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

Memoria para optar al grado de Doctor en Ciencias de la Salud, en la rama de Investigación Biomédica, por la Universidad del País Vasco - Euskal Herriko Unibertsitatea. Realizada bajo la codirección de:

Dr. Luís Martínez, doctor por la Facultad de Ciencia y Tecnología de la UPV/EHU, profesor titular del Departamento de Matemáticas de la UPV/EHU, especialista en biología de sistemas, álgebra y combinatoria.

&

Dr. Ildefonso Martínez de la Fuente, doctor por la Facultad de Medicina y Odontología de la UPV/EHU, científico titular del Consejo Superior de Investigaciones Científicas (CSIC), investigador académico del Departamento de Matemáticas de la UPV/EHU, pionero en biología de sistemas en el País Vasco.

Men wanted for hazardous journey.
Low wages, bitter cold, long hours of complete darkness.
Safe return doubtful. Honour and recognition in event of success.
Ernest Shackleton, 1901.

Agradecimientos

Desde hace años, mi objetivo ha sido el de hacer uso de las matemáticas para mejorar la salud de las personas y el bienestar de la sociedad. Para cumplir mi propósito elegí el camino de la investigación, lo que dio comienzo a esta tesis doctoral. La memoria que aquí presento ha sido fuente de largas jornadas de trabajo, intensas lecciones y un continuo esfuerzo por perfeccionar mis resultados investigadores. Sin embargo, también me ha hecho sentirme orgulloso de mí mismo y ha servido para iniciarme como científico, permitiéndome así avanzar en pos de mi ideal.

Sin duda, esta andadura no podría haberla hecho solo. Muchas personas han estado ahí para ayudarme en los momentos más difíciles, y es por eso que quiero dedicarles la presente tesis y este sincero agradecimiento.

A Luis e Ildefonso, mis directores, por guiarme en este camino, por anteponer mi formación y mi futuro a todo lo demás. Si tuviera que volver a realizar esta tarea, los elegiría otra vez sin dudarlo.

A Carlos, por compartir la carga de mi tesis. No puedo imaginar lo duro que habría sido este trabajo sin su apoyo.

A Roberto, Jesús, Larraitz, Leire, Carmen, y muchos otros grandísimos científicos que he tenido la suerte de conocer. Porque sin sus enseñanzas no habría podido llegar tan lejos.

A Lola, por creer en mí desde el principio, y darme la oportunidad de comenzar esta andadura.

A Aingeru, por ser siempre un referente y ayudarme cuando lo he necesitado.

A mis amigos Amaia, Ander, Campillo, David, Erik, Idoia, Inazio, López, Mikel, Miren, Toño y Tundi, por infundirme la fuerza necesaria para llegar hasta el final.

Por último quiero dar las gracias a mi familia, la pieza más importante. Su apoyo incondicional es la piedra angular de esta tesis. No puedo describir con palabras la trascendencia de su papel. El que se sientan orgullosos de mí hace que todo esto haya valido la pena.

Durante la realización de esta tesis, he contado con el apoyo económico de las siguientes fuentes:

- Programa predoctoral del Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco. Ref: PRE-2015-1-194.

- Proyecto de investigación "Groups, topology and applications" financiado por el Gobierno Vasco. Ref: IT974-16.

- Proyecto de Innovación Educativa "El proyecto de los tres pilares con propósito, aprendizaje por indagación en materias relacionadas con el Álgebra" financiado por la UPV/EHU.

Además, quiero agradecer tanto el apoyo técnico como humano del servicio IZO-SGIker (UPV/EHU, MICINN, GV/EJ, ESF).

Tome I

Contents of Tome I

Abstract

There is nothing more important than preserving life, and the thesis here presented is framed in the field of quantitative biomedicine (or systems biomedicine), which has as objective the application of physico-mathematical techniques in biomedical research in order to enhance the understanding of life's basis and its pathologies, and, ultimately, to defend human health.

In this thesis, we have applied physico-mathematical methods in the three fundamental levels of Biomedical Research: basic, translational and clinical.

At a basic level, since all pathologies have their basis in the cell, we have performed two studies to deepen in the understanding of the cellular metabolic functionality. In the first work, we have quantitatively analyzed for the first time calcium-dependent chloride currents inside the cell, which has revealed the existence of a dynamical structure characterized by highly organized data sequences, non-trivial long-term correlation that last in average 7.66 seconds, and "crossover" effect with transitions between persistent and anti-persistent behaviors.

In the second investigation, by the use of delay differential equations, we have modeled the adenylate energy system, which is the principal source of cellular energy. This study has shown that the cellular energy charge is determined by an oscillatory non-stationary invariant function, bounded from 0.7 to 0.95.

At a translational level, we have developed a new method for vaccine design that, besides obtaining high coverages, is capable of giving protection against viruses with high mutability rates such as HIV, HCV or Influenza.

Finally, at a clinical level, first we have proven that the classic quantitative measure of uterine contractions (Montevideo Units) is incapable of predicting preterm labor immediacy. Then, by applying autoregressive techniques, we have designed a novel tool for premature delivery forecasting, based only in 30 minutes of uterine dynamics.

Altogether, these investigations have originated four scientific publications, and as far as we know, our work is the first European thesis which integrates in the same framework the application of mathematical knowledge to biomedical fields in the three main stages of Biomedical Research: basic, translational and clinical.

Resumen de la tesis

Las ciencias cuantitativas son esenciales para el desarrollo de la biomedicina. Los métodos físico-matemáticos han probado ser no sólo útiles, sino necesarios para avanzar en muchos campos de las Ciencias de la Vida, como por ejemplo la compresión de las dinámicas moleculares celulares, el funcionamiento de las redes neuronales, la secuenciación genética y la simulación de muchos procesos fisiológicos humanos.

Se entienden por ciencias cuantitativas aquellas que basan sus análisis en técnicas numéricas y métodos matemáticos. Éstas abarcan un amplísimo abanico de disciplinas, como por ejemplo: geometría fractal, teoría de sistemas, bioinformática, mecánica estadística, cálculo diferencial, álgebra computacional, inteligencia artificial, teoría de la información o biología de sistemas.

Las herramientas cuantitativas se han introducido progresivamente en todos los campos fundamentales de la medicina, volviéndose esenciales para las Ciencias de la Vida; como ejemplo, basta decir que recientemente se han utilizado para: modelar el sistema circulatorio y así entender determinadas enfermedades (Müller & Toro, 2014) en *cardiología*; probar la eficacia de nuevos fármacos contra la soriasis (Papp et al., 2013) en *dermatología*; descubrir nuevas formas de reducir la transmisión de la malaria (Govella et al., 2010) en *epidemiología*; estimar la respuesta a la medicación contra el VIH (Xiao et al., 2013) en *inmunología*; estudiar las similitudes entre diferentes secuencias genéticas del virus Zika (Wang et al., 2016) en *genética médica*; predecir la velocidad del crecimiento bacteriano en un huésped (Huang, 2013) en *microbiología*; determinar los factores de riesgo del Alzheimer (Norton et al., 2014) en *neurología*; discriminar entre fetos sanos y enfermos (Splika et al., 2014) en *obstetricia*; o estudiar algunas características de la metástasis del cáncer de páncreas (Yachida et al., 2010) en *oncología*.

Las ciencias cuantitativas también han mostrado jugar un papel fundamental en los mayores proyectos dedicados a las Ciencias de la Vida. Por ejemplo, el proyecto Human Genome Project (1990-2003) contó con una amplísima colaboración internacional multidisciplinar, surgiendo de él la Bioinformática como herramienta clave para la secuenciación genética.

Inspirado por el éxito del HGP, se propuso como gran reto de la ciencia del siglo XXI el profundizar en la comprensión del cerebro, lo que ha supuesto una competición entre Europa y Estados Unidos por conseguir dicho objetivo.

Así, en 2005 Suiza lanzó el proyecto Blue Brain con la finalidad de conseguir simular ciertas regiones del cerebro de los mamíferos, y poder estudiar su funcionamiento y el efecto de posibles patologías. Ese proyecto sirvió de base para que en Octubre de 2013 se iniciara a nivel europeo el Human Brain Project (HBP), que tiene como meta uno de los desafíos científicos más complejos de la ciencia actual: conseguir una simulación computarizada completa del cerebro humano. El HBP es un claro ejemplo de la necesidad de multidisciplinariedad, ya que necesita tanto de expertos en biomedicina que aporten su conocimiento del funcionamiento del cerebro, como de especialistas en ciencias cuantitativas que puedan modelizarlo y computarizarlo.

Por otro lado, en 2009 Estados Unidos impulsó el Human Connectome Project (HCP), con el fin de construir la red completa de las conexiones neuronales del cerebro humano. En Abril de 2013, como paso siguiente y necesario al HCP, se inició otro proyecto en el que la sinergia entre las ciencias cuantitativas y la biomedicina es fundamental: el proyecto BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies). Esta propuesta, liderada por Estados Unidos, tiene como propósito dotar a los científicos de herramientas para obtener una imagen dinámica del cerebro en acción, lo que permitiría, entre otras cosas, comprender cómo aprendemos y memorizamos, o entender el funcionamiento de enfermedades como el Alzheimer o el Parkinson.

La tesis que aquí se presenta se enmarca, como hemos dicho, en el campo de la biomedicina cuantitativa (o biomedicina de sistemas), y en ella, nosotros hemos desarrollado una actividad investigadora en la que se aplican diversas técnicas físico-matemáticas a los tres niveles fundamentales de la Investigación Biomédica: básica, traslacional y clínica.

A nivel básico, dado que todas las patologías tienen como base la célula, hemos realizado dos estudios sobre la funcionalidad metabólica celular, es decir, sobre la regulación de los procesos enzimáticos celulares.

El primer trabajo se ha centrado en estudiar las propiedades dinámicas de las corrientes de cloro calcio-dependientes. Como es sabido, el Cl^- es el anión permeable más abundante dentro de la célula, y participa en una gran variedad de procesos fisiológicos, estando implicado en diversas enfermedades humanas (pag. 16). En nuestra investigación, las corrientes de cloro calcio-dependientes fueron obtenidas en ovocitos de *Xenopus laevis* por medio de técnicas electro-fisiológicas y posteriormente fueron analizadas a través de técnicas de la mecánica estadística, tales como el root-mean square fluctuation, bridge detrended Scaled Window Variance, Dispersion Analysis o Detrended Fluctuation Analysis.

Entre los principales resultados obtenidos en este trabajo, cabe resaltar los siguientes:

- Las corrientes de cloro calcio-dependientes presentaron correlaciones no triviales positivas a largo plazo de duraciones comprendidas entre 3 y 13 segundos (con una media de 7.66 segundos), que abarcaron entre 1,500 y 6,500 valores temporales.

- Dichas corrientes mostraron transiciones de correlaciones positivas a negativas. Las correlaciones positivas se dieron en intervalos temporales cortos, mientras que las negativas se dieron en intervalos más largos.

- Las series de Cl^- se categorizaron como fractional Brownian motion (fBm).

- Todas las corrientes de cloro calcio-dependientes presentaron memoria anti-persistente a largo plazo en intervalos largos.

El resultado fundamental del estudio es que las concentraciones de cloro calcio-dependientes presentan una estructura dinámica compleja caracterizada por: secuencias de datos altamente organizadas, correlaciones no triviales a largo plazo de 7.66 segundos de duración media, y efecto crossover con transiciones entre comportamientos persistentes y anti-persistentes.

Este trabajo ha sido publicado en Nature Scientific Reports: *Dynamic properties of calcium-activated chloride currents in Xenopus laevis oocytes (2017). Nature Scientific Reports, 7, 41791; por: De la Fuente, I. M., Malaina, I., Pérez-Samartín, A., Boyano, M. D., Pérez-Yarza, G., Bringas, C., Villarroel, A., Fedetz, M., Arellano, R., Cortés, J.M. & Martínez, L.*

La segunda investigación básica se ha enfocado en comprender algunos de los elementos que regulan el estado energético celular. Para ello, hemos modelizado el sistema del adenilato, que es la principal fuente de energía de la célula. Concretamente, a partir de datos experimentales, hemos desarrollado un modelo matemático basado en ecuaciones diferenciales con retardo que es capaz de representar diferentes escalas temporales en las que ocurren las principales reacciones metabólicas que involucran al ATP/ADP/AMP. Además, hemos utilizado este modelo para analizar las fluctuaciones del ATP, el ADP, el AMP, la energía libre de Gibbs de la hidrólisis del ATP y las fluctuaciones del total de nucleótidos de adenina. Por último, a través del estudio de los niveles de carga energética del adenilato (AEC), hemos estimado el nivel crítico de ATP a partir del cual el sistema del adenilato colapsa, lo que conlleva la muerte celular.

El resultado fundamental del estudio mostró que, en condiciones celulares normales la carga energética celular (esto es, la relación entre ATP, ADP y AMP) está determinada por una función oscilante no estacionaria, acotada entre 0.7 y 0.95.

Este trabajo ha sido publicado en PloS one: *On the dynamics of the adenylate energy system: homeorhesis vs homeostasis (2014). PloS one, 9(10), e108676; por: De la Fuente, I. M., Cortés, J. M., Valero, E., Desroches, M., Rodrigues, S., Malaina, I., & Martínez, L.*

A nivel traslacional, el propósito ha sido el de desarrollar un nuevo método de diseño de vacunas capaz de combatir la mutabilidad viral. Las técnicas de vacunación tradicional se han mostrado incapaces de atajar de forma eficiente los virus con alta tasa de mutabilidad como el VIH, VHC o Influenza. Además, los algoritmos de diseño de vacunas actuales se centran principalmente en maximizar el número de epítopos cubiertos, lo que lleva a no considerar los epítopos menos frecuentes, y por ende no ofrecer protección contra muchas variantes de los virus. Basándonos en el concepto

combinatorio de λ-supercadena introducido por nuestro grupo, hemos desarrollado un método de diseño de vacunas peptídicas capaz de ofrecer una protección equilibrada contra todas las variantes del virus consideradas. El problema combinatorio que sustenta nuestro criterio de λ-supercadena se ha resuelto tanto de forma exacta como de forma aproximada, a través de algoritmos basados en programación entera y en hill-climbing, respectivamente. Dichas técnicas han sido utilizadas posteriormente para obtener potenciales candidatos a vacuna contra la Influenza y contra el VIH. Finalmente, hemos comparado nuestros resultados con los obtenidos por otros métodos de diseño computacional de vacunas enfocados en maximizar el recubrimiento. Asimismo, hemos estimado los máximos valores de recubrimiento alcanzables, y los hemos contrastado con los valores obtenidos por nuestras técnicas.

Los resultados de este estudio mostraron que:

- El algoritmo basado en programación entera obtuvo mejores resultados que el método hill-climbing. Sin embargo, su coste computacional es mucho mayor.

- A pesar de no ser su objetivo principal, el nivel de recubrimiento obtenido por nuestras técnicas (62% para Nef y 82% para Gag) estuvo próximo al máximo posible (67.8% y 85.4%, respectivamente).

- Nuestros algoritmos alcanzaron el mismo recubrimiento (62% para Nef y 82% para Gag) que los algoritmos diseñados para optimizar dicha cantidad, cumpliendo al mismo tiempo el criterio de ser λ-supercadenas.

El resultado fundamental de la investigación es que el enfoque de las λ-supercadenas mejora sustancialmente los métodos de diseño computacional de vacunas, ya que en las pruebas *in silico*, además de obtener un nivel de recubrimiento tan alto como los algoritmos que maximizan dicha cantidad, ofreció un equilibrio adecuado entre los epítopos de todas las variantes de virus seleccionadas. Por lo tanto, este método abre la posibilidad de diseñar vacunas contra virus con alta tasa de mutabilidad como el VIH, VHC o Influenza, que actualmente no disponen de una vacuna eficiente.

Esta investigación ha sido publicada en Journal of mathematical biology: *A combinatorial approach to the design of vaccines (2015). Journal of mathematical biology, 70(6), 1327-1358; por: Martínez, L., Milanič, M., Legarreta, L., Medvedev, P., Malaina, I., & De la Fuente, I. M.*

A nivel clínico, nuestra meta ha sido obtener una herramienta capaz de predecir el parto pre-término a través del uso de técnicas estadísticas. Para ello, hemos revisado más de 400 historias clínicas de mujeres que acudieron al hospital por amenaza de parto pre-término y hemos estudiado los 30 primeros minutos de su dinámica uterina. Con el objeto de anticiparnos con suficiente antelación al problema, hemos dividido el grupo de pacientes en dos poblaciones: las mujeres que dieron a luz en menos de una semana desde su visita, y las que parieron más adelante. Primero, hemos evaluado si el método cuantitativo clásico de análisis de las contracciones (las Unidades de Montevideo) es capaz de predecir el parto prematuro. Después, hemos modelizado las series temporales de presión uterina a través de técnicas autorregresivas, y hemos estudiado si existen

diferencias significativas en los coeficientes de dichos modelos en función de la inmediatez del parto.

Destacamos de entre los resultados más importantes de esta investigación los siguientes:

- Las Unidades de Montevideo no mostraron ser significativamente distintas entre las mujeres que dieron a luz en menos de una semana desde su visita, y las que dieron a luz más tarde.

- Al utilizarse para indicar qué mujeres estaban listas para el parto (esto es, cuáles tenían más de 200 Unidades de Montevideo), se observó que el porcentaje era similar en ambos grupos (el 14% de las que parieron en menos de una semana y el 12% para el otro grupo).

- Al modelar las tocografías con técnicas autorregresivas, encontramos diferencias significativas en los coeficientes de dichos modelos dependiendo de la inmediatez del parto.

- La media de sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo del modelo autorregresivo superó en 0.144 a la media correspondiente para las Unidades de Montevideo cuando se utilizaron para predecir el parto pre-término.

El resultado fundamental del estudio es que, a diferencia de las Unidades de Montevideo, los modelos autorregresivos aplicados a las dinámicas uterinas son sensibles a la inmediatez del parto prematuro. Por ende, esta herramienta podría utilizarse para ayudar a la prognosis y detección del parto pre-término.

Este trabajo ha sido publicado en: *Montevideo Units Vs Autoregressive Models on Preterm Labor Detection (2016). ITISE Proceedings 2016, 799-807; por: Malaina, I., Matorras, R., Martínez, L., Fernandez-LLebrez, L., Bringas, C., Aranburu, L. & De La Fuente, I. M.*

Además de las investigaciones anteriormente señaladas, en la tesis se incluye un Anexo en el cual se resumen brevemente algunos de los aspectos más relevantes de la aplicación de técnicas físico-matemáticas a las Ciencias de la Vida a lo largo de la historia.

Por último quisiera señalar que, debido a la extensión de esta tesis, el presente trabajo ha sido dividido en dos tomos. En el Tomo I, describimos los aspectos más relevantes de nuestras investigaciones, y al final de ellas he añadido un anexo donde se incluye un breve resumen histórico de la aplicación de técnicas físico-matemáticas a las Ciencias de la Vida. En el Tomo II se presentan, de forma íntegra, los cuatro artículos publicados en los que se ha basado ésta tesis, de forma que algunos gráficos, tablas, métodos y explicaciones no incluidos en el Tomo I sirvan de complemento a la tesis que se presenta.

Dada la importancia de las ciencias cuantitativas, concluimos esta tesis con una breve reflexión sobre la necesidad de que en una gran parte de los grupos de investigación biomédica, debería haber al menos un experto en técnicas cuantitativas. No es esencial que los especialistas médicos conozcan las herramientas físico-matemáticas. El que las conozcan es deseable, pero no es imprescindible. Lo necesario es que en los grupos de investigación haya al menos una persona capaz de entender y trasladar los problemas médicos al lenguaje físico-matemático, y eso sólo puede conseguirse con expertos en ciencias cuantitativas biomédicas.

