds of measurement) was considered the unit. Then, the data were discretized to obtain approximately 2000 values, taking one point every 0.0291457 units.

This study was approved by our center investigation board (CEIC-E16/13).

2.2 Autoregressive (AR) models.

We define a model as Autoregressive when a variable Y_t can be explained by its previous p observations, adding an error term. We describe an AR model of order p (AR(p)) as:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \dots + \varphi_p Y_{t-p} + a_t \quad (1)$$

being φ_0 a constant, φ_i (for $i=1, \dots, p$) the coefficients of the previous p variables and a_t the error term.

3 Results

In order to study the existence of significant differences on uterine activity depending on the immediacy of labor, 417 cases of suspected threatened premature delivery (STDP) with gestational ages comprehended between 24.0 and 37.0 weeks were collected and analyzed. These cases were subdivided into two groups, depending on the number of days between the STPD diagnose and the labor date: women who gave birth seven or more days after their visit (delayed group, $n=89$), and women who gave birth during the seven days that followed their visit (anticipated group, $n=328$). Then, the tocograms were transformed on time series of approximately 2000 time points by a digitizing program.

First, we estimated the Montevideo Units [24](MVU), a classical method for quantifying uterine activity. In practice, these units are calculated by adding the uterine pressure of all the contractions above baseline tone in a ten minute period. For adequate labor, more than 200 Montevideo Units are considered necessary. In our study, MVUs were calculated selecting the ten minute period of maximum activity, and a contraction was considered when the pressure of a wave increased at least 20 mmHg above baseline. The result of this calculation for the delayed group was 87.81 ± 84.49 (mean \pm SD) while for the anticipated group was 94.13 ± 95.45 . The distributions of MVU values are illustrated in Figure 1 by a box plot. In order to test if significant differences between groups existed, a test to compare the distributions of both groups was performed. To select the appropriate test, we checked if the MVU values followed a normal distribution by a Kolmogorov-Smirnov normality test. The hypothesis of normality was rejected, so we performed a non-parametric test (Wilcoxon rank sum test) to compare both groups, obtaining a p-value of 0.8035. This result implies that we cannot reject that the MVUs of the anticipated and the delayed group come from continuous distributions with equal medians. Then, we calculated the percentages of women with more than 200 MVUs, a threshold related to adequate delivery, to check

whether the proportions in both groups were significantly distinct or not. The results were 12.35% for the delayed group and 14.32% for the anticipated group, indicating that the relative number of cases ready for adequate labor according to the MVUs was very similar between both groups. Thus, MVUs were unable to discern between the cases whose labor resulted in seven days or less and the ones who delivered later.

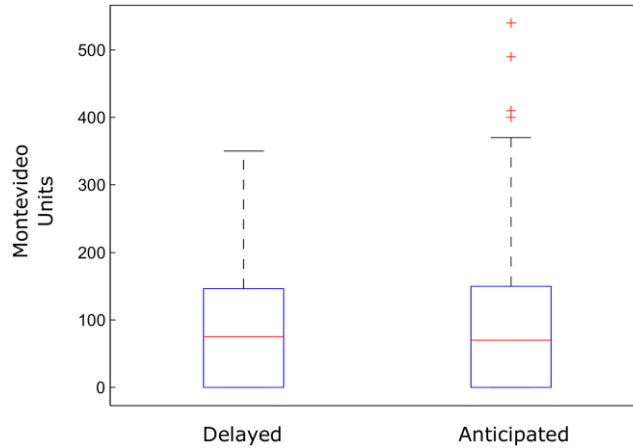


Fig 1. Box plot of the Montevideo Units for delayed and anticipated labor groups. Box plot illustration of the distributions of the MVU values calculated for the delayed and the anticipated labor groups. The blue boxes represent the distribution of the central 50% of the values and the red lines represent the medians. The rest of the values are represented by the arms, or in the case of atypical values, by red crosses. As it can be observed in the figure, there were no significant differences between the distributions of the values of the delayed and the anticipated labor groups.