La unión entre las ciencias cuantitativas y la medicina permitirá un gran avance en lo que es más importante y esencial: la defensa y el cuidado de la salud humana.

Introduction

Quantitative sciences are essential for the development of biomedicine. Physical-mathematical methods have proven to be not only useful, but also necessary to advance in many fields of Life Sciences such as the comprehension of molecular cellular dynamics, the functioning of neural networks, genetic sequencing and the simulation of several human physiological processes.

The term "quantitative sciences" covers all those disciplines which base their analyses in numerical techniques and mathematical methods. They encompass a huge amount of sciences, as for instance: fractal geometry, systems theory, bioinformatics, statistical mechanics, differential calculus, computational algebra, artificial intelligence, information theory or systems biology.

Quantitative tools have entered progressively into all the fundamental fields of medicine, becoming essential for Life Sciences. As an example, these methods have been recently used to: design models of the circulatory system which can be used to study many cardiovascular diseases (Müller & Toro, 2014) in *cardiology*; test the effectiveness of new drugs against psoriasis (Papp et al., 2013) in *dermatology*; find new ways to reduce the transmission of malaria (Govella et al., 2010) in *epidemiology*; foretell the response to a human immunodeficiency virus treatment (Xiao et al., 2013) in *immunology*; study the similarities between different genetic sequences of the virus Zika (Wang et al., 2016) in *medical genetics*; predict the speed of bacterial growth in a host (Huang, 2013) in *microbiology*; determine several risk factors of Alzheimer disease (Norton et al., 2014) in *neurology*; discriminate between healthy and ill fetuses (Splika et al., 2014) in *obstetrics*; or estimate the metastasis-free period in pancreatic cancer (Yachida et al., 2010) in *oncology*.

The role of quantitative sciences has also proven to be fundamental in the biggest projects of humanity dedicated to Life Sciences. For instance, through the Human Genome Project (1990-2003), which included an extensive international multidisciplinary collaboration, Bioinformatics emerged as a key tool for genetic sequencing.

The success of HGP inspired the proposal of deepening in the comprehension of the brain as one of the big challenges of the 21st Century, which has lead to a competition between Europe and the United States to accomplish this goal first.

Thus, in 2005, Switzerland launched the Blue Brain project, with the purpose of obtaining an accurate simulation of several regions of mammals' brains, therefore enabling the study of its behavior and the effect of possible pathologies. This research laid the foundations for the initiation of a new project at European level, namely, the Human Brain Project (HBP), which started in October 2013. The main objective of the HBP is one of the most complex challenges of current science, which is to achieve a full computerized simulation of the human brain. This project stands as an example of the necessity of multidisciplinarity, since it demands both experts in biomedicine to provide their knowledge of the performance of the brain, and specialists in quantitative sciences to model and computerize it.

On the other hand, in 2009 the United States boosted the Human Connectome Project (HCP) with the purpose of building a complete map of human brain's neural networks. In April 2013, and as a necessary next step of the HCP, another project in which the synergy between quantitative sciences and biomedicine is fundamental emerged: the BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies). This project is lead by the United States, and its purpose is to endow the scientists with tools that allow obtaining a dynamical image of the brain in action, which in turn will lead to, for example, the understanding of how we learn and memorize, or to comprehend the mechanism of diseases such as Alzheimer or Parkinson.

The thesis here presented is framed, as we said, within the quantitative biomedicine (or systems biomedicine) field, and in it, we have performed an exercise of investigation in which diverse physico-mathematical techniques have been applied in the three fundamental levels of Biomedical Research: basic, translational and clinical.

At a basic level, since all pathologies have as basis the cell, we have performed two studies about the cellular metabolic functionality, i.e., regarding the regulation of cellular enzymatic processes.

The first work was centered on studying the dynamical properties of calcium-dependent chloride currents in the cell. As it is known, chloride (Cl^-) is thought to be the most abundant permeable anion in the cell; it participates in a wide variety of important local and systemic physiological processes, being involved in a variety of human diseases (pg. 16). In our research, calcium-dependent chloride currents were obtained in *Xenopus Laevis* oocytes through electro-physiological techniques, and subsequently analyzed by techniques belonging to statistical mechanics, such as root mean square fluctuation, bridge detrended Scaled Window Variance, Dispersion Analysis or Detrended Fluctuation Analysis. This study has been published in Nature Scientific Reports: Dynamic properties of calcium-activated chloride currents in Xenopus laevis oocytes (2017). *Nature Scientific Reports*, *7*, 41791; by: De la Fuente, I. M., Malaina, I., Pérez-Samartín, A., Boyano, M. D., Pérez-Yarza, G., Bringas, C., Villarroel, A., Fedetz, M., Arellano, R., Cortés, J.M. & Martínez, L.

The second work has focused on understanding some key elements that regulate the cellular energetic status. For that purpose, we have modeled the adenylate energy system, which is the main source of cellular energy. Specifically, by the use of experimental data, we have developed a mathematical model based on delay differential equations. These equations allow us to represent the different time scales in which the

major metabolic reactions involving ATP/ADP/AMP occur. Additionally, we have applied this model to analyze the fluctuations of ATP, ADP, AMP, Gibbs free energy charge for ATP hydrolysis and total adenine nucleotide pool. Finally, by studying the levels of AEC (adenylate energy charge), we have estimated the critical level of ATP from which the adenlylate system collapses, which leads to the dead of the cell. This study has been published in PloS one: On the dynamics of the adenylate energy system: homeorhesis vs homeostasis (2014). *PloS one*, *9*(10), e108676; by: De la Fuente, I. M., Cortés, J. M., Valero, E., Desroches, M., Rodrigues, S., Malaina, I., & Martínez, L.

At a translational level, the purpose of this thesis has been to develop a new method for vaccine design, capable of fighting the viral mutability. Traditional vaccination methods have been proven incapable of offering enough protection against viruses with high mutability rate, such as HIV, HCV or Influenza. Furthermore, current vaccine design algorithms focus mainly in maximizing the number of covered epitopes, which leads to leaving aside the less frequent ones, and therefore not to give protection against many variants of those viruses. Based on the combinatorial concept of λ-superstring introduced by our group, we developed a vaccine design method that, in addition to covering as many epitopes as the algorithms focused on maximizing that amount, offers a balanced protection against all the considered virus variants. The combinatorial problem which integrates our criterion has been solved in both exact and approximate ways, through an integer programming algorithm and a hill-climbing technique, respectively. Then, we have applied these techniques to obtain potential vaccine candidates against Influenza and HIV. Afterwards, we have compared our results with the ones obtained by other computational vaccine design methods focused on maximizing the coverage. Finally, we have estimated the highest achievable coverage values, and contrasted them with the values obtained by our algorithms. This work has been published in Journal of Mathematical Biology: A combinatorial approach to the design of vaccines (2015). *Journal of mathematical biology*, *70*(6), 1327-1358; by: Martínez, L., Milanič, M., Legarreta, L., Medvedev, P., Malaina, I., & De la Fuente, I. M.

At a clinical level, the goal of this thesis has been to obtain a tool capable of predicting preterm delivery, by means of statistical techniques. For that purpose, we have revised more than 400 medical records of women admitted to the hospital because of suspected threatened premature delivery and we have studied their uterine dynamics. In order to detect the problem with enough anticipation, we have divided the group of patients in two populations: the first one, constituted by those women who gave birth a week or less since their visit to the obstetrical emergency unit, and the second, comprised by those who delivered later. On one hand, we have analyzed if the classic quantitative method of contraction analysis (Montevideo Units) is capable of predicting premature labor. On the other hand, we have modeled the uterine pressure time series by means of autoregressive techniques, and then we have studied if there were significant differences between the coefficients of those models depending on the labor immediacy. This work has been published in: Montevideo Units Vs Autoregressive Models on Preterm Labor Detection (2016). *ITISE Proceedings* 2016, 799-807; by: Malaina, I., Matorras, R., Martínez, L., Fernandez-Llebrez, L., Bringas, C., Aranburu, L. & De La Fuente, I. M.

Besides the abovementioned researches, we include an Annex in which we briefly summarize some of the most relevant aspects of the application of physico-mathematical techniques to Life Sciences through history.

Finally, we would like to note that, given the extension of this work, the current thesis has been divided in two tomes. In Tome I, we describe the most relevant aspects of our investigations, and in the end of it we have added an Annex where we give a brief summary of the application of mathematical methods to biology and medicine through history. In Tome II, we gather, as they were published, the four papers in which this thesis has been based, so some graphs, tables, references and explanations that are not included in Tome I serve as complement for the thesis here presented.

As a result, we present a thesis integrating investigations in the three main levels of Biomedical Research, illustrating the power and utility of quantitative methods in modern Life Sciences.

Basic biomedical quantitative investigation

Research n°1: Intracellular dynamics of calcium-dependent chloride currents

1.1. Main objective

To study the dynamic properties of the chloride currents belonging to calcium-activated chloride channels (CaCCs) of *Xenopus laevis* oocytes.

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.

This study has been published in Nature Scientific Reports: Dynamic properties of calcium-activated chloride currents in Xenopus laevis oocytes (2017). *Nature Scientific Reports*, *7*, 41791; by: De la Fuente, I. M., Malaina, I., Pérez-Samartín, A., Boyano, M. D., Pérez-Yarza, G., Bringas, C., Villarroel, A., Fedetz, M., Arellano, R., Cortés, J.M. & Martínez, L.

1.2. Importance

Chloride (Cl^-) 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.

1.3. Brief background

Chloride (Cl^-) is thought to be the most abundant free anion in the cell (Huang, 2012), 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 (Jentsch & Günther, 1997; Jentsch et al., 2005).

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 vesicles (Jentsch et al., 2002; Nilius & Droogmans, 2003; O'Rourke, 2007; Stauber & Jentsch, 2007). 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 (Jentsch et al., 2002; Tang & Chen, 2011). In addition, different studies have recognized the Cl^- channels' contributions to apoptosis (Okada et al., 2006), signal transduction (Gonzalez-Silva et al., 2013), cell cycle (Mao et al., 2009), cell adhesion and motility (Kim et al., 2004), among other complex cellular processes.

Intracellular chloride currents also play important roles in a variety of physiological processes (Berg et al., 2012), including epithelial secretion (Frizzell & Hanrahan, 2012), neuronal excitability (Voglis & Tavernarakis, 2006), repolarization of the cardiac action potential (Duan, 2013), modulation of light responses (Endeman et al., 2012) and olfactory transduction (Pifferi et al., 2012).

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 (Huang et al., 2012; Jentsch, 2008; Planells-Cases & Jentsch, 2009).

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 (Li et al., 2009; Peretti et al., 2015).

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 (Jentsch et al., 2002; Suzuki et al., 2006). 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 channels (Verkman & Galietta, 2009).

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 levels (Dickson et al., 2014). These channels are ubiquitously expressed, in both excitable and non-excitable cells (Hartzell et al., 2005).

The relationship between chloride currents and intracellular calcium fluctuations gives CaCCs a crucial role in many cellular processes, and many studies show the great importance and broad physiological role of these channels (Hoffmann et al., 2014).

Historically, chloride channels have been less studied than cation channels. Considerable progress has been made in the knowledge of their molecular structures and functions (Dickson et al., 2014), but there seems to be practically no quantitative studies of the dynamics of chloride currents.

1.4. Calcium-activated chloride currents

In order to analyze 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 (numbered from n1 to n21), each one of them formed by 130,000 discrete data points, with a sampling interval of 2 milliseconds. Figure 1.1 shows three representative experimental signals obtained by means of the patch-clamp technique (voltage clamped at -60 mV), under three different pH conditions, Ringer's solution at pH 5.0, 7.0 and 9.0 (acid, neutral and basic pH).

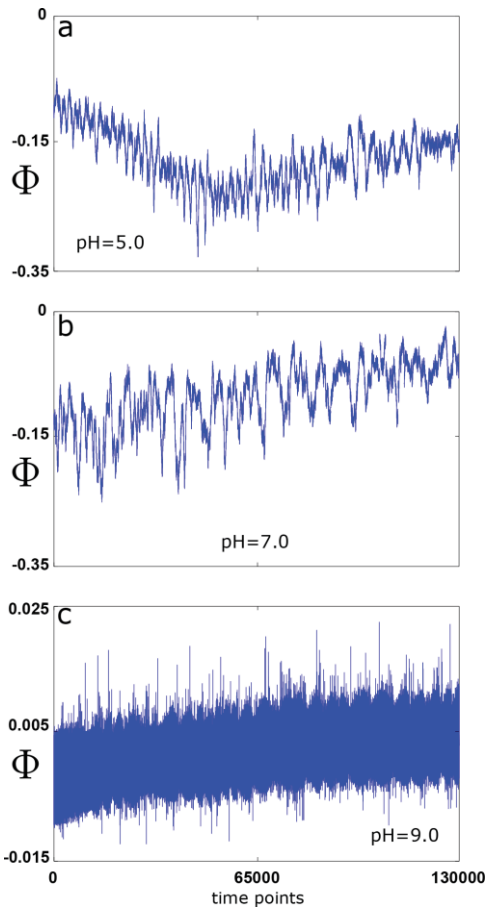

Figure 1.1. Calcium-activated chloride currents in *Xenopus laevis* oocyte. Three prototype experimental Cl^- currents obtained from the same cell at different conditions: (**a**) pH 5.0 (n10), (**b**) pH 7.0 (n11), (**c**) pH 9.0 (n12).

To confirm that oscillations monitored in *Xenopus* oocytes by application of Fetal bovine serum (FBS) 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 is illustrated in Fig.1.2a, 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.1.2b. 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 was 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 (Fig.1.2c).

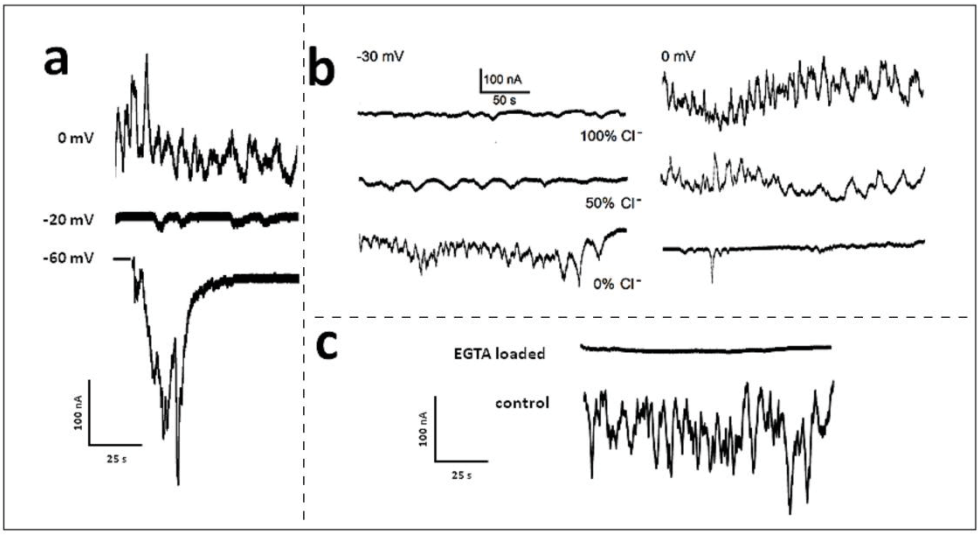

Figure 1.2. Ca^{2+}-dependent Cl^- current validation. (**a**) *Xenopus* oocyte held at either −60, −40, −20 or 0 mV. Reversal potential of oscillatory currents corresponded to a value close to −23 mV. (**b**) Oscillatory current reversal potential were dependent on external Cl^- concentration; traces 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.

1.5. Results

First, to test for the presence of long-term correlations in the experimental chloride data we used the root-mean square (rms) fluctuation *F(l)* (Stanley et al., 1996). 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 long-term anti-correlations. According to this method, we 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 $l=1,...,l_{max}$ (see Tome II 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 of the fit.

Second, in order to discern whether the experimental Cl^- currents exhibit non-trivial correlations, we 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±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±1,298.8), which correspond to periods of time ranging between 3 and 13 seconds (mean of 7.66±2.6). Boundary times were achieved on the series n17 (experiment 6, pH=7.0) and n2 (experiment 1, pH=7.0), respectively.

Figure 1.3 shows an example of rms fluctuation analysis applied to three C^{2+}-activated Cl^- responses of the same oocyte (n1, n2 and n3 time series belonging to experiment 1) for their correlation durations. 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.

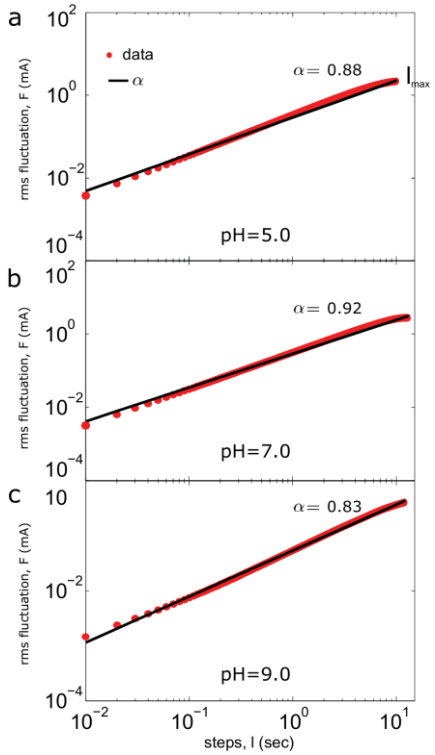

Figure 1.3. Root mean square fluctuation analysis applied to experiment 1 on a single window. Log-log plot of the rms fluctuation F versus l step. The red points depict the results of the original data for each value of l, while the black lines represent the regression lines. (**a**) $\alpha = 0.88$ (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.

Next, we 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, l_{max} = 1,200). The global average $\bar{\alpha}$ was 0.774±0.108. It can be observed that the values of α decrease slowly as l_{max} increases (Fig. 1.4a).

In addition, we observed a critical transition around l_{max} = 28 seconds, where the behavior of the Cl^- currents changes from positive to negative correlations (Fig. 1.4b). 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±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.

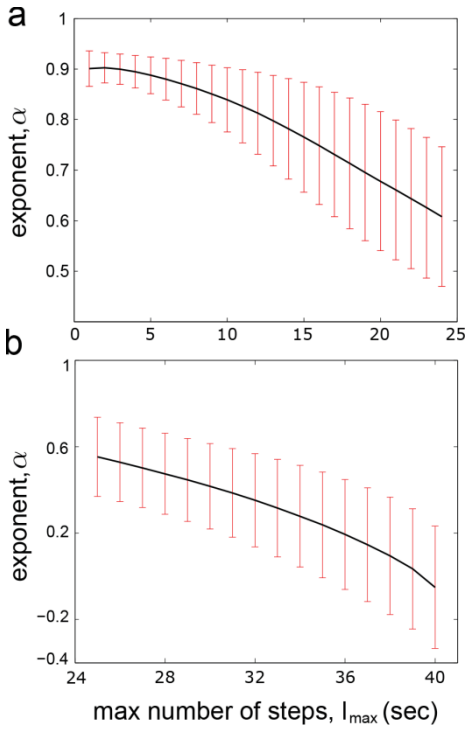

Figure 1.4. Long-term correlations across different windows lengths. (**a**) Global average $\bar{\alpha}$ versus different values of l_{max} (varying from 1 to 24 seconds). (**b**) $\bar{\alpha}$ as a function of l_{max} (varying from 25 to 40 seconds). The error bars represent the standard deviation at each step. It can be observed that all Cl^- time series change from positive to negative correlation near l_{max} = 28 seconds.

Moreover, we 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 plot (Eke et al., 2000). 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.

Next, we checked whether the chloride time series show persistent or anti-persistent long-term memory by calculating the Hurst exponent (Hurst, 1951). Although several tools exist for estimating the long-term memory from fBm time series, one of

the most reliable methods is the bridge detrended Scaled Windowed Variance analysis (bdSWV) (Cannon et al., 1997) (see Tome II 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.

In order to estimate the significance of our results, we performed a shuffling procedure that defines the null-hypothesis. If the original time series exhibits a memory structure (H≠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 were completely lost. Notice that after shuffling the series became Gaussian white noise (fGn series with $\overline{\beta} = -0.0006 \pm 0.004$), and for this case the use of bdSWV is not justified. Instead, Dispersion Analysis is the most recommendable tool for this kind of series (Eke et al., 2000; Caccia et al., 1997).

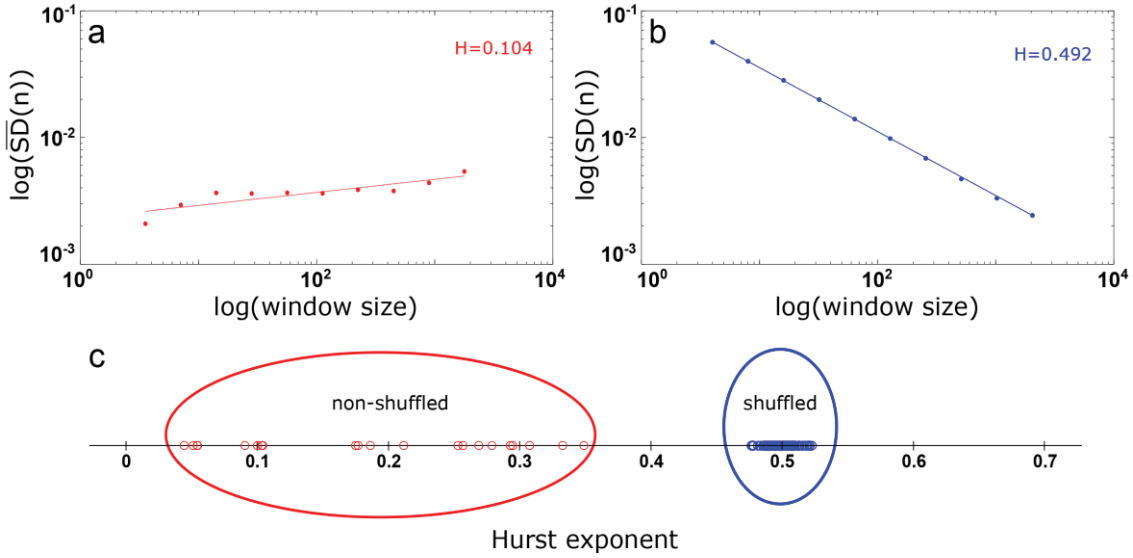

Figure 1.5. Hurst exponents obtained by the bdSWV analysis. (**a**) The slope of a log-log plot of the $\overline{SD}(n)$ versus the window size for a bdWSV applied to an evoked chloride series (n13, experiment 5, pH=5.0) gives H = 0.104, indicating the presence of long-term memory. (**b**) The slope 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.

Figure 1.5a 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. 1.5b). In Fig. 1.5c, 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 ($\overline{H}$=0.498±0.01). For illustration purposes, Fig. 1.5c 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) (Peng et al., 1995), 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 unreliable (Cannon et al., 1997). 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 $\overline{H}$=0.697±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.

1.6. Conclusions

1. The results of the root mean square fluctuation analysis revealed non-trivial correlations in all experimental time series. The α exponent has a mean of 0.927 (R^2≥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 conditions (here defined by the pH of the cellular external medium).

2. Ca^{2+}-dependent Cl^- data present transitions from positive to negative correlations. 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.

3. By calculating the slope of the Power Spectral Density plot, we have concluded that the Cl^- data sets can be categorized as fractional Brownian motion.

4. 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.

5. By applying a shuffling procedure (21,000 shuffled time series in total), we have shown that the Hurst exponent values measured from the original experimental series ($\overline{H}$=0.191±0.101) were significantly different from the ones obtained after shuffling ($\overline{H}$=0.498±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.

Basic biomedical quantitative investigation

Research nº2: Intracellular dynamics of the adenylate energy system

2.1. Main objective

To understand some of the key elements that determine the cellular energy status of cells.

This work has been published in PloS one: On the dynamics of the adenylate energy system: homeorhesis vs homeostasis (2014). *PloS one*, *9*(10), e108676; by: De la Fuente, I. M., Cortés, J. M., Valero, E., Desroches, M., Rodrigues, S., Malaina, I., & Martínez, L.

2.2. Importance

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.

Our pioneer 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.

2.3. Brief background

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 (Jeong et al., 2000; Sear, 2005).

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 selforganized enzymatic functions and to maintain their complex biomolecular structures.

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 (Knowles, 1980; Nelson et al., 2008; Hardie, 2011), and the levels of ATP, ADP and AMP reflect roughly the energetic status of the cell (Nelson et al., 2008). 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. A characteristic of the temporal evolution of ATP, ADP and AMP concentrations is their complexity (Ataullakhanov & Vitvitsky, 2002). 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 (Ataullakhanov & Vitvitsky, 2002; Özalp et al., 2010; Ytting et al., 2012). In fact, under normal conditions inside the cell, the time evolution of the adenosine-59-triphosphate is subjected to marked variations presenting transitions between quasi-steady states and oscillatory behaviors (Özalp et al., 2010; Ytting et al., 2012).

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 (Getty-Kaushik et al., 2005), NAD(P)H concentration (Rosenpire et al., 2001), biosynthesis of phospholipids (Marquez et al., 2004), cyclic AMP concentration (Holz et al., 2008), actin polymerization (Rengan & Omann, 1999), ERK/MAPK metabolism (Shankaran et al., 2009). In addition, experimental observations in *Saccharomyces cerevisiae* during continuous culture have shown that the majority of metabolome also shows oscillatory dynamics (Murray et al., 2007).

At a global metabolic level, experimental studies have shown that the cellular metabolic system resembles a complex multioscillator system (Murray et al., 2007; Lloyd & Murray, 2005; Lloyd & Murray, 2006), what allows for interpretation that the cell is a complex metabolic network in which multiple autonomous oscillatory and quasi-stationary activity patterns simultaneously emerge (De la Fuente et al., 1999a; De la Fuente et al., 2008; De la Fuente et al., 2009; De la Fuente et al., 2010a; De la Fuente et al., 2011; De la Fuente et al., 2013).

Cells are open dynamic systems (De la Fuente, 2010b, De la Fuente, 2014), and when they are exposed to unbalanced conditions, such as metabolic stress, or physiological processes that produce drastic variations both in the concentration of the adenosine nucleotides (Özalp et al., 2010; Ytting et al., 2012; Edwards et al., 2012; Boender et al., 2011) and in their molecular turnovers (Lim et al., 2010).

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 (Atkinson & Walton, 1967), defined as:

$$AEC = \frac{[ATP] + 0.5[ADP]}{[ATP] + [ADP] + [AMP]} \qquad (1)$$

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 work 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 (Chapman et al., 1971; Ball & Atkinson, 1975; Swedes et al., 1975; Chapman & Atkinson, 1977; Walker-Simmons et al., 1977).

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 (Privalle & Burris, 1983), *Escherichia coli* (Weber et al., 2005), neurons (Chen et al., 2007), erythrocytes (Suska & Skotnicka, 2009), and fungi (Rakotonirainy & Arnold, 2008) among many others.

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 (Sel'kov, 1968), later developed by Goldbeter (Goldbeter, 1974), as well as by Heinrich and Rapoport (Rapoport et al., 1976). 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 (Sel'kov, 1975).

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

Here, in order to further understanding some 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.

2.4. The adenylate energy model

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 2.1):

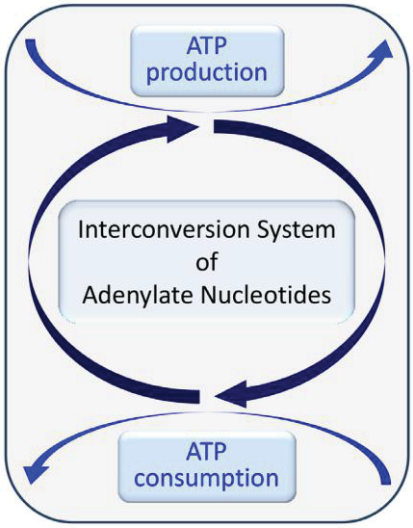

Figure 2.1. Elemental biochemical processes involved in the energy status of cells. The synthesis sources of ATP are coupled to energy-consumption processes through a network of enzymatic reactions which, interconverting ATP, ADP and AMP, shapes a permanent cycle of synthesis-degradation for the adenine nucleotides.

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

I. First, we have assumed the oxidative phosphorylation as the main synthesis source of ATP in the cell. 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) (Nelson et al., 2008).

$$ADP + P_i + nH^+_{inter-membrane} \rightarrow ATP + H_2O + nH^+_{matrix} \quad (2)$$

where n indicates the H^+/ATP ratio, with values between 2 and 4 which has been reported as a function of the organelle under study (Steigmiller et al., 2008).

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 (Nelson et al., 2008).

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.

$$2Mg^{2+} \cdot ADP \leftrightarrows Mg^{2+} \cdot ATP + AMP \qquad (3)$$

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):

$$Sbs + ATP \rightarrow Sbs(P) + ADP \qquad (4)$$

where Sbs and $Sbs(P)$ are the substrate and the product of the catalytic process, respectively.

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:

$$A + B + ATP \rightarrow AB + AMP + PP_i \qquad (5)$$

Figure 2.2 schematically shows the enzymatic processes of the ATP consuming-generating system:

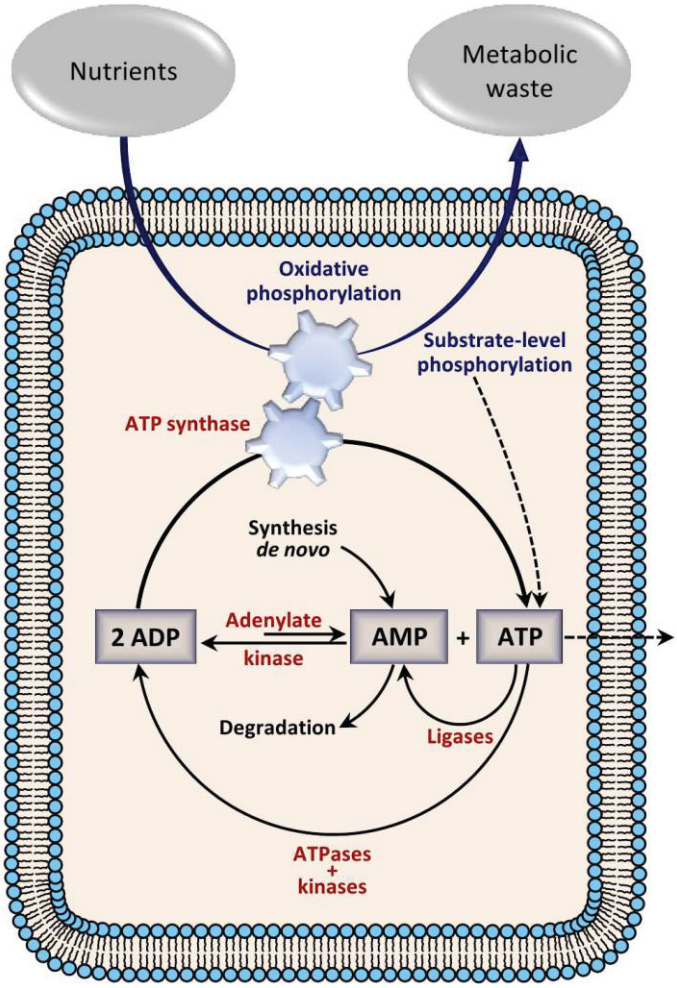

Figure 2.2. The Adenylate energy system. Oxidative phosphorylation and substrate-level phosphorylation generate ATP which is degraded by kinases (also ATPases) and ligases yielding ADP and AMP, respectively. The three adenine nucleotides are catalytically interconverted by adenylate kinase according to the needs of the metabolic system. AMP is also subjected to processes of synthesis-degradation, some AMP molecules are de novo iosynthesized, 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.

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 (Bønsdorff et al., 2004); whilst a part of AMP does not continue in the reactive system due to its hydrolysis, forming adenine and ribose 5-phosphate (Versées & Steyaert, 2003). 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 (Erlinge, 2010; Tanaka et al., 2010; Falzoni et al., 2013; Forsyth et al., 2011; Burnstock, 2012).

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:

$$ADP + P_i \xrightarrow{\Phi_1} ATP + H_2O \qquad (6)$$

$$ATP + AMP \xrightarrow{\Phi_2} 2ADP \qquad (7)$$

$$2ADP \xrightarrow{\Phi_3} ATP + AMP \qquad (8)$$

$$ATP \xrightarrow{\Phi_4} ADP + P_i \qquad (9)$$

$$ATP \xrightarrow{\Phi_5} AMP + PP_i \qquad (10)$$

$$\xrightarrow{v_1} ATP \xrightarrow{v_2} \qquad (11)$$

$$\xrightarrow{v_3} AMP \xrightarrow{v_4} \qquad (12)$$

where Φ_i (i =1,...,5) are the rates of the enzymatically-catalyzed reactions (6) to (10), 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 (Erlinge, 2010; Tanaka et al., 2010; Falzoni et al., 2013; Forsyth et al., 2011; Burnstock, 2012), being $v_2 = k_2[\mathrm{ATP}]$, v_3 is the rate of the biosynthesis *de novo* of AMP and v_4 is the rate of the sink of AMP, being $v_4 = k_4[\mathrm{AMP}]$. The reversible adenylate kinase reaction (3) has been described by its corresponding reactions (7) and (8) linked by a control parameter (see below for more details) allowing to move the reactive process to either of the two reactions according 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 (7), (9), (10) and (11), as well as a production of AMP, which is regulated by steps (6), (10) and (12).

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 (Nath & Jain, 2000). An iso-random Bi Bi mechanism has been reported for adenylate kinase kinetics (Valero et al., 2006; Sheng et al., 1999). 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 (Goldbeter & Lefever, 1972; Goldbeter & Prigogine, 1990), and finally, for the ligase family we have chosen threonyl-tRNA synthetase, which shows Michaelis–Menten kinetics (Curien et al., 2009).

The time-evolution of the ATP consuming-generating system (Figure 2.2) can be described by the following three differential equations:

$$\begin{aligned}
\frac{d\alpha}{dt} &= v_1 + \lambda\sigma_1\Phi_1 - \Delta'\sigma_2\Phi_2 + \Delta''\sigma_3\Phi_3 - \sigma_4\Phi_4 - \sigma_5\Phi_5 - v_2, \\
\frac{d\beta}{dt} &= -\lambda\sigma_1\Phi_1 + \Delta'\sigma_2\Phi_2 - \Delta''\sigma_3\Phi_3 + \sigma_4\Phi_4\,, \qquad (13) \\
\frac{d\gamma}{dt} &= v_3 - \Delta'\sigma_2\Phi_2 + \Delta''\sigma_3\Phi_3 + \sigma_5\Phi_5 - v_4,
\end{aligned}$$

where the variables α, β and γ denote the ATP, ADP and AMP concentrations, respectively, σ_1,…, σ_5 correspond to the maximum rates of the reactions (6) – (10), 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. The numbers Δ' and Δ'' are also control parameters of the system regulating adenylate kinase activity towards the synthesis or the consumption of ADP, respectively, with $\Delta' = 2 - \Delta''$.

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 (Ellis, 2001), 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 (Nenninger et al., 2010; Zuo, 2007; Mori et al., 1986; Ueda et al., 1987).

Moreover, experimental studies have shown that metabolism exhibits complex oscillations in the concentrations of individual adenine nucleotides, with periods from seconds to several minutes (Özalp et al., 2010; Ytting et al., 2012), 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 (De la Fuente et al., 1996a; De la Fuente et al., 1996b; De la Fuente et al., 1998; De la Fuente, 1999b; De la Fuente & Cortes, 2012).

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 $t - r_i$ where r_i are the delays and t is the time, and consequently the metabolite supplies to the enzymes (substrates and regulatory molecules) are not instantaneous; other dependent variables may be evaluated in 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 (14)) 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 (7) is used by the reaction (8) in the same place, and therefore, we have also considered that ADP concentration ($\beta(t)$) is not delayed in this process (Eq (16)). Therefore, the enzymatic rate functions are written as follows:

$$\Phi_1 = \frac{\beta(t - r_2)}{\beta(t - r_2) + K_{m,1}\left(1 + \frac{\alpha(t)}{K_{I,1}}\right)} \tag{14}$$

$$\Phi_2 = \frac{\alpha(t - r_1)\gamma(t - r_3)}{K_2 + K_{m,2}^{ATP}\gamma(t - r_3) + K_{m,2}^{AMP}\alpha(t - r_1) + \alpha(t - r_1)\gamma(t - r_3)} \tag{15}$$

$$\Phi_3 = \frac{\beta(t)^2}{K_3 + 2K_{m,3}^{ADP}\beta(t) + \beta(t)^2} \tag{16}$$

$$\Phi_4 = \frac{\alpha(t - r_1)\big(1 + \alpha(t - r_1)\big)(1 + \beta(t - r_2))^2}{L_4 + (1 + \alpha(t - r_1))^2(1 + \beta(t - r_2))^2} \tag{17}$$

$$\Phi_5 = \frac{\alpha(t - r_1)}{K_{m,5} + \alpha(t - r_1)} \tag{18}$$

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. (17) are divided by 1 µM 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 2.1.

Table 2.1. Values of the kinetic parameters used to simulate some of the dynamics of the adenylate energy system.

Param.	*Value*	*Reference*	*Param.*	*Value*	*Reference*
$\boldsymbol{\sigma_1}$	7.14µmol s^{-1}	Soga et al., 2011	$\boldsymbol{\sigma_3}$	800µmol s^{-1}	Thuma et al., 1972
$\boldsymbol{K_{m,1}}$	30 µmol	Nath & Jain, 2000	$\boldsymbol{K_3}$	1,360 μmol^2	Sheng et al., 1999
$\boldsymbol{K_{I,1}}$	25 µmol	Nath & Jain, 2000	$\boldsymbol{K_{m,3}^{ADP}}$	29 µmol	Sheng et al., 1999
$\boldsymbol{\sigma_2}$	800 µmol s^{-1}	Abrusci et al., 2007	$\boldsymbol{\sigma_4}$	100µmol s^{-1}	Goldbeter, 1974
$\boldsymbol{K_2}$	71,000 μmol^2	Valero et al., 2006	$\boldsymbol{L_4}$	10^6	Blangy et al., 1986
$\boldsymbol{K_{m,2}^{ATP}}$	25 µmol	Valero et al., 2006	$\boldsymbol{\sigma_5}$	0.43 µmol s^{-1}	Curien et al., 2009
$\boldsymbol{K_{m,2}^{AMP}}$	110 µmol	Valero et al., 2006	$\boldsymbol{K_{m,5}}$	100 µmol	Curien et al., 2009

For these values, the preliminary integral solutions of the system of differential equations (13) 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:

$$\alpha_0(t) = C + D\sin(\frac{2\pi}{P}) \quad (19)$$

$$\beta_0(t) = E + F\sin(\frac{2\pi}{P}) \quad (20)$$

$$\gamma_0(t) = G + H\sin(\frac{2\pi}{P}) \quad (21)$$

with C=6 μmol, D=2 μmol, E=4 μmol, F=1 μmol, G=7 μmol, H=3 μmol and P=200 s. The values of the other parameters were v_1 =35 x 10^{-3} μmol s^{-1}, k_2 =9 x 10^{-5} s^{-1}, v_3 =1.4 μmol s^{-1}, k_4 =0.69 s^{-1}, Δ'=1.98, r_1=5 s, r_2=27 s and r_3=50 s.