Next, we calculated the autoregressive approximation. As has been shown in a previous work [25], the best model within the ARIMA family in order to model tocograms to discern between preterm cases depending on labor immediacy is the AR(2) (in fact, both φ_1 and φ_2 have been demonstrated to be significantly different between these groups, for gestational ages between 24.0 and 35.0). Thus, we estimated the first parameter φ_1 by maximum likelihood for the 417 time series obtained from the tocograms. The results for the delayed group were $\varphi_1=1.61\pm 0.16$, while for the anticipated group were $\varphi_1=1.49\pm 0.21$, with a p-value of $4\cdot 10^{-8}$ on the rank sum test. This indicates a remarkably different behavior between both groups on the autoregressive coefficient, and improves the results from [25], where the p-value was $1\cdot 10^{-5}$.

Then, we calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the autoregressive coefficient used as a predictor. To calculate this statistical measure, we had to fix a threshold to discriminate between both groups, which in the case of φ_1 was set to 1.547 (following the best accuracy

criterion, i.e., the one with the highest $\frac{\text{sensitivity}+\text{specificity}}{2}$). Considering the value of the indicator as a test to confirm preterm labor in less than seven days, a positive outcome was associated to φ_1 values above the threshold, and negative outcome was associated to values below 1.547. Thus, we determined that the sensitivity of the test ($\frac{\text{number of true positives}}{\text{number of true positives}+\text{number of false negatives}}$) was 0.764, while the specificity ($\frac{\text{number of true negatives}}{\text{number of true negatives}+\text{number of false positives}}$) was 0.585. The positive predictive value ($\frac{\text{number of true positives}}{\text{number of true positives}+\text{number of false positives}}$) was 0.33, while the negative predictive value ($\frac{\text{number of true negatives}}{\text{number of true negatives}+\text{number of false negatives}}$) was 0.901. To illustrate the effect of varying the threshold on the sensitivity and the specificity, in Figure 2, the receiver operating characteristic (ROC) curve for φ_1 is represented.

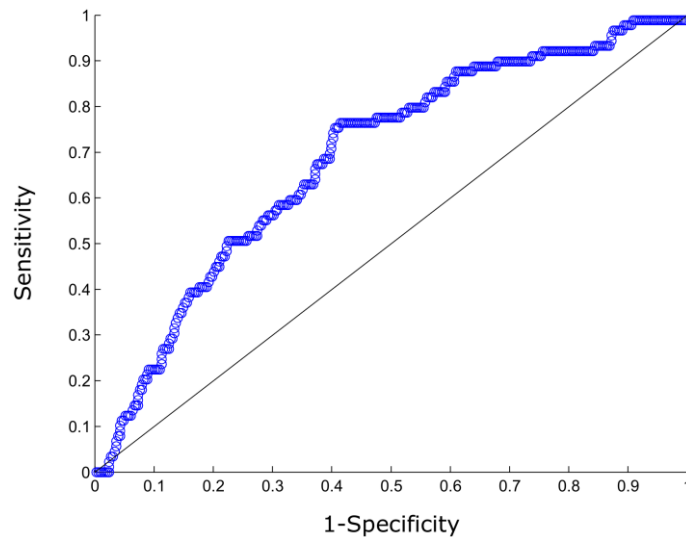


Fig 2. ROC curve of the first coefficient of the AR(2) approximation. The Y axis represents the sensitivity, while the X axis represents 1- specificity. Dots above the black line indicate good balance between sensitivity and specificity.

Finally, we calculated these predictive parameters for the Montevideo Units used as a preterm labor immediacy predictor. The threshold was set to 90, where a positive outcome was associated to values above 90, and a negative outcome was associated to values below 90. In this case, the sensitivity was 0.461, the specificity was 0.543, the PPV was 0.215, and the NPV was 0.788. These values were compared with the autoregressive predictive parameters, and it was observed that the AR parameters were on average, 0.144 ± 0.11 higher than the MVUs ones.