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} (Hagen, 2006) and the turnover number for RNA polymerase II is less rapid, about 0.16 s^{-1} (Jin et al., 1998). However, many cellular processes occur on a time scale of minutes (Richard et al., 1996; Weber et al., 2005; Petty & Kindzelskii, 2001). 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 and 120 seconds, covering a wide range of cellular physiological processes.

2.5. Results

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

- In Scenario I, λ is the 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.

- In Scenario II, the delay r_2 is the control parameter, modeling the time constants for the time delays of ADP (this Scenario is described in Tome II).

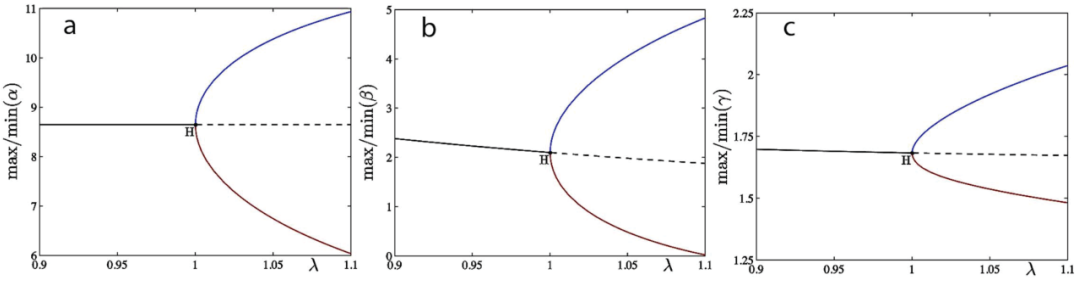

Figure 2.3. Numerical analysis for the model of the adenylate energy system. In y-axis we are plotting the max and the min of the variables α, β and γ. For situations with no oscillations (stable fixed point colored in solid black lines) the max and the min are coincident. For situations with oscillations, the max and the min of the oscillations are plotted separately; we are coloring in blue the max of the oscillation, in red, its minimum value, and λ is the control parameter.

The numerical integration illustrated in Figure 2.3 shows that at small λ values, for $0.9 \leq \lambda \leq 1$, 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 μmol s^{-1} (Soga et al., 2011). For values of λ bigger than 1 the attractor of the system is a stable limit cycle (therefore, the Hopf bifurcation is supercritical).

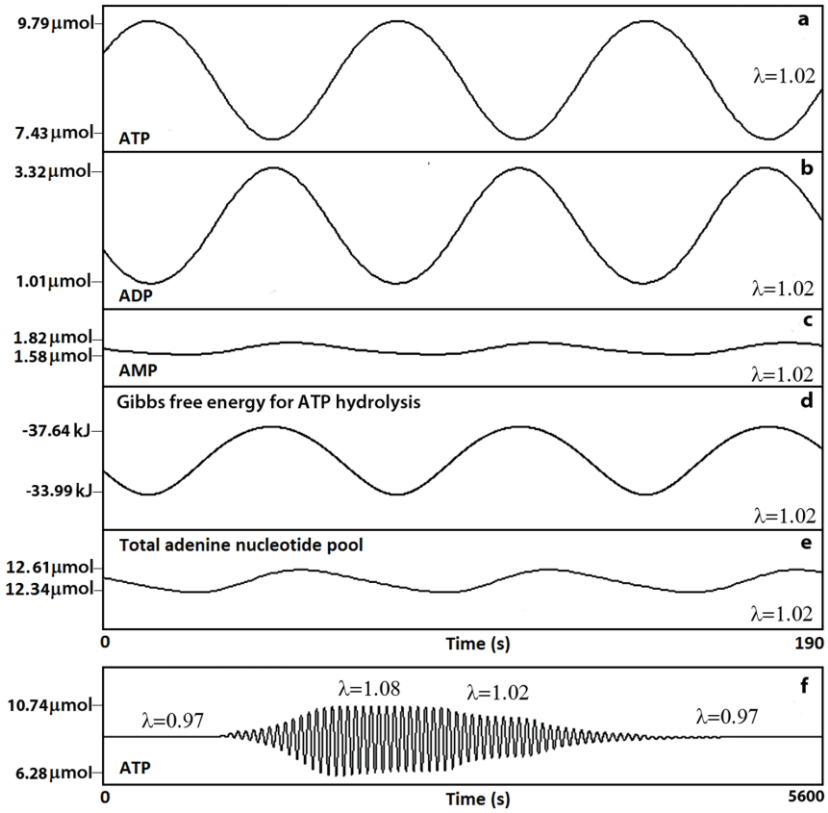

Figure 2.4. Dynamical solutions of Scenario I. For $\lambda = 1.02$ (normal activity for the ATP synthesis), periodic oscillations emerge. (**a**) ATP concentrations. (**b**) ADP concentrations. (**c**) AMP concentrations. (**d**) The Gibbs free energy change for ATP hydrolysis to ADP. (**e**) The total adenine nucleotide (TAN) pool. (**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.

Figure 2.4 shows three time series belonging to ATP, ADP and AMP (panels a, b and c, respectively), for $\lambda = 1.02$. It can be 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 $\Delta G'_{reaction} = \Delta G'^{0}_{reaction} + RT \ln(\beta/\alpha)$. The change of the standard Gibbs free energy for this reaction was previously evaluated by Alberty and co-workers (Alberty & Goldberg, 1992) 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 2.4d 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 $kJmol^{-1}$ and -33.99 $kJmol^{-1}$, 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 levels of the adenine nucleotide pool occur under different physiological conditions (Bonzon et al., 1981). We have estimated the total adenine nucleotide (TAN) pool as [ATP] + [ADP] + [AMP], and Figure 2.4e 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.

Figure 2.4f illustrates ATP transitions between different periodic oscillations and a steady state pattern for several values of λ (0.97, 1.08, 1.02, and 0.97).

Next, to analyze the dynamics of the energetic status of the system we calculated the energy charge level. Figure 2.5 shows different oscillatory patterns for AEC.

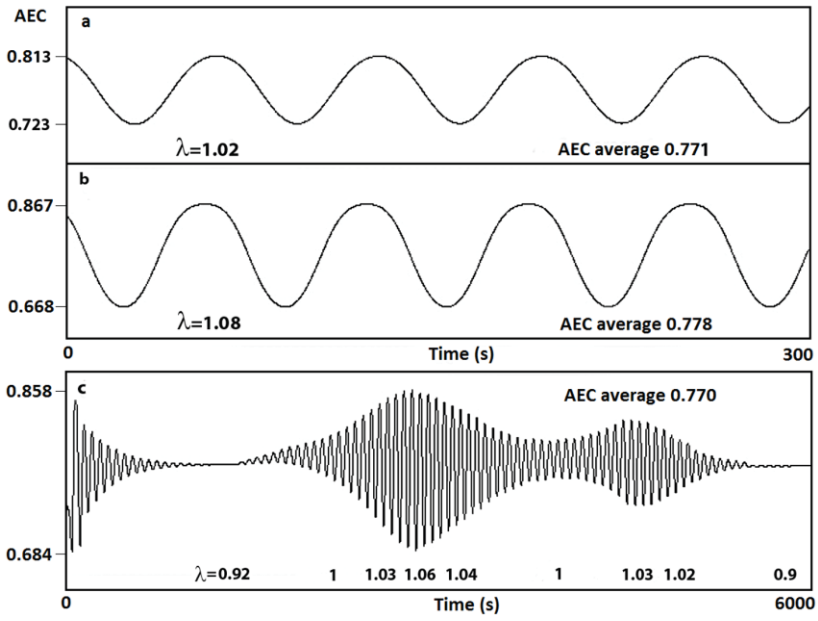

Figure 2.5. Emergence of oscillations in the AEC. Different oscillatory behavior appears when varying λ, the modifying factor for the ATP synthesis. (**a**) For $\lambda = 1.02$ (normal activity of ATP synthesis) the AEC periodically oscillates with very low relative amplitude of 0.09. (**b**) At higher 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, and 0.9).

For $\lambda = 1.02$ the AEC periodically oscillates with a low relative amplitude of 0.09 ($max = 0.813$ and $min = 0.723$) (panel a). At higher values of ATP synthesis (an increment of 8%) larger oscillations emerge ($max = 0.867$ and $min = 0.668$) (panel

b). Finally, panel c 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, and 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.

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 (Chapman et al., 1971; Ball & Atkinson, 1975; Swedes et al., 1975; Chapman & Atkinson, 1977; Walker-Simmons et al., 1977) (Figure 2.6).

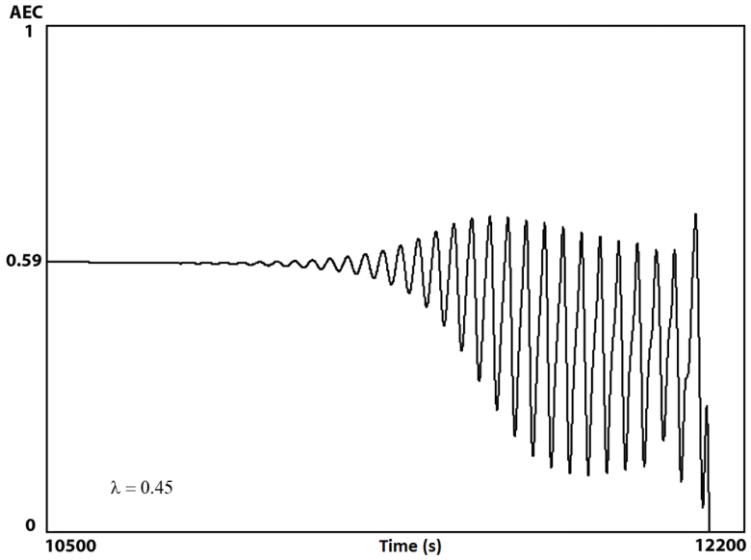

Figure 2.6. AEC dynamics under low production of ATP. At very small λ values, for $\lambda \leq 0.45$, the AEC exhibits values below 0.6, which are gradually descending up to reach very small energy values, when the system finally collapsed.

Lastly, 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* (Richard et al., 1996). These cells were quenched 5 min after adding 3 mM-KCN and 20 mM-glucose at time intervals of 5 s. Figure 2.7a 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 (Richard et al., 1996). Figure 2.7b shows an AEC oscillatory pattern at high values of ATP synthesis ($\lambda = 1.1$), $max = 0.873$, $min = 0.656$ and a period of 65 s.

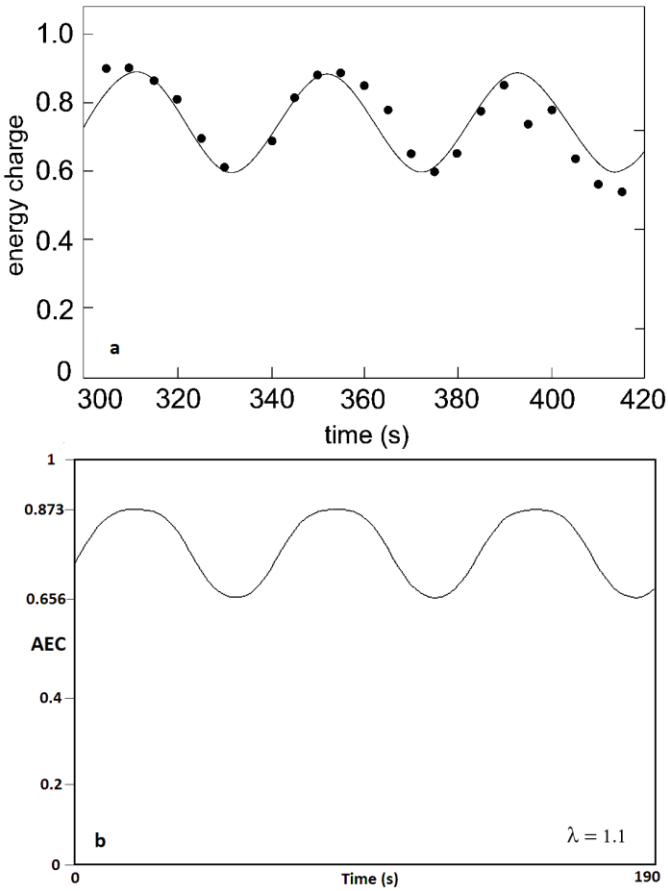

Figure 2.7. Experimental vs numerical results of AEC oscillations. Figure 2.7a illustrates a classical study of the intracellular adenine nucleotides in a population of intact cells belonging to the yeast *Saccharomyces cerevisiae* (Richard et al., 1996) which exhibits AEC rhythms, with $max = 0.9$, $min = 0.6$ and a period around 50 s. The authors fitted the experimental points to a sinusoidal curve. Figure 2.7b 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.

2.6. Conclusions

1. The adenylate energy system exhibits complex dynamics, with steady states and oscillations including multi-stability and multifrequency 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 (Özalp et al., 2010; Ytting et al., 2012).

2. The model is in agreement with previous experimental observations (Özalp et al., 2010; Ytting et al., 2012; Richard et al., 1996), showing oscillatory solutions for adenine nucleotides under different ATP synthesis conditions, at standard enzymatic concentrations, and for different ADP delay times.

3. 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 (Richard et al., 1996; Weber et al., 2005).

4. During the oscillatory patterns, ATP and ADP exhibit antiphase oscillations (the maxima of ATP correspond with the minima of ADP) also experimentally observed in (Richard et al., 1996).

5. As a consequence of the rhythmic metabolic behavior, the total adenine nucleotide pool exhibits oscillatory patterns (see experimental examples of this phenomenon in (Bonzon et al., 1981; Ching, 1972), as well as the Gibbs free energy change for ATP hydrolysis (see (Richard et al., 1996)). In agreement with these results, we have found that the oscillation for the Gibbs free energy has maximum and minimum values per period of -37.64 $kJmol^{-1}$ and -33.99 $kJmol^{-1}$, the same order of magnitude as in experimental observations (about -50 $kJmol^{-1}$ in rat hepatocytes) (Moran et al., 2011).

6. 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 (Richard et al., 1996).

7. The adenylate energy charge (AEC) does not substantially change during the simulations, indicating that it is strongly buffered against the perturbations, in agreement with experimental data (Chapman et al., 1971; Ball & Atkinson, 1975; Swedes et al., 1975; Chapman & Atkinson, 1977; Walker-Simmons et al., 1977).

8. When the maximum of the energy charge oscillations reaches a very small value (AEC~0.28) the adenylate system suddenly collapses, which leads to the death of the cell.

9. There exists an invariant of the energy function which restricts the values that the adenylate pool dynamics can take, and the equation of Atkinson is the manifestation of that invariant function.

Translational biomedical quantitative investigation

Research nº3: Vaccine design through combinatorial methods

3.1. Main objective

To introduce a new criterion in the design of vaccines which, besides complementing the approach of getting high coverage, guarantees that in all viral sequences at least a minimum number of epitopes is covered.

This work has been published in Journal of Mathematical Biology: A combinatorial approach to the design of vaccines (2015). *Journal of mathematical biology*, *70*(6), 1327-1358; by: Martínez, L., Milanič, M., Legarreta, L., Medvedev, P., Malaina, I., & De la Fuente, I. M.

3.2. Importance

Traditional vaccination methods have been proven unable to protect efficiently against viruses with high mutability rate. The usual criterion of including the most frequent epitopes in order to maximize the coverage leads to an unprotection against several variants of the viruses, which end up escaping from the immune system and perpetuating the infection.

In this work, we present a method which offers a balanced protection against all the considered virus variants. As a consequence, it opens a new scenario to design vaccines against viruses lacking of effective vaccines, such as HIV, HCV or Influenza.

3.3. Brief background

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 & 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 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 & 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 work 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 support 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 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, 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.

3.4. Shortest λ-Superstring, and Shortest λ-Cover Superstring problems

In this work 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 key definition of this section.

Definition 1. Let $S_1, \dots, S_k$ be in A^*, let $T \subset A^*$ be a set of *target strings*, and let $\lambda \in \mathbb{N}$. A λ- *superstring* for $(S_1, \dots, S_k, T)$ is a string $\boldsymbol{v} \in A^*$ such that for every $i \in \{1, \dots, k\}$, at least λ different target strings are common substrings of both S_i and $\boldsymbol{v}$.

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

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

Definition 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

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

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

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

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

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

Then:

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

We will give here a formal presentation of the Shortest λ-Superstring problem.

Shortest λ-Superstring problem

Instance: Strings $S_1, \ldots, S_k$ over a finite alphabet A, a finite set $T \subset A^*$ of *target strings*, and a *coverage requirement* $\lambda \in \mathbb{N}$.
Task: Find a λ-superstring for $(S_1, \ldots, S_k, T)$ of minimum length.

The Shortest λ-Superstring problem is a minimization problem, with the set of feasible solutions given by all λ-superstrings for $(S_1, \ldots, S_k, T)$. Clearly, the problem defined with an instance $(S_1, \ldots, S_k, T, \lambda)$ is feasible if and only if at least λ different substrings of each S_i belong to T. Since this condition can be efficiently tested, we will assume in the rest of the work 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 $(S_1, \ldots, S_k, T)$ is also a λ_1-superstring. Consequently, denoting by $\alpha(S_1, \ldots, S_k, T; \lambda)$ the minimum length of a λ-superstring for $(S_1, \ldots, S_k, T)$, it holds that

$$\alpha(S_1, \ldots, S_k, T; \lambda_1) \leq \alpha(S_1, \ldots, S_k, T; \lambda_2),$$

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 $S_1, \ldots, S_k$ represent the relevant protein sequences, and the set 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 **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, a 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.

Example 2: Let A be an alphabet of cardinality 20 representing amino acids, $T = A^9$ representing the 9-mer epitopes, and

$$S_1 = \text{ARNDCQEGHPLQQFTSTTQV},\ S_2 = \text{AIRDVIEGHILKMFPSTWWV},$$
$$S_3 = \text{AINDCQEGTITKMFPSTWYV}.$$

Then:

ARNDCQEGHILKMFPSTWYV is a 1-superstring for (S_1, S_2, S_3, T) of length 20 (Figure 3.1).

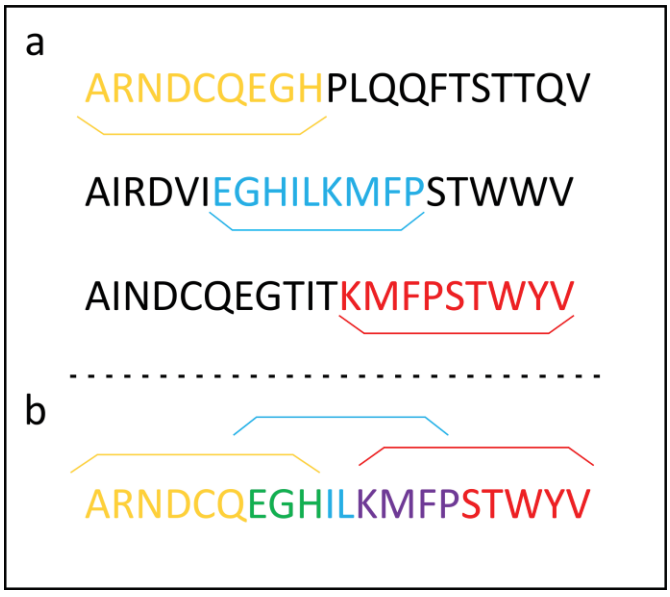

Figure 3.1. 1-superstring example. In panel (**a**), the strings S_1, S_2 and S_3, with the *epitopes* of the 1-superstring highlighted in yellow, blue and red respectively. (**b**) The resulting 1-superstring of length 20, where the overlapping segments are colored in green (the common *amino acids* of the *epitopes* of S_1 and S_2) and in purple (the coincident *amino acids* of the *epitopes* of S_2 and S_3).

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

Proposition 1. $\alpha(S_1, \dots, S_k, T; \lambda) \leq k\lambda\tau$, *where* τ *denotes the maximum length of a target string.*

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

Proposition 2. $\alpha(S_1, \dots, S_k, A^l; \lambda) \leq k(l + \lambda - 1)$.

Definition 3. For given strings $S_1, \dots, S_k$ and a target string $\in T$, we define the frequency of $\mathbf{t}$ in $\{S_1, \dots, S_k\}$ to be $\mathrm{f}(\mathbf{t}) = |\{i \mid \mathbf{t}\ is\ a\ substring\ of\ S_i\}|$.

Definition 4. If $\boldsymbol{v}$ is a λ-superstring for $(S_1, \dots, S_k, T)$, we define the *coverage* of $\boldsymbol{v}$ to be

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

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

Proposition 3. *Let* $T = A^l$ *and* $S_1, \dots, S_k \in A^m$ *for some positive integers* l, m. *Then, the coverage of every* λ*-superstring* $\boldsymbol{v}$ *satisfies* $c(\boldsymbol{v}) \geq \frac{\lambda}{m-l+1}$.

As we have shown in Proposition 3, 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 1 we gave λ-superstrings for λ from 1 to 4 for $A = \{0, 1\}$, $T = A^3$ 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 $\frac{11}{27}$. (As a matter of fact, $\frac{11}{27}$ is the maximum coverage attainable by a string of length 5 for this choice of A, T and $S_1, \dots, S_5$.) The string $s = 01001$ has also length 5 and coverage $\frac{11}{27}$, 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.

Definition 5. 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 $\boldsymbol{v} \in A^*$ such that for every i, at least λ elements of X_i are substrings of $\boldsymbol{v}$.

With the previous definition, we can now describe formally the Shortest λ-Cover Superstring problem.

Shortest λ-Cover Superstring problem

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

More formally, the requirement of the Shortest λ-Cover Superstring problem is to find $\boldsymbol{v} \in A^*$ minimizing $l(\boldsymbol{v})$ such that for all $i \in \{1, \ldots, n\}$, it holds that

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

The Shortest λ-Superstring and Shortest λ-Cover Superstring problems are computationally difficult: not only are they NP-hard, they are also hard to approximate. In Tome II, we prove that the two problems are polynomially equivalent. In addition, we 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.

3.5. Solving the Shortest λ-Cover Superstring problem

3.5.1. An integer programming approach

In this section, we describe how to solve the Shortest λ-Cover Superstring problem using integer programming (IP). 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 $(A, X_1, \ldots, X_n, \lambda)$ be an instance of the Shortest λ-Cover Superstring problem. We construct a complete directed edge-weighted graph $G = (V, E, w)$, called the *distance graph*, as follows:

– The vertex set of G is $V = T \cup \{s^*\}$ where $T = \bigcup_{i=1}^{n} X_i$, and s* is a new vertex. We identify the vertices of G other than s^* with the corresponding strings from T. In particular, for $i, j \in T$, notation $i \subseteq j$ means that i is a substring of j and $i \subset j$ that i is a proper substring of j (that is, $i \subseteq j$ and $i \neq j$).

– For every two distinct vertices $s, t \in V \backslash \{s^*\}$, 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 $s \in V \backslash \{s^*\}$, add the arc (s, s^*) to E and assign to it weight $w_{s,s^*} = l(s)$.

– For every vertex $s \in V\backslash\{s^*\}$, add the $arc\ (s^*, s)$ to E and assign to it weight $w_{s^*,s} = 0$.

To express the Shortest λ-Cover Superstring problem in graph theoretic terms, we need one more definition. A subgraph H of G is said to *cover* a string $s \in T$ if there exists a vertex $t \in V\ (H) \cap T$ such that $s \subseteq t$. For each $i \in \{1, \dots, n\}$, we will denote the set of all strings in X_i covered by H by $V_i(H)$.

Definition 6. A directed cycle C in the distance graph G is said to be feasible if it satisfies the following conditions:

1. $s^* \in V(C)$.

2. For every two distinct vertices $\mathbf{s}, \mathbf{t}$ *from* $V(C) \cap T$, $\mathbf{s}$ *is not a substring of* $\mathbf{t}$.

3. For every $i \in \{1, \dots, n\}$, *we have* $|V_i(C)| \geq \lambda$.

The following proposition establishes the connection between λ-cover superstrings and feasible directed cycles in the derived distance graph. The *weight* of a directed cycle C in G with edge set F is defined as $\sum_{e \in F} w_e$.

Proposition 4. There exists a λ-cover superstring for $(X_1, \dots, X_n)$ of length at most l if and only if G contains a feasible directed cycle C of weight at most l.

We seek for a shortest feasible directed cycle C in G. Therefore, we formulate the IP algorithm for the Shortest λ-Cover Superstring problem as follows:

Define the variables:

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

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

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

where i ranges over all elements of V, and

$$z_i = \begin{cases} 1, & if\ C\ covers\ the\ string\ corresponding\ to\ vertex\ i, \\ 0, & otherwise, \end{cases}$$

where i ranges over all elements of T. The IP formulation is the following:

$$\min \sum_{i,j} w_{ij} x_{ij} \quad // \quad s.t.\ y_{s^*} = 1 \quad // \quad \sum_{i \in V\ :i \neq j} x_{ij} = y_j \quad \forall j \in V \quad //$$

$$\sum_{j \in V\ :j \neq i} x_{ij} = y_i \quad \forall i \in V \quad // \quad \sum_{i \in X_j} z_i \geq \lambda \quad \forall j \in \{1, \dots, n\} \quad // \quad \sum_{i \subseteq j} y_i \geq z_i \quad \forall i \in T \quad //$$

$$y_i + y_j \leq 1 \quad \forall i, j \in T\ such\ that\ i \subset j \quad // \quad 0 \leq x_{ij} \leq 1,\ \ x_{ij} \in \mathbb{Z} \quad //$$

$$0 \leq y_i \leq 1,\ \ y_i \in \mathbb{Z} \quad // \quad 0 \leq z_i \leq 1,\ \ z_i \in \mathbb{Z}.$$

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, we are only interested in solutions that consist of a single directed cycle. There are several ways to eliminate these subtours. Here, we have used the so called Miller–Tucker–Zemlin (MTZ) formulation (Miller et al. 1960); for more details see Tome II.

3.5.2. A hill-climbing approach

We have developed a hill-climbing algorithm to find short λ-superstrings for given strings $S_1, \dots, S_k$, a given set 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 $\boldsymbol{v} = \boldsymbol{v}_1 +' \dots +' \boldsymbol{v}_k$, where each $\boldsymbol{v}_i$ is likewise an overlapping sum of λ consecutive different substrings of S_i from 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 $\boldsymbol{v}_{i,j}$ is deleted from $\boldsymbol{v}$. In the transformations of the second kind, each substring $\boldsymbol{v}_{i,j}$ is changed for every possible substring of $S_1, \dots, S_k$ from T that is not already a substring of $\boldsymbol{v}$. Changes of the first kind and of the second kind are applied consecutively to $\boldsymbol{v}$ (but each one of them only to $\boldsymbol{v}$, 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 $\boldsymbol{v}'$ diminishes, then we replace $\boldsymbol{v}$ with $\boldsymbol{v}'$ 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 $\boldsymbol{v}$ 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.

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 from an alphabet of cardinality 20. For the set T of target strings we took the set of all strings of length $l = 5$. 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 in 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 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. The lengths of the obtained λ-superstrings and their coverages are shown in

Fig. 3.2a and b, respectively (the numerical values are given in Tome II). 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 the Results section, this phenomenon of good performance for intermediate values of λ seems to happen also when experimental data are considered.

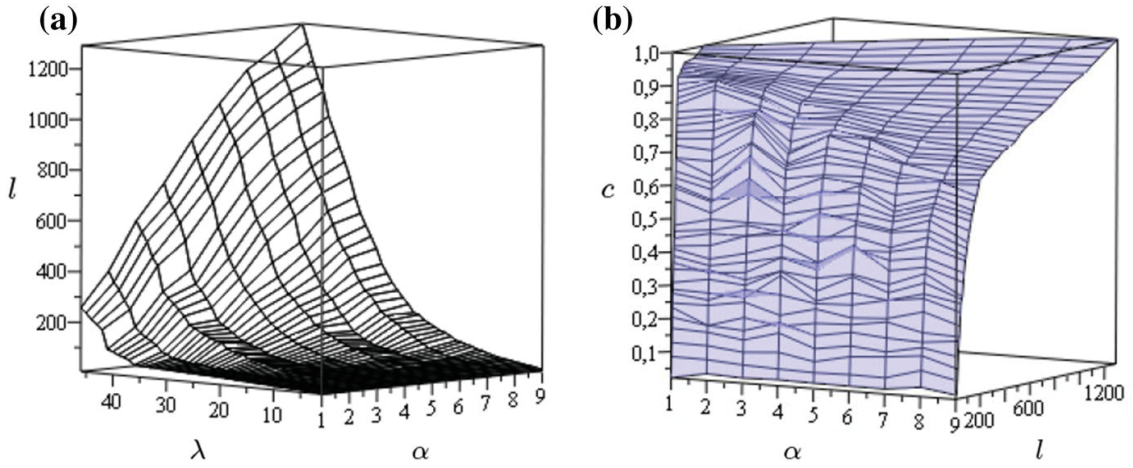

Figure 3.2. Numerical simulations of the hill-climbing algorithm. Panel (**a**) represents the λ-superstring lengths, the α (the percentage of mutation) values, and the λ values of each simulation. In panel (**b**), we depict the λ-superstring lengths, the α values and the coverage of each simulation.

3.6. Results

3.6.1. 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 did not include them in our calculations so that we can work with high values of the parameter λ. 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 $T = A^{10}$ as the set of target strings. The lengths of the λ-superstrings obtained, their coverages, and the GenBank (2013) IDs of the sequences corresponding to the hemagglutinin (HA) genes are given in Tome II. These lengths and coverages are plotted in Fig. 3.3a and b, respectively. As shown in the figures, 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. 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%.

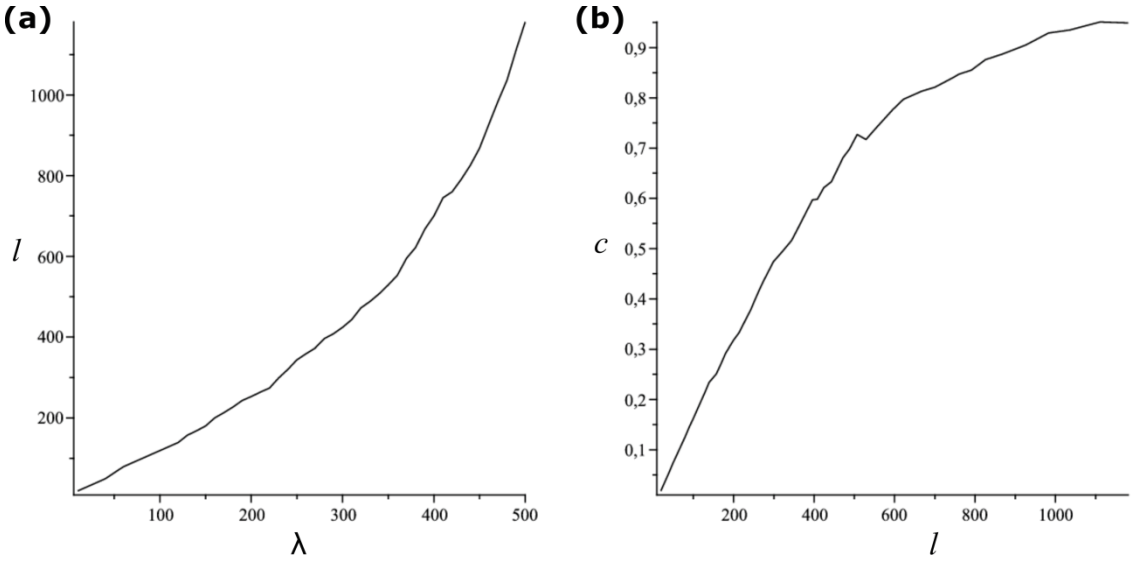

Figure 3.3. Results of hill-Climbing algorithm for hemagglutinin. Panel (**a**) represents lengths l versus the λ values of each λ-superstring solution for hemagglutinin. In panel (**b**), we depict the λ-superstring lengths against the respective coverage value.

3.6.2. Hill-climbing algorithm 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 added 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-gene length, because beyond that value the increase in coverage was lower with respect to increase in length. They obtained sequences of three-gene 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 in the case of Nef protein and for a subset of 166 sequences (shown in Tome II) in the case of Gag protein, the same level of coverage after 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.

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 Section 3.5.2. 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 T_v of epitopes in the λ-superstring. More specifically, we assigned to each epitope e not in T_v the fitness

$$\phi(e) = p \cdot rf(e) + \frac{1-p}{18|T_v|} \cdot \sum_{e' \in T_v} \Big(ov(e, e') + ov(e', e) \Big),$$

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 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 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 & 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:

YTPGPGTRFPLTFGWCFKLVPVDPEEVGFPVKPQVPLRPMTYKAAVDLSHFLQNYTPGPGTRYPLTFGWCFKLVP
VEPDQNYTPGPGVRYPLTFGWPTVRERMRRAEPAAEGVGAVSRDLERHGAITSSNTAATNADCAWLERPMTYKAA
LDLSHFLREKGGLEGLIHSQKRQDILDLWIYHTQGYFPAADGVGAASRDLEKHGMDDPEREVLEWRFDSRLAFH
HVARELHPEYYKDCFKLVPVEPEKIEEANEGENNSLLHPMSLHGMEDPEKEVLVWKFDSRLVPVEPEKVEEANEG
ENNCLLHPMSQHMGGKWSKRSVEKANEGENNAACAWLEAQEDEEVGFPVRPQVPLRPMTYKGALDLSHFLKEA
REKHPEYYKRQEILDLWVYHTQGYFPDWMGGKWSKSSITSSNTAANNADCAWLEAQEEEEVGFPVRPMTYKGAV
DLSHFLKEKGGLEGLVYSQRRQDILDLWVYHNSLLHPMSQHGMDDPEKEVLMWKFDSRLAFHHMARELHPEYY
KNCLLHPMSLHGMDDPEKGGLEGLIYSQKRQDILDLWVYNTQGYFPDWQNYTPGPGIRYPLTFGWPAVRERMRR
AEPAADGVGAVSRDLEKHGAITSSNTAT

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 $c(\boldsymbol{v})$ are as in Definition 4, and $T = A^l$.

Proposition 5. *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*

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

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 $T = A^l$ 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-gene length. The solution that we found (depicted in Tome II) had length 1,495 and obtained coverage close to the upper bound given by Proposition 5, which is 85.4.

3.6.3. Integer programming algorithm for Nef

An optimal solution to the integer programming problem derived in Section 3.5.1, 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 that the hill-climbing method from Section 3.5.2. However, denoting $t = |T|$, 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 $T = A^l$, with $|A| = 20$ and $l \in \{9, 10\}$ (as was done in Subsections 3.6.1 and 3.6.2), 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 Section 3.5.1 (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 87:

IRYPLTFGWCFKLVPVGFPVRPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIYSQKRQDILDLWVYHTQGYFPDWQ NYTPGPGVRYPL

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 was applied 10^5 times.

3. A suboptimal integer programming algorithm was used (the integer program described in section 3.5.1 is an optimal corrected version of the one appearing in the original paper collected in Tome II. However, we have maintained the suboptimal solution of the original paper to illustrate the comparative study in Figure 3.4, because even the suboptimal version achieved much better results than the other techniques).

The lengths of the λ-superstrings are showed in Fig. 3.4. 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. The suboptimal 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 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 suboptimal solution found by the integer programming algorithm is 19% of the length of the one found by the brute force algorithm.

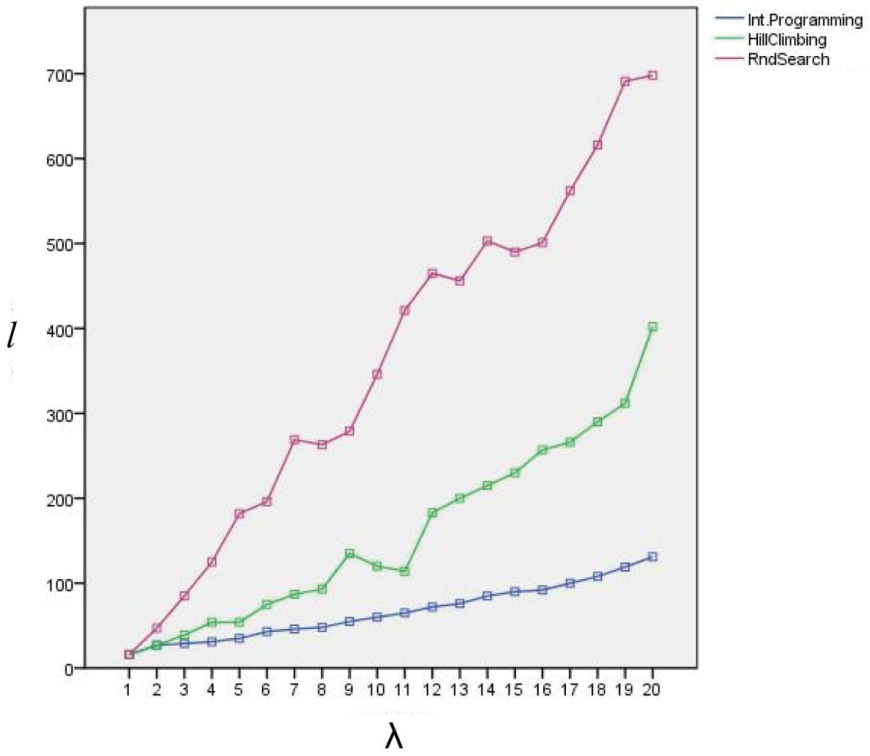

Figure 3.4. Comparative study on Nef protein. Length l of the λ-superstrings as a function of λ for a random concatenation of epitopes (*pink*), the hill-climbing algorithm (*green*) and the integer programming algorithm (*blue*).

3.7. Conclusions

1. The shortest λ-superstring and the shortest λ-cover superstring problem are NP-hard and polynomially equivalent.

2. The λ-superstring approach improves computational vaccine design by selecting an adequate balance of epitopes over the selected sample strings.

3. Since the hypothetical maximum coverages (67.8% for Nef and 85.4% for Gag) are close to the coverages achieved by our algorithms (62% and 82% respectively), we conclude that our methods are also a good approach to achieve good coverages.

4. Our algorithms have obtained the same coverage (62% for Nef and 82% for Gag) as the algorithms focused in optimizing this measure. Therefore we conclude that including the λ-superstring criterion does not necessarily have a cost on the coverage.

5. The integer programming approach gives optimal solutions, and outperforms the hill-climbing algorithm. However, it has a much bigger computational cost than the other methods, which restricts its use.

6. This new criterion opens the possibility for designing efficient vaccines against viruses with high mutability rate such as HIV, HCV or Influenza.

Clinical biomedical quantitative investigation

Research nº4: Preterm labor prediction by autoregressive models

4.1. Main objective

To introduce a new tool to forecast which patients admitted to the hospital because of Suspected Threatened Premature Delivery will give birth in less than seven days.