4 Discussion

In this paper, in order to evaluate some predictive techniques for preterm labor detection, 417 tocograms of women suspected of threatened premature delivery were analyzed. The CTGs were divided into two groups depending on the remaining time until labor, and then the Montevideo Units and the autoregressive parameters of these time series were estimated and compared.

First, we calculated the MVUs and observed that there were no significant differences between the values of women who delivered in seven days or less since their CTG recording, and the values of women who delivered later. Moreover, we observed that the number of cases related to adequate labor, that is, with more than 200 MVUs, was very similar between both groups, highlighting the inaccuracy of this method.

Second, the tocograms were modeled by autoregressive models of second order, and the parameters of both groups were compared. In this case, the delayed and the anticipated populations presented a considerably different distribution of φ_1 values, with a p-value of $4 \cdot 10^{-8}$ on the Wilcoxon rank test. This indicates that autoregressive parameters are sensitive to labor immediacy for women with gestational ages under 37.0 weeks.

Finally, we estimated the classical predictive parameters (i.e., sensitivity, specificity, positive predictive value and negative predictive value) of the AR model and the MVUs. It can be observed that the estimation of preterm delivery immediacy by the autoregressive coefficient outperformed the estimation by MVUs in all of these parameters.

In conclusion, in this article we have confirmed that the use of autoregressive coefficients to detect preterm labor [25] can be extended to women with gestational ages between 35.0 and 37.0 weeks. In addition, we have evaluated the forecasting capacity of both the autoregressive approach, and the classical measure used to quantify uterine contractions, i.e., the Montevideo Units. We have observed that the autoregressive coefficient achieved higher predictive parameters than the MVU approach, which suggests that its use on hospitals could improve preterm labor detection.

Acknowledgements. Work by the first author was supported by the Basque Government, grant PRE-2015-1-194. Authors would also like to thank the Archive Service of Cruces University Hospital for their continuous help and effort on the sample acquisition.

References

1. Behrman R, Butler A. Preterm birth. Washington, D.C.: National Academies Press; 2007.
2. Goldenberg R L, Iams J D, Mercer B M., Meis P J, Moawad A H, Copper R L, Das A, Thom E, Johnson F, McNellis D, Miodovnik M, Van Dorsten J P,

- Caritis S N, Thurnau G R, and Bottoms S F: The preterm prediction study: the value of new vs standard risk factors in predicting early and all spontaneous preterm births. NICHD MFMU Network. *American Journal of Public Health* 1998; 88 (2): 233-238.
3. Iams J D, Goldenberg R L, Meis P J, Mercer B M, Moawad A, Das A, Thom E, McNellis D, Copper R L, Johnson F, Roberts J M: The length of the cervix and the risk of spontaneous premature delivery. *New England Journal of Medicine* 1996; 334(9): 567-573.
 4. Newman R B, Goldenberg R L, Iams J D, Meis P J, Mercer B M, Moawad A H, Thom E, Miodovnik M, Caritis S N, Dombrowsky M, & Thurnau G R: Preterm prediction study: comparison of the cervical score and Bishop score for prediction of spontaneous preterm delivery. *Obstetrics and gynecology* 2008; 112(3): 508.
 5. Devedeux D, Marque C, Mansour S, Germain G, Duchane J. Uterine electromyography: A critical review. *American Journal of Obstetrics and Gynecology* 1993;169(6):1636-1653.
 6. Pajntar m, Rosl̄kar E, Rudel D. Electromyographic observations on the human cervix during labor. *American Journal of Obstetrics and Gynecology* 1987;156(3):691-697.
 7. Buhimschi C, Boyle M, Garfield R. Electrical activity of the human uterus during pregnancy as recorded from the abdominal surface. *Obstetrics & Gynecology* 1997;90(1):102-111.
 8. Rabotti C, Mischi M, van Laar J, Oei G, Bergmans J. Estimation of internal uterine pressure by joint amplitude and frequency analysis of electrohysterographic signals. *Physiol Meas* 2008;29(7):829-841.
 9. Vinken M, Rabotti C, Mischi M, Oei S. Accuracy of Frequency-Related Parameters of the Electrohysterogram for Predicting Preterm Delivery. *Obstetrical & Gynecological Survey* 2009;64(8):529-541.
 10. Hadar E, Melamed N, Aviram A, Raban O, Saltzer L, Hirsch L et al. Effect of an oxytocin receptor antagonist (atosiban) on uterine electrical activity. *American Journal of Obstetrics and Gynecology* 2013;209(4):384.e1-384.e7.
 11. Iams J, Newman R, Thom E, Goldenberg R, Mueller-Heubach E, Moawad A et al. Frequency of Uterine Contractions and the Risk of Spontaneous Preterm Delivery. *New England Journal of Medicine* 2002;346(4):250-255.
 12. Berghella, V, Iams J, Newman R, MacPherson C, Goldenberg R, Mueller-Heubach E et al. Frequency of uterine contractions in asymptomatic pregnant women with or without a short cervix on transvaginal ultrasound scan. *American Journal of Obstetrics and Gynecology* 2004;191(4):1253-1256.
 13. Iams J. What have we learned about uterine contractions and preterm birth? The HUAM prediction study. *Seminars in Perinatology* 2003;27(3):204-211.
 14. Newman R. Uterine Contraction Assessment. *Obstetrics and Gynecology Clinics of North America* 2005;32(3):341-367.
 15. Lucovnik M, Maner W, Chambliss L, Blumrick R, Balducci J, Novak-Antolic Z et al. Noninvasive uterine electromyography for prediction of preterm delivery. *American Journal of Obstetrics and Gynecology* 2011;204(3):228.e1-228.e10.
 16. Maul H, Maner W, Olson G, Saade G, Garfield R. Non-invasive