This work has been published in: Montevideo Units Vs Autoregressive Models on Preterm Labor Detection (2016). *ITISE Proceedings* 2016, 799-807; by: Malaina, I., Matorras, R., Martínez, L., Fernandez-Llebrez, L., Bringas, C., Aranburu, L. & De La Fuente, I. M.

4.2. Importance

Preterm delivery affects about one tenth of human births and is associated with an increased perinatal morbimortality as well as with remarkable costs. Even if there are a number of predictors and markers of preterm delivery, none of them has a high accuracy.

Here, for the first time, we have used mathematical modeling to predict preterm labor immediacy basing only in 30 minutes of uterine pressure recordings. By the use of autoregressive parameters, we have been able to predict correctly almost 70% of the cases. This approach could be used as a new tool to assist in the prognosis and treatment jointly with other clinical proofs.

4.3. Brief Background

According to the World Health Organization, every year 15 million babies are born preterm, i.e., before the 37th gestational week. This problem affects more than one tenth of all human births, and the complications that it entails lead to the death of nearly one million children each year, making its early detection and tracing a crucial matter for society. The risk factors for preterm birth are diverse, ranging from pre-eclampsia, vascular disease, uterine overdistension, previous preterm births, and low maternal body-mass to stress (Goldenberg et al., 2008; Copper et al., 1996), among many others.

In order to detect threatened premature delivery, obstetrical emergency units usually perform a systemic and obstetric examination, a blood analysis, a vaginal fibronectin determination, a vaginal ultrasound and an external cardiotocography (CTG). Although some indicators have been associated with premature delivery (Goldenberg et al., 1998; Iams et al., 1996; Newman et al., 2008), the available protocols to predict this complication are still far from perfect. In fact, in a review of 22 different tests, it was concluded that their quality and accuracy was generally poor (Honest et al., 2009). Current screening tests for the prediction of preterm labor can be divided into three general categories: risk factor assessment, cervical measurement, and biochemical markers (Georgiou et al., 2015).

First, prediction based on risk factor assessment alone is unreliable (Georgiou et al., 2015). While, for instance, a previous delivery before the 34th week of pregnancy poses a risk factor of 13 (Esplin et al., 2008), basing the prognosis in such factors alone would lead to misprediction in more than 50% of the cases (Georgiou et al., 2015).

Cervical measurement by transvaginal ultrasound is the most employed technique in the prediction of preterm delivery. Specifically, a short cervical length is associated to a relative risk of 6 (Goldenberg et al., 1998). A meta-analysis of 28 studies reported sensitivities ranging from 53% to 67% and specificities ranging from 89% to 92% for delivery within one week (Sotiriadis et al., 2010). However, due to limitations in ultrasound availability and operator expertise, cervical length cannot be universally and reliably utilized to routinely predict preterm labor on its own (Georgiou et al., 2015; Berghella et al., 2009).

Lastly, cervicovaginal fetal fibronectin and phosphorylated insulin-like growth factor binding protein-1 (phIGFBP1) are among the most widely utilized biomarkers of preterm delivery, in spite of their shortcomings. A systematic review of 13 studies showed that when fibronectin was used to predict premature labor, its sensitivity and specificity were highly variable, ranging between 23-92%, and 59-97% respectively (Leitich et al., 1999; Revah et al., 1997). On the other hand, even if the protein phIGFBP1A is a good negative predictor of preterm birth with a specificity of 90.5-91.8%, it lacks of suitable sensitivity (22.2-69.2%) and positive predictive value (11.8-50%) for precise forecasting (Paternoster et al., 2007).

In addition to these three important prognosis tools, a crucial indicator of labor beginning and progression is the presence of uterine contractions. In clinical practice, they are evaluated through the recording of the intrauterine pressure by tocography devices. These contractions are characterized by a triple descending gradient: 1) propagation of the contractile wave in a descending direction; 2) longer duration of the

contraction at the uterine fundus than in the inferior parts of the organ; 3) stronger contractions in the upper parts of the uterus than in the lower areas (Alvarez, & Caldeyro-Barcia, 1954). In addition to these physical changes, several biochemical and electrophysiological processes are altered during uterine contractions (Tong et al., 2011; Lammers, 2013; Zhang et al., 2016). Nevertheless, there is universal agreement concerning the importance of some measures of uterine activity such as frequency, intensity and amplitude of contractions (Stout & Cahill, 2011).

Since labor occurs as a consequence of the combination of increased longer-lasting depolarizations, raised myocyte to myocyte connectivity, and the activation of intracellular contractile machinery (Smith et al., 2015), we hypothesized that some subtle changes in uterine dynamics (small modifications in the myocyte to myocyte coordination pattern) as well as in cell to cell electrophysiological alterations (specially changes related to bursting-type action potentials) (Zhang et al., 2016) could have an impact on preterm delivery, and that they could be ascertained in conventional cardiotocograms through a variability analysis of uterine pressure recordings.

4.4. Sample acquisition and processing

4.4.1. Sample acquisition

During a four-year period (from 2010 to 2013), from the 47,671 women who were assisted in the Obstetric and Gynecology section at Cruces University Hospital (Basque Country, Spain) 1,643 were assisted because of suspected threatened premature delivery (STPD). In Cruces hospital, this condition 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. In cases of discrepancy, if the disagreement was of less than five days, the LMP prevailed, whereas if the discrepancy was greater, the estimation by vaginal ultrasound was chosen. The STPD protocol included medical history, systemic and obstetric examination, blood analysis, vaginal fibronectin determination, vaginal ultrasound and at least 30 minutes of external cardiotocography (CTG) (recorded indistinctively by Philips Avalon FM30, Philips Avalon FM20, Hewlett Packard Vidria 50XM and Hewlett Packard 50IP cardiotocographs).

For the purpose of this study, threatened premature delivery (TPD) was considered when the following two conditions were given in a pregnant woman with less than 37.0 gestational weeks:

1. The recording of at least 4 uterine contractions in a 30 minute period, each lasting at least 30 seconds at the CTG, or 8 uterine contractions in a 60 minute period.
2. A cervical effacement ≥80% and at the same time a cervical dilatation ≥2cm. In cases in which the cervical effacement was <80% and/or cervical dilatation was

<2cm, a cervical shortening <25mm at 24-31 weeks (or <15mm at 32-34 weeks) was required.

This pathology was treated when, in addition to any of the previous criteria, the gestational age was ≥24.0 and <35.0 (or 34.0 if membranes are ruptured) gestational weeks. In Cruces University Hospital, TPD treatment consists in the administration of the oxytocin receptor blocker atosiban (Tractocile, Ferring Pharmaceuticals A/S, Copenhagen, Denmark). This drug was administered by the attending obstetrician in the delivery room on the following dose: an initial bolus (6.75 mg) using Tractocile 7.5 mg/ml solution for injection, immediately followed by a continuous high dose infusion (loading infusion 300 µg/min) of Tractocile 7.5 mg/ml concentrate for solution for infusion over three hours, continued by a lower dose of Tractocile 7.5 mg/ml concentrate for solution for infusion (subsequent infusion 100 µg/min) for up to 45 hours. In the remaining cases, patients (either not fulfilling gestational age criteria or uterine dynamic criteria) were sent home after 2-12 hours of observation. In cases with a cervical dilatation greater than >5cm, premature delivery was considered inevitable and no treatment was provided.

Two specific populations were considered for the study:

a) The "Delayed Delivery" population (DD), constituted by all those women whose delivery occurred more than seven days after the initial consultation (N=123 cases). 34 of these cases were excluded because the CTG was unavailable, leaving 89 recordings for the study.
b) The “Delivery Before 7 Days” population (DB7D) constituted by the women whose delivery occurred in the following seven days since their visit. This population was constituted by a subset of 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) from the total of women consulting because of TDP who delivered in seven days or less since the CTG recording. Finally the group was reduced to 328 cases, since 152 of the CTG's were not available.

The sample acquisition procedure is illustrated in the following figure:

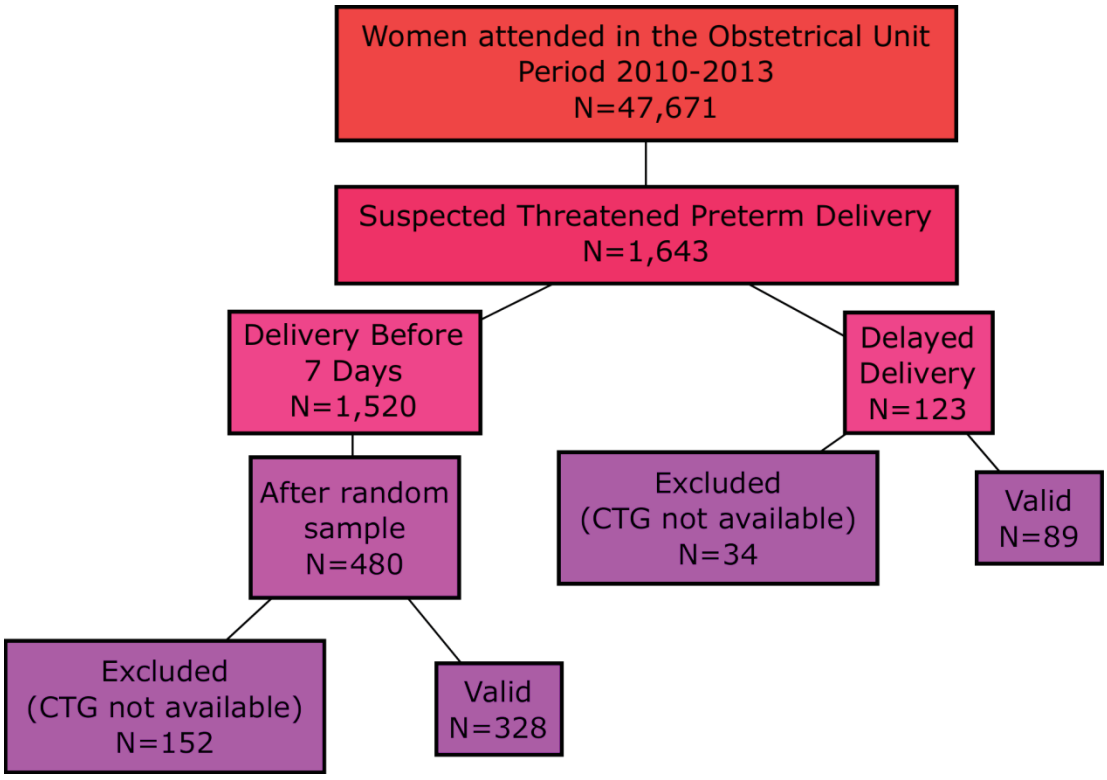

Fig 4.1. Flow chart of population selection. Flow chart representing the population selection. Earliest stages are represented in red, while latest are colored in purple.

All these patients signed the informed consent to participate in this study, which was approved by our center's investigation board (CEIC-E16/13).

4.4.2. Digitization process:

The first 30 minutes of the tocographies recorded in the first visit of the patient due to STPD were digitized by the open source program Engauge Digitizer 4.0. The original proportions of the CTG were maintained by placing the Cartesian coordinate system origin on the first square, that is, the one on the south-west of the measurement, and by considering the length of a square of the tocography to be the unit of the CTG. Then, the data were discretized to obtain approximately 2,000 time points, equidistant by 0.0291457 units. In Figure 4.2, we illustrate the digitization procedure:

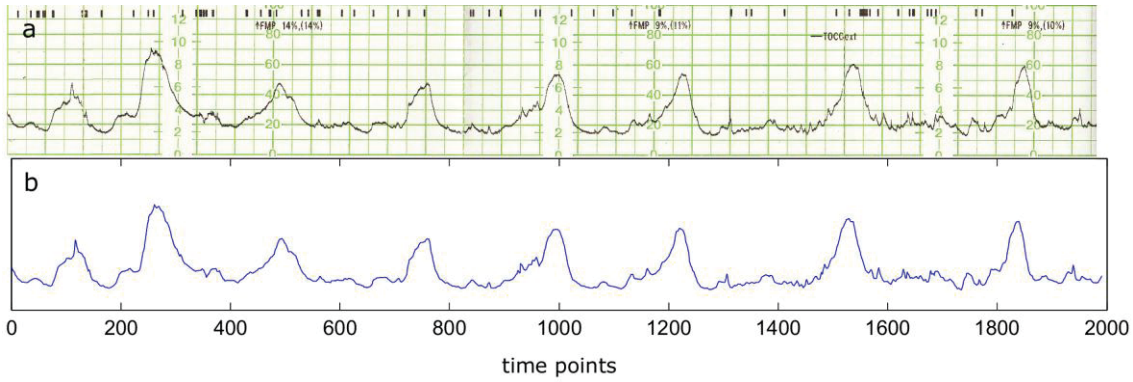

Fig 4.2. Digitization process applied to a generic tocography. (**a**) Scanned generic tocography recording (Delayed Delivery group, tocography nº1) of 30 minutes of duration. (**b**) Digitization of the same tocography by Engauge Digitizer 4.0, represented by 1,991 time points.

4.5. Results

First, we estimated the Montevideo Units (MVU) (Caldeyro-Barcia et al., 1957), 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. In Figure 4.3, we illustrate an example of its calculation.

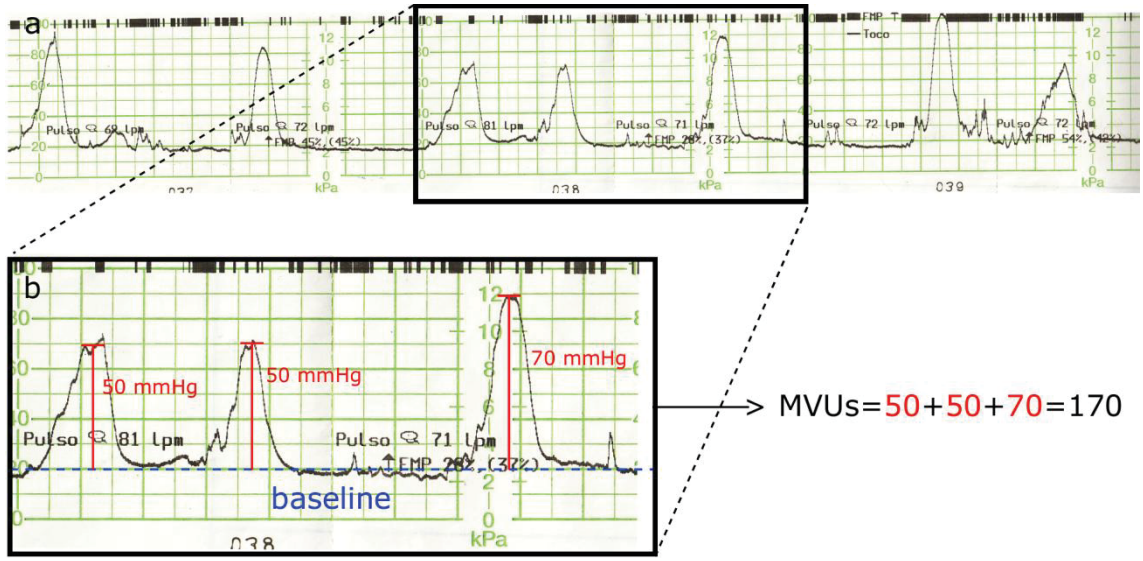

Fig 4.3. MVU calculation example. (**a**) Prototypical tocography (Delayed Delivery group, tocography nº2) Framed in black, the 10 minutes of maximum activity. (**b**) MVU calculation, in blue the baseline and in red the intensities of the contractions.

The result of this calculation for the Delayed Delivery group was 87.81±84.49 (mean±SD) while for the Delivery Before 7 Days group was 94.13±95.45. The

distributions of MVU values are illustrated in Figure 4.4 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 DB7D and the DD groups come from continuous distributions with equal medians.

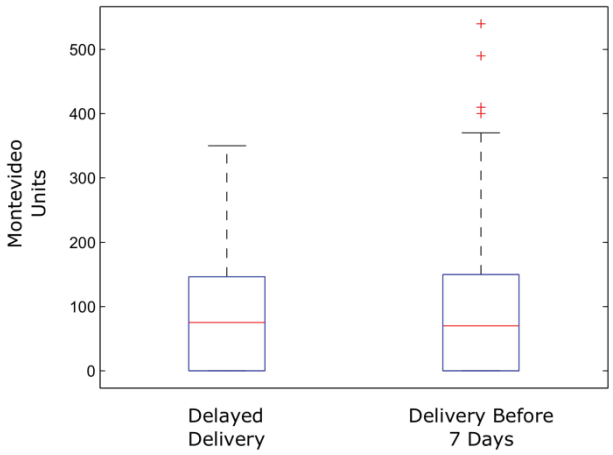

Fig 4.4. Box plot of the Montevideo Units for Delayed Delivery and Delivery Before 7 Days groups. Box plot illustration of the distributions of the MVU values calculated for the DD and the DB7D groups. The blue boxes represent the distribution of the central 50% of the values and the red lines represent the medians. The length of the arms is 1.5 times the size of the central values (the blue box), and the values outside these arms are represented 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 Delivery and the Delivery Before 7 Days groups.

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 DD group and 14.32% for the DB7D group, indicating that the relative number of cases ready for adequate labor according to the MVUs was very similar between both groups.

Next, we calculated the autoregressive approximation. As has been shown in a previous work (Malaina et al., 2016), the best model within the ARIMA family in order to model tocographies 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 DD group were $\varphi_1 = 1.61 \pm 0.16$, while for the DB7D group were $\varphi_1 = 1.49 \pm 0.21$, with a p-value of $4 \cdot 10^{-8}$ on the rank sum test.

Besides, we estimated predictive capacity of the autoregressive coefficient used as a predictor by calculating the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). To calculate these statistical measures, 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 balance between sensitivity and specificity criterion, i.e., the

one with the highest (sensitivity+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 below the threshold, and negative outcome was associated to values above 1.547.

Then, the predictive parameters were calculated as follows:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$\text{PPV} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}}$$

$$\text{NPV} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false negatives}}$$

In this case, we obtained a sensitivity of 0.764, a specificity of 0.585, a positive predictive value of 0.33 and a negative predictive value of 0.901. To illustrate the effect of varying the threshold on the sensitivity and the specificity, in Figure 4.5, the receiver operating characteristic (ROC) (Verhulst, 1838) curve for φ_1 is represented. This curve had an AUC (Area Under the Curve) of 0.685.