- transabdominal uterine electromyography correlates with the strength of intrauterine pressure and is predictive of labor and delivery. *J Matern Fetal Neonatal Med* 2004;15(5):297-301.
17. Garfield R, Maner W, Maul H, Saade G. Use of uterine EMG and cervical LIF in monitoring pregnant patients. *BJOG: An International Journal of Obstetrics & Gynaecology* 2005;112:103-108.
 18. Maner W. Predicting term and preterm delivery with transabdominal uterine electromyography. *Obstetrics & Gynecology* 2003;101(6):1254-1260.
 19. Marque C, Terrien J, Rihana S, Germain G. Preterm labour detection by use of a biophysical marker: the uterine electrical activity. *BMC Pregnancy Childbirth* 2007;7(Suppl 1):S5.
 20. Verdenik I, Pajntar M, LeskoÅek B. Uterine electrical activity as predictor of preterm birth in women with preterm contractions. *European Journal of Obstetrics & Gynecology and Reproductive Biology* 2001;95(2):149-153.
 21. Kandil M, Abdel-Sattar M, Abdel-Salam S, Saleh S, Khalafallah M. Abdominal electromyography may predict the response to tocolysis in preterm labor. *European Journal of Obstetrics & Gynecology and Reproductive Biology* 2012;160(1):18-21.
 22. Bennett P. Preterm labour. *Dewhurst's Textbook of Obstetrics & Gynaecology* 2007;7: 177-191.
 23. Frey H A, Tuuli M G, Roehl K A, Odibo A O, Macones G A, Cahill A G. Can contraction patterns predict neonatal outcomes?. *The Journal of Maternal-Fetal & Neonatal Medicine* 2014; 27(14): 1422-1427.
 24. Caldeyro-Barcia R, Pose S, Alvarez H. Uterine Contractility in polyhydramnios and the effects of withdrawal of the excess of amniotic fluid. *Obstetrical & Gynecological Survey* 1957;12(5):652-654.
 25. Malaina I, Matorras R, Bringas C, Aranburu L, Fernández-LLevréz L, Arana I, Gonzalez L, Martinez de la Fuente I. Precocious diagnosis of preterm labor immediacy by autoregressive integrated moving average models. *IWBBIO* 2016, ISBN: 978-84-16478-75-0.