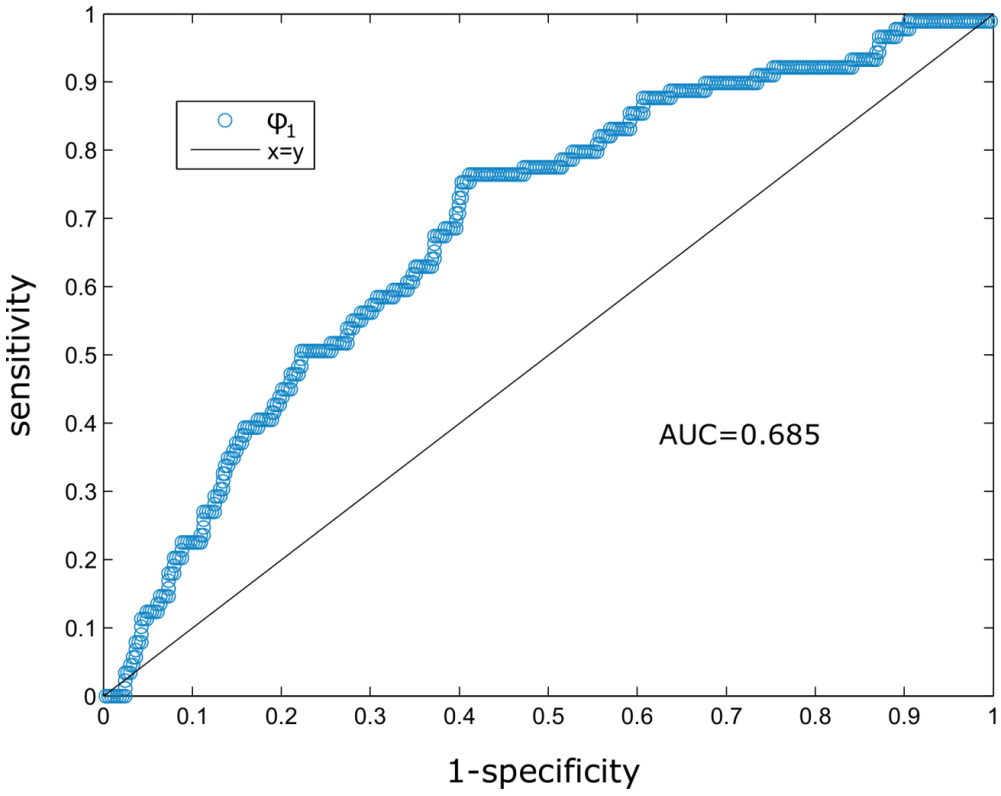

Fig 4.5. ROC curve of the first coefficient of the AR(2) approximation. The Y axis represents the sensitivity, while de X axes represents 1- specificity. Dots above the black line indicate good balance between sensitivity and specificity. The Area Under the Curve was 0.685.

Finally, we calculated these predictive parameters for the Montevideo Units used as a preterm labor immediacy predictor. Following the previous criterion, 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.

4.6. Conclusions

1. No significant differences were found between the Montevideo Units of women who will deliver within a week and women who will not. Thus, this measure seems unable to predict preterm labor immediacy.

2. Montevideo Units indicated that a very similar percentage of women of both groups (12% for the Delayed Delivery group and 14% for the Delivery Before 7 Days group) were ready for an adequate labor (i.e., had more than 200 MVUs). Therefore, we conclude that this measure should not be used to identify adequate labor condition.

3. There are significant differences between the autoregressive coefficients of first order of patients depending on the labor immediacy.

4. When delivery was nearer (the case of Delivery Before 7 Days group), the autoregressive parameter was closer to 1 than in the cases in which it was farther (the DD group). Therefore, the time series of the DB7D group were more dependent on the past values, than the ones of the DD group. Consequently, we consider that the degree of dependency on past values could be used as an indicator of the immediacy of labor.

5. Since the highest predictive parameter of the autoregressive approach was the Negative Predictive Value, the best way to use this indicator would be to confirm that a woman will not deliver within seven days.

6. The predictive parameters of the autoregressive model were 0.144±0.11 higher (on average) than the ones obtained by Montevideo Units. Thus, we suggest the use of an autoregressive approach instead of the MVU calculation.

Fundamental conclusions

There is nothing more important than preserving life, and the thesis here presented is framed in the field of quantitative biomedicine (or systems biomedicine), which has as objective the application of physico-mathematical techniques in biomedical research in order to enhance the understanding of life's basis and its pathologies and, ultimately, to defend human health.

With this purpose, we have performed a research activity in which physico-mathematical methods are applied in the three fundamental levels of Biomedical Research: basic, translational and clinical.

In the first basic research, we have found for the first time a dynamical structure characterized by highly organized data sequences, non-trivial long-term correlation that last in average 7.66 seconds, and "crossover" effect with transitions between persistent and anti-persistent behaviors in calcium-dependent chloride currents. Lastly, we disarrayed the time series by a shuffling procedure and observed that all these properties disappeared, which indicates that this dynamic structure is intrinsic to calcium-dependent chloride currents and did not occur by chance.

Through the second basic investigation, by modeling the adenylate energy system, we have found that in normal cellular conditions, the cellular energy charge (i.e., the relationship between ATP, ADP and AMP) is determined by a non-stationary oscillating function, bounded between 0.7 and 0.95.

At a translational level, through the use of combinatorial optimization techniques, we have developed a new method of vaccine design capable of fighting the viral mutability. We have proven that our method is able to achieve identical coverage level (62% and 82% in the case of HIV's Nef and Gag proteins, respectively) as algorithms focused in maximizing this quantity. At the same time, our technique offers a balanced coverage with respect to all the selected variants of the virus. Therefore, we consider that the λ-superstring approach offers the possibility of designing vaccines against viruses that, until date, due to their high mutability rate, do not have an efficient vaccine for fighting them, as is the case of the HIV, HCV or Influenza.

In the research at a clinical level, on the one hand, we have demonstrated that the classical quantitative method to measure uterine pressure (Montevideo Units) is incapable of predicting accurately preterm labor. On the other hand, we have designed a

new tool for premature delivery forecasting, based only in the recording of 30 minutes of uterine dynamics. Through the analysis of more than 400 tocographies, we have found significant quantitative differences in the uterine activity recordings of the patients, depending on the immediacy of labor. This study has shown that as the labor approaches, uterine pressure series show more dependency with respect to the past, which could be used as an indicator of labor proximity.

As a result, this thesis integrates investigations in the three main levels of Biomedical Research, and altogether, these investigations have originated four scientific publications. Besides, as far as we know, our work is the first European thesis which integrates in the same framework the application of mathematical knowledge to basic, translational and clinical biomedicine.

The scientific community is becoming increasingly aware of the importance of quantitative sciences in biomedicine. Just a few months ago, *Science* journal published an article entitled *Convergence: The future of health*. In this work (signed on behalf of more than 100 scientist), Phillip Allen Sharp (awarded with the Nobel Prize in 1993) and Susan Hockfield (MIT's president from 2004 to 2012) discuss the importance of the integration of life sciences, physical sciences, mathematics, engineering and information technology (so-called Convergence) in order to achieve new medical breakthroughs. Among the benefits of this integration, they highlight that: "Investing in Convergence, from early research through clinical applications, will transform health and provide health care cost savings. For example, Convergence can enhance early diagnosis." (e.g. our work at a clinical level); "Convergence can also increase the effectiveness of treatments. New immunotherapies and vaccines will enable our own bodies to better fight disease." (e.g. our translational research); or "Convergence can advance fundamental knowledge. New computational models of complex systems, advanced imaging at every scale (from subcellular processes to the whole body), and detailed characterization of protein, RNA, and DNA of single cells will expand our understanding of what makes us healthy or sick." (e.g. our basic investigation).

Although the level of inversion in Convergence is still small (for example, only about 3% of the budget of National Institutes of Health (NIH) of the U.S. goes to principal researchers in mathematics or physical sciences), there is an increasing trend of scientific institutes that realize the importance of investing in Convergence to develop new therapies and promote health innovation, which, ultimately, is the key to increase the quality of medical care.

Given the importance of quantitative sciences, which has been demonstrated in numerous current investigations, we would like to conclude this thesis with a brief reflection about the necessity of including at least one expert in quantitative techniques in a part of biomedical research teams. It is not imperative for medical specialists to know mathematical methods. It is better if they do, but it is not indispensable. What is necessary is that investigation teams include at least one person capable of understanding and transferring medical problems to mathematical language, which can only be done by an expert in biomedical quantitative sciences.

The merge of quantitative sciences and medicine will allow an effective advance of the most important and crucial issue in our modern world: the preservation and care of health and human life.

Annex

Mathematical applications to biomedicine through history, a brief summary

Philosophy [i.e. physics] is written in this grand book, which stands continually open before our eyes (I say the 'Universe'), but cannot be understood without first learning to comprehend the language and know the characters as it is written. It is written in mathematical language, and its characters are triangles, circles and other geometric figures, without which it is impossible to humanly understand a word; without these one is wandering in a dark labyrinth. (Galilei, 1623).

Mathematics is essential for the development of biomedicine. Nowadays, quantitative techniques are being applied to improve diagnosis, design of new treatments and prediction of diseases' progression, yet the mathematization of biology and medicine dates back to long ago.

This historical summary is a short review of the importance of mathematics in the growth of biomedicine, and is organized as follows: first, we go back to the classical era, when quantitative methods were born; then, we summarize various biomedical fields that have been able to advance thanks to mathematics; next, we highlight several examples of new disciplines within quantitative biomedicine resulting as a merge of mathematics and biomedicine; and finally we describe some of the most important biomedical projects, in which mathematics have played a fundamental role.

A.1. Classical era, the origin of concepts

This section summarizes the origin of two fundamental ideas in this thesis: the necessity of applying mathematics to biology and medicine, and the development of quantitative methods.

The need of mathematics to describe the world that surrounds us has been present since, at least, the classical era. Due to the influence of the studies on Number Theory (a branch of mathematics developed by the school founded by Pythagoras (*ca*560- *ca*480 B.C.)), the notion of numbers being essential to the study of the physical and biological world arose in the minds of the scientist of the time (Lancaster, 1994). An example of this line of thought is the ancient Greek philosopher Plato (427-347 B.C.), who believed that the laws that regulate causalities in nature could be defined in mathematical terms.

Pursuing this idea, the Greeks were able to widely develop the disciplines of physics and astronomy by applying mathematics to those fields. However, they did not have the same success solving problems related to biology, because, among other reasons, this subject was insufficiently grown at the time.

One of the main developments of biology came by the hand of Aristotle (384-322 B.C.), who was able to distinguish and classify various sub-disciplines of this field. Besides, he is considered the founder of the comparative method (Mayr, 1982), a technique of analysis to describe, organize and explain data by making observations of the similarities and differences between subjects. Aristotle's method has been widely used in medicine, for instance, to study the etiology of scrotal cancer (Pott, 1775), the etiology of melanoma (Lancaster, 1956) or the association of tobacco with lung cancer (Wynder et al., 1959). In addition, comparative method led inevitably to measuring numerically the degree of distinction, which ended up giving birth to quantitative techniques (Harvey& Pagel, 1991).

A.2. Modern era, the development of mathematics in biomedicine

There were some important contributors to the mathematization of biology and medicine during the Dark Ages, such as Leonardo of Pisa (*ca*1175-*ca*1250) (also known as Fibonacci) due to his famous model of rabbit population growth (published in Liber Abaci in 1202 and translated later to modern English in (Sigler, 2003)) or Nicolaus Cusanus (1401-1464), who proposed the idea of quantifying the pulse through the use of a water clock (Moore, 1908) and promoted describing experimental observations in mathematical terms (Cusanus, 1450).

However, it was not until the modern era (1500-present) that mathematics was found really indispensable for the development of biomedicine. In the following section, we summarize the evolution of some fields that, without the fundamental role played by mathematics, would not be the sciences that we know today.

A.2.1. Genetics

One of the greatest geneticists of all times is Gregor Johann Mendel (1822-1884), mostly known due to the laws of Mendelian inheritance. He studied agriculture and physics, as well as mathematics. In fact, he attended a course in combinatorial mathematics, which was fundamental in his revolutionary approach to genetics. His interest in the hereditary properties led him to experiment by combining (breeding) different lines of plants and observing the properties of the resulting offspring. Besides being able to explain why some characteristics are inherited, he quantified the proportion of a property that passes to the next generation.

Another indispensable figure in the development of genetics and in the mathematization of biology is Francis Galton (1822-1911). Although his main interest was in genetics, he was able to notice and criticize that there was a huge lack of quantitative methodology in biology (Galton, 1874). One of his main contributions to genetics is known as Galton's ancestral law, which is related to Mendel's laws and is also based in combinatorial concepts. In short, this law states that the contribution of two parents to the total heritage of the offspring is 1/2; the four grandparents contribute in 1/4th of the total, and so on (Galton, 1894).

Francis Galton had a close relationship with a mathematician named Karl Pearson (1857-1936). Pearson greatly contributed to the application of mathematical techniques to biomedicine and has been credited with establishing the discipline of mathematical statistics (Bronowsky, 1978). In his greatest series (Contributions to the Mathematical Theory of Evolution, Vol I-II and Mathematical Contributions to the Theory of Evolution, Vol. III-XVI), Pearson was able to bring together and give mathematical form to numerous problems of evolution and genetics. Despite the efforts of Pearson and Galton to introduce quantitative techniques in biology, the Royal Society stated that mathematics and biology were not to be discussed together (Lancaster, 1994). As a consequence, in 1901 Pearson, Galton and Walter Weldon (1860-1906) (a zoologist who contributed to the mathematization of evolution theories) founded *Biometrika,* a journal dedicated to promote the study of mathematics and statistics in biology.

Pearson's series about mathematics in evolution inspired a great mathematician and biologist named Ronald Fisher (1890-1962) to focus on evolutionary and genetic problems. By the use of mathematics, he was able to combine Mendelian genetics and natural selection, developing the new Darwinist synthesis of evolution known as the Modern Evolutionary Synthesis. Besides that, one of the biggest achievements of Fisher was the development of the statistical models known as Analysis of variance (ANOVA). Ronald Fisher introduced the term variance in 1918 (Fisher, 1918) and performed his first application of ANOVA in 1921 (Fisher, 1921). Ever since, these methods are one of the most used techniques to compare populations. In addition, the ANOVA test and his recent work (Fisher, 1925) influenced the chemist Frank Wilcoxon (1892-1965) to develop two new statistical non-parametric proofs, i.e., Wilcoxon rank-sum test and Wilcoxon signed-rank test (Wilcoxon, 1945). Nowadays both tests are extensively used in biomedicine, and have been applied in the researches nº1 and nº4 of this thesis as well.

The aim to understand genetics and heredity from a quantitative point of view also lead to the development of two fundamental pieces in current science: regression method and correlation coefficients (which are also applied in research nº1).

Francis Galton, as a consequence of reading his cousin's (Charles Darwin) book Origin of Species, came up with the idea that the normal curve could describe the variability of observations in both mental and physical characteristics of humans. After accumulating data from hundreds of individuals, he was able to confirm years later that physical characteristics follow a normal distribution (Galton, 1889). Nevertheless, his real interest was to demonstrate that intelligence was inherited; in other words, he wanted to prove that the intelligence of children was "co-related" with the intelligence of their parents. Even so, he realized that something as unquantifiable as human mental characteristics was not the best variable to the develop regression and correlation techniques, so he studied both concepts in human physical characteristics and in sweet peas.

Despite the merit of applying the first regression method corresponded to Legendre (Legendre, 1805), Galton was the one who, through his experiments combining biology and mathematics, was able to discover and explain the use of the regression line (Galton, 1886). Besides, he gave a first definition of the correlation coefficient:

"Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction.... It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes... If they were in no respect due to common causes, the co-relation would be nil." (Galton, 1888).

Later, Karl Pearson provided a mathematical framework to this correlation coefficient, which nowadays is known as Pearson correlation coefficient. One of the earliest example of the application of this statistic in biology was provided by Walter Weldon (co-founder of Biometrika), in his study to measure the physical characteristics of shrimps (Weldon, 1892), and till date, this coefficient has been applied in more than 200,000 published works.

A.2.2. Epidemiology

One of the best examples where mathematical basis are applied to develop biomedicine is the field of epidemiology. This discipline analyzes the patterns and effects of diseases in populations, and is usually approached by mathematical modeling.

While the mathematician Pierre-François Verhulst (1804-1849) was studying the population growth in Belgium, he realized that sigmoid curves (also called S-shape curves) were accurate representations of growth's progression (Verhulst, 1838). Years later, he called this curve Logistic Curve (Verhulst, 1845), but its modern mathematical use did not occur until Galton fit several experimental data sets to this curve (Galton, 1875). The regression model derived from the logistic curve, named Logistic Regression, is probably the most utilized model in biomedicine, and has been applied,

for instance, to model biomarker's dynamics in Alzheimer's disease (Jack et al., 2013) or to predict immune cell response (Pradeu et al., 2013).

The logistic curve also played a fundamental role in the development of epidemiology. One of the most important figures in the mathematical modeling of diseases is the physician Ronald Ross (1857-1932). Ross successfully recognized the mosquito that transmits the malaria and was awarded with a Nobel Prize in 1902 for discovering the life cycle of malarial parasite. He showed a great interest in mathematics since he was a child, which gave him the necessary background to make one of the most important contributions to mathematical modeling, which is the introduction of differential equations in this discipline (Ross, 1915). By using the logistic curve and ordinary differential equations, he was able to develop a model (named Ross Model) to explain the relationship between the number of mosquitoes and the incidence of malaria in humans (Ross, 1916). At the same time, Ross used his model to show that the reduction of mosquito numbers "below a certain figure" was sufficient to counter malaria, introducing for the first time the concept of Transmission Threshold.

Since then, differential equations have been widely applied to model biomedical processes like, for example, cancer growth (Wang et al., 2015), drug responses to HIV (Xiao et al., 2013) or the dynamics of the adenylate energy system (De la Fuente et al., 2014), which is one of the works resulting from this thesis (explained extensively in research nº2).

A.2.3. Cardiology and pressure quantification

Mathematical modeling has also been indispensable in the development of cardiology and the understanding of human circulatory system, through the quantification of pulse and blood pressure.

One of the earliest figures who developed a mathematical model of the pulse is the physician Bryan Robinson (1680-1754), also known for being a mathematical writer. By studying the pulse in different species, Robinson found that the pulse rates per minute n were inversely proportional to the length l of the animal (according to (Plackett, 1988)). He shaped these results as the mathematical model:

$$n = 1606.6 \cdot l^{-0.75},$$

where l is measured in inches.

At that time, the arterial blood pressure was first quantified by the cleric Stephen Hales (1677-1761), who performed his measurements by fastening a glass tube inside a horse's artery (Hales, 1733). Later, the physicist and mathematician Daniel Bernoulli (1700-1782) related the fluid pressure to its velocity (Bernoulli, 1738), a result known as Bernoulli's Principle. Within medical field, he applied his principle to the circulatory system and estimated the blood pressure depending on its speed.

Bernoulli's observations set the basis to the research of Adolf Eugen Fick (1829-1901), a physician who started studying mathematics before medicine. Fick is known

for describing the laws of diffusion (Fick, 1855), but also for defining the *Fick's Principle*, which states that the blood flow to an organ can be calculated by using a marker substance. Among its multiple applications, Fick used this principle to measure the cardiac output, i.e., the volume of blood being pumped by the heart per unit of time (Fick, 1870).

The measurement of fluid pressure extended from cardiology to other medical areas such as neurology, where the pressure of cerebrospinal fluid is measured in order to diagnose neurological diseases, or obstetrics, where the pressure of uterine contractions is quantified to study labor progression.

A.2.4. Particle dynamics

In addition to its influence in cardiology, Fick's mathematical model of diffusion has played a fundamental role in the field of particle dynamics, which is essential for studying processes like drug delivery (Grief & Richardson, 2005) or cancer metastasis (Namimoto et al., 1997).

Robert Brown (1773-1858) was a Scottish botanist who besides giving the earliest detailed description of the cell nucleus (Brown, 1833), observed an important type of particle motion, namely, the Brownian motion. While he was observing the particles trapped in cavities inside pollen grains in water through a microscope, he realized that those particles were moving across the water. However, in spite of his finding, he was unable to describe the mechanisms that originated this motion.

These dynamics were not explained mathematically until the astronomer Thorvald Nicolai Thiele (1838-1910) described them years later (Thiele, 1880). Ever since, Brownian motion model has been widely applied in medicine, for instance to analyze medical images (Chen et al., 1989) or to estimate the speed of retroviral infections (Chuck et al., 1996).

In addition to this achievement, Thiele is known for making a substantial contribution to mathematical modeling by introducing the Likelihood Function (Thiele, 1889). This function is especially important for methods based on the estimation of parameters, and has been widely applied in biomedical research, for example, to model indicator-dilution curves (Kuenen et al., 2014), quantify the abundance of protein conformations (Onuk et al., 2015) or study the similarities between different sequences of Zika Virus (Wang et al., 2016).

Some years later, due to his studies on the diffusion equation described by Fick (Fick, 1855), Albert Einstein (1879-1955) was able to describe the Brownian motion by means of the diffusion equation (Einstein, 1905). In that work, he stated that the diffusion coefficient was related to the mean square displacement of the Brownian particle. In addition, in that same work (Einstein, 1905), Einstein provided empirical evidence for the existence of atoms and molecules. Specifically, he was able to determine the size of atoms, the number of atoms in a mole, or the molecular weight of a gas in grams, by relating the diffusion coefficient to measurable quantities.

In 2015, another Einstein theory was confirmed, namely, the existence of gravitational waves. This verification has occurred thanks to the researches of the mathematician Yves Meyer (1939-present), due to his work on wavelet theory (for which he has been granted with the 2017 Abel prize). Meyer's mathematical theory has been applied to a huge amount of sciences, as is the case of biomedicine, where the wavelet analysis has been used to study EEG, ECG, brain rhythms, DNA, proteins and so on. In fact, due to his contributions, nowadays we can obtain digital images of our organs.

A.3. Multidisciplinarity, creating new areas of science

The previous section highlights that the evolution of several fields in biomedicine has been linked to the application of physical and mathematical tools. Most of the scientists mentioned in the previous section were able to apply more than one area of knowledge in their research, which spotlighted the advantages of approaching real world problems by multiple disciplines.

Consequently, there has been an increasing emphasis in teamwork involving multidisciplinarity (Poulton & West, 1993), and especially in health research (Stokols et al., 2005). In fact, it has been accepted that there are problems that cannot be adequately addressed by single disciplines alone (Choi & Pak, 2006), and due to the technologization of sciences, multidisciplinarity has become the norm (Alvargonzález, 2011). Some of the fields approached by this perspective are, for instance, molecular biology, neurosciences, microscopic cytology, or nanotechnology (Schmidt, 2008).

In the following lines, we focus on describing three multidisciplinary fields integrated within quantitative biomedicine, which are also the basis of the three levels of research (basic, translational and clinical) performed in this thesis, namely, systems biology, immunoinformatics and quantitative diagnosis.

A.3.1. Systems biology

Systems biology is the framework for the investigations carried out in the basic level. This discipline started as a new area of biology resulting from the combination of molecular and cellular biology with computational and mathematical approaches (Wolkenhauer et al., 2013), and aims to develop a system-level understanding of biological interactions (Kitano, 2000). Thus, systems biology's objective is to understand the relationships between genes, proteins and metabolites in a global manner, instead of locally. Indeed, one of the virtues of systems biology is that it can model the complexity of the cell and its environment robustly (Werner et al., 2014).

Norbert Weiner (1894-1964) is one of the first developers of system-level analysis in biological sciences (Kitano, 2002). His work (Wiener, 1948) gave birth to biological cybernetics, which in words of the famous mathematician Andrey Kolmogorov (1903-1987) can be defined as:

"Science concerned with the study of systems of any nature which are capable of receiving, storing and processing information so as to use it for control." (Kolmogorov, 1958).

Even if the basis were set, systems biology did not awake much interest among the scientific community. This can be attributed to the fact that molecular biology was still an emerging discipline at that time, so there were no sufficient data of biological systems performance to deepen in this approach (Kitano, 2002).

The revolution of molecular biology came when Watson and Crick identified the structure of the DNA (Watson & Crick, 1953). Ever after, the field of molecular biology has made enormous progresses, promoted by technologies for making comprehensive measurements on gene expression profiles, protein-protein interactions, etc. (Kitano, 2002).

As a consequence, systems biology gained renewed interest and has been extensively developed since. This discipline can be beneficial to several clinical research aspects like, for example, to understand drug resistance, predict effective combination therapies (Werner et al., 2014) or analyze the effect of external stimuli in a system (Kitano, 2000). In this thesis, we have applied a systems biology's approach to study the informational structure of experimental calcium-activated chloride fluxes belonging to *Xenopus laevis* oocytes under different pH stimuli (De la Fuente et al., 2017) and also to model the adenylate energy system (De la Fuente et al., 2014), which are fully discussed in researches nº1 and nº2.

The logical next step and necessary expansion of systems biology is its extension to biomedicine, i.e., the mathematical modelization of medical systems. Since many diseases originate due to cellular systemic malfunction, a deeper understanding of the mechanisms underlying cell metabolism and functions is necessary for developing new therapies. Besides, the emergence of diseases is a nonlinear dynamical phenomenon, which requires quantitative monitoring of key biological parameters at molecular, cellular and physiological levels (Wolkenhauer et al., 2013). As a consequence, the systemic perspective has become a widespread methodology to confront different diseases.

An example of the systemic approach to medical field is the modeling of cancer. Cancer is one of the greatest killers in the world, and appears from sub-cellular to macroscopic scale, operating in a systemic manner (Bellomo et al., 2004). In cancer systems biology, the analysis is focused on understanding how intracellular networks of normal cells are perturbed during carcinogenesis, in order to develop effective predictive models which assist clinicians in the validation of new drugs and therapies (Werner et al., 2014).

One of the first mathematical models related to cancer was built by the zoophysiologist Schack August Steenberg Krogh (1874-1949). Based in Fick's diffusion equation (A.2.3), he developed a mathematical model to describe the diffusion into a cylinder of tissue from a blood vessel (Krogh, 1919). Likewise, that model lead to the development of another important diffusive model considering both the distribution of oxygen and the distance to the center of the tumor (Burton, 1966). Few years later, the first model of the dynamics of metastatic processes was proposed (Liotta et al., 1974).

In that work, they represented the systemic interactions between tumor volume, vascularization, number of single tumor cells, number of pulmonary metastasis and tumor cell clumps in tumor venous fluent.

From there on, clinicians became increasingly aware that current medical techniques were often unable to approach the systemic complexity responsible of tumor growth, and realized that mathematical modeling offers a new way to approach these difficulties (Kunz-Schughart et al., 1998). Besides, applied mathematics had demonstrated to be capable of preventing excessive experimentation, and likewise, giving valuable insight into the mechanisms that control the development of tumors (Byrne, 1999). Due to this recognition, the importance of mathematical models based in cancer systems biology has been increasing till date, being applied in several medical studies. For instance, they have been used to find the optimal way to combine anticarcinogenic drugs (Bozic et al., 2013), or to estimate the period until metastasis is detected in pancreatic cancer (Yachida et al., 2010).

A.3.2. Immunoinformatics

Immunoinformatics is the framework for the translational investigation performed in research nº3. Thanks to the recent advances in genomic and proteomic technologies such as the sequencing of human genome (explained in more detail in the section A.4.1), the volume of immunological data has increased considerably in recent years (Tong & Ren, 2009). The necessity to process these amounts of information encouraged the development of a discipline called immunoinformatics, the branch of bioinformatics that provides a mathematical framework to traditional immunology.

The origin of this area is associated to Ross's theoretical model of malaria epidemiology (explained in A.2.2), which is considered the first mathematical model of immunology. At the time Ross's model appeared, mathematics were used to study disease expansion, but since then, they have expanded to cover all other aspects of immune system processes (Tong & Ren, 2009).

Systems immunology is one of those aspects. This area combines immunoinformatics with systems biology approach (section A.3.1), and aims to construct comprehensive network models that accurately describe all of the elements regulating the immune system (Arazi et al., 2013). In this field, mathematical models are used to suggest new functional roles of metabolites, proteins and genes, by describing the potential connections among intra and intercellular components (Kidd et al., 2014). Systems immunology has been extensively applied, for example, to provide maps of the regulatory circuit controlling the differentiation of hematopoietic cell lineage (Novershtern et al., 2011), or to describe the transcriptional regulators of mouse immune system cell types (Jojic et al., 2013).

Another essential branch of immunoinformatics is the one responsible for describing the immunological system by means of combinatorial methods. The idea of using combinatorics to describe biological processes can be traced back to Mendel's and Galton's laws of heredity (section A.2.1), which highlighted the combinatorial diversity of cellular and human processes. In recent years, combinatorial techniques have been

applied in immunology to describe protein sequence diversity of dengue virus (Khan et al., 2006), measure the variability of influenza A virus proteome (Heiny et al., 2007) or find escape mutations in HIV-1 Gag protein (Peters et al., 2008), among others. Besides, genetic heritage process inspired the creation of a type of combinatorial algorithms named genetic algorithms (Holland, 1975), which are currently used for approaching numerous optimization problems.

Vaccine design, a central subject within immunology, has also been substantially influenced by mathematics since the emergence of computers. Vaccination's aim is to prime the immune system in order to generate immunological memory, so that a strong immune response will happen upon exposure to a specific pathogen (Tong & Ren, 2009). In the last decades, several mathematical algorithms have been designed to predict T and B cell epitopes (Tong et al., 2007), MHC-binding peptides (Tong et al., 2007), protein structures (Kim et al., 2004) and protein functions (Zhang, 2008). These tools have revolutionized the vaccine design process, allowing to optimize the vaccine candidates in an *in silico* stage, which is performed before *in vitro* experiments, and saves consequently a lot of time and money to researchers. In research nº3, we merge two branches of immunoinformatics, namely, combinatorial algorithms and computational vaccine design, in order to build optimal vaccine candidates (Martínez et al., 2015).

A.3.3. Quantitative diagnosis

Quantitative diagnosis is the basis of the investigation implemented in research nº4. Even if the first medical diagnosis dates before ancient Egypt (3100 B.C.), the use of quantitative techniques and instruments to detect pathological states did not spread until the 18th century. One of the first to build a quantitative diagnosis tool was Stephen Hales (section A.2.3), who made the first manometer, a device used to quantitatively estimate the capacity of the heart, blood pressure and velocity of blood current (Berger, 1999). Since then, quantitative tools have entered clinical practice, becoming a necessary piece in medical diagnosis.

One of the most extended examples is the technique known as flow cytometry. This method quantifies the nuclear DNA and has been used for instance, to diagnose leukemia (Peters & Ansari, 2011) and paroxysmal nocturnal hemoglobinuria (Borowitz et al., 2010), or to count the remaining number of CD4+ T cells in HIV infected patients (Burdo et al., 2011).

Syntactic structure analysis is another quantitative diagnostic technique that, based on graph theory, is able to provide quantitative information on tissue architecture (Van Diest et al., 1995). This method has been used, for example, to diagnose ovarian cancer (Sprindzuk et al., 2011), predict papillary thyroid carcinoma (Shih et al., 2013) or for colorectal cancer prognosis (Eynard et al., 2009).

In the obstetrics field, the use of cardiotocograph (also called electronic fetal monitor) has become the rule when assessing labor and delivery (Stout & Cahill, 2011). This device quantifies both the Fetal Heart Rate and the activity of the uterine muscle, and it was first conceived in 1960s with the purpose of decreasing morbidity and

mortality, both in the newborn and in the mother. Recently, cardiotocography recordings have been analyzed by numerical methods in order to distinguish between normal and pathological fetuses (Huhn et al., 2011; Signorini et al., 2003).

Till date, there have been two outstanding attempts to quantitatively diagnose labor by mathematical models based on uterine pressure, namely, the Montevideo Units (Caldeyro-Barcia & Poseiro, 1960) and the Alexandria units (El-Sahwi et al., 1967). Montevideo Units can be calculated multiplying the average amplitude (mm Hg) of the contractions by the average frequency (per 10 minutes). Alexandria units are an "improvement" of Montevideo Units, where the average duration (in minutes) is also considered. However, these two parameters have not been extensively used because of their lack of accuracy. In fact, the inability of Montevideo Units to predict preterm labor has been highlighted by an experimental study (Malaina et al., 2016), which is extensively explained in research nº4.

A.4. Projects derived from multidisciplinary approach.

Since the last century, the biggest projects of humanity are only conceived with a multidisciplinary approach. For instance, the Apollo Program (1969-1972) involved more than 400,000 scientists, engineers and technicians of diverse fields and culminated landing the first human in the moon. Another example of multidisciplinarity is the design of the Large Hadron Collider (1998-2008), which was developed by more than 10,000 scientists (physicists, mathematicians, chemists, engineers, etc.) from over 100 countries, and was used in 2013 to confirm the existence of Higgs boson (also known popularly as God's Particle).

A.4.1. Human Genome Project

Biomedical projects are no different, as is the case, for instance, of the Human Genome Project (1990-2003). The HGP was an international, collaborative research program whose goal was the complete understanding and mapping of the whole human genome. This project was developed by integrating mathematical and computational techniques with biological knowledge in order to sequence and comprehend massive amounts of data, and it remained as the world's largest collaborative biological project (Tripp & Grueber, 2011). It was originally publicly funded by the US Government but in the end, numerous other groups from around the world contributed to its financing.

On the other hand, Celera Corporation started a parallel private project of human genome mapping in 1998, competing with the HGP. This company managed to sequence human genome parts much more rapidly and cheaper than the public project. However, in 2000, former president Bill Clinton announced that the human genome sequence could not be patented, making Celera's stock plummet and generating $50 billion losses to the biotechnology sector. Ultimately, this company changed his policy and made their sequences available for non-commercial use, which contributed to achieving the objective of mapping the complete human genome.

In addition to sequencing the approximately 20,500 genes of human DNA (International Human Genome Sequencing Consortium, 2004), The Human Genome Project resulted in the development of several techniques that have greatly improved various disciplines such as bioinformatics or systems biology.

A.4.2. Human Brain Project

In 2005, Switzerland launched the project Blue Brain, with the purpose of simulating several parts of mammals' brains, thereby enabling the study of its behavior and the effect of possible pathologies. This research laid the foundations for the initiation of a new project at European level, namely, the Human Brain Project (HBP), which started in October 2013, and is scheduled to run for 10 years. The main objectives of the HBP are the following ones:

- Create and operate a European scientific Research Infrastructure for brain research, cognitive neuroscience, and other brain-inspired sciences.
- Gather, organize and disseminate data describing the brain and its diseases.
- Simulate the brain.
- Build multi-scale scaffold theory and models for the brain.
- Develop brain-inspired computing, data analytics and robotics.
- Ensure that the HBP's work is undertaken responsibly and that it benefits society.

This investigation relies on the collaboration of more than 750 scientist and engineers from over 24 countries, and has been founded with 1 billion Euros. In its first phase (2013-2016), this project has supported 274 publications. The HBP is a clear example of the multidisciplinary necessity, since it demands both experts in biomedicine to provide their knowledge of the performance of the brain, and specialists in quantitative sciences that can model and computerize it.

A.4.3. BRAIN Initiative

In 2009, the National Institutes of Health of the United States boosted the five-year research named Human Connectome Project (HCP), with an initial budget of $30 million. The main purpose of this project was to build a complete map of human brain's neural networks. In April 2013, and as a necessary next step of the HCP, another project in which the synergy between quantitative sciences and biomedicine is fundamental emerged: the BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies). This investigation lead by the United States had an initial budget of $110 million for its first year, and is scheduled until 2025. In the words of the former president Barack Obama, its goal is to “accelerate the development and application of

new technologies that will enable researchers to produce dynamic pictures of the brain that show how individual brain cells and complex neural circuits interact at the speed of thought". As a result of this investigation, scientists would be able to, for example, understand how we learn and memorize, or comprehend the functioning of diseases such as Alzheimer or Parkinson.

The scientific community is becoming increasingly aware that these huge projects require to be approached by multidisciplinary teams. For instance, in the case of BRAIN Initiative, this necessity has been stated as one of the core principles of the project:

"Cross boundaries in interdisciplinary collaborations. No single researcher or discovery will crack the brain's code. The most exciting approaches will bridge fields, linking experiment to theory, biology to engineering, tool development to experimental application, human neuroscience to non-human models, and more, in innovative ways." (Jorgenson et al., 2015).

Besides, there is a growing trend of investigation groups which have noticed that the integration of mathematics and medicine has the potential to achieve new medical and technological breakthroughs. As a consequence, it is expected that in the next few years this merger will lead to develop new therapies and promote health innovation, which ultimately is the key to increase the quality of medical care (Sharp & Hockfield, 2017).

In short, history shows the long path in which quantitative sciences and biomedicine converge. The future will keep moving towards a profound and continuous synergy between both disciplines, for the sake of the improvement of human health and quality of life.

References

Resumen / Introduction

- Govella, N. J., Okumu, F. O., & Killeen, G. F. (2010). Insecticide-treated nets can reduce malaria transmission by mosquitoes which feed outdoors. *The American journal of tropical medicine and hygiene*, *82*(3), 415-419.
- Huang, L. (2013). Optimization of a new mathematical model for bacterial growth. *Food Control*, *32*(1), 283-288.
- Müller, L. O., & Toro, E. F. (2014). A global multiscale mathematical model for the human circulation with emphasis on the venous system. *International journal for numerical methods in biomedical engineering*, *30*(7), 681-725.
- Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K., & Brayne, C. (2014). Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *The Lancet Neurology*, *13*(8), 788-794.
- Papp, K. A., Langley, R. G., Sigurgeirsson, B., Abe, M., Baker, D. R., Konno, P., ... & Richards, H. B. (2013). Efficacy and safety of secukinumab in the treatment of moderate-to-severe plaque psoriasis: a randomized, double-blind, placebo-controlled phase II dose-ranging study. *British Journal of Dermatology*, *168*(2), 412-421.
- Spilka, J., Abry, P., Goncalves, P., & Doret, M. (2014, September). Impacts of first and second labour stages on Hurst parameter based intrapartum fetal heart rate analysis. In *Computing in Cardiology Conference (CinC),* (pp. 777-780). IEEE.
- Wang, L., Valderramos, S. G., Wu, A., Ouyang, S., Li, C., Brasil, P., ... & Aliyari, R. (2016). From mosquitos to humans: genetic evolution of Zika virus. *Cell host & microbe*, *19*(5), 561-565.
- Xiao, Y., Miao, H., Tang, S., & Wu, H. (2013). Modeling antiretroviral drug responses for HIV-1 infected patients using differential equation models. *Advanced drug delivery reviews*, *65*(7), 940-953.

° Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., ... & Velculescu, V. E. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, *467*(7319), 1114-1117.

Research n°1: Intracellular dynamics of calcium-dependent chloride currents

- Berg, J., Yang, H., & Jan, L. Y. (2012). Ca2+-activated Cl− channels at a glance. *J Cell Sci*, *125*(6), 1367-1371.
- Caccia, D. C., Percival, D., Cannon, M. J., Raymond, G., & Bassingthwaighte, J. B. (1997). Analyzing exact fractal time series: evaluating dispersional analysis and rescaled range methods. *Physica A: Statistical Mechanics and its Applications*, *246*(3-4), 609-632.
- Cannon, M. J., Percival, D. B., Caccia, D. C., Raymond, G. M., & Bassingthwaighte, J. B. (1997). Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A: Statistical Mechanics and its Applications*, *241*(3-4), 606-626.
- Dickson, V. K., Pedi, L., & Long, S. B. (2014). Structure and insights into the function of a Ca2+-activated Cl-channel. *Nature*, *516*(7530), 213-218.
- Duan, D. D. (2013). Phenomics of cardiac chloride channels. *Comprehensive Physiology*.
- Eke, A., Herman, P., Bassingthwaighte, J., Raymond, G., Percival, D., Cannon, M., ... & Ikrényi, C. (2000). Physiological time series: distinguishing fractal noises from motions. *Pflügers Archiv*, *439*(4), 403-415.
- Endeman, D., Fahrenfort, I., Sjoerdsma, T., Steijaert, M., Ten Eikelder, H., & Kamermans, M. (2012). Chloride currents in cones modify feedback from horizontal cells to cones in goldfish retina. *The Journal of physiology*, *590*(22), 5581-5595.
- Frizzell, R. A., & Hanrahan, J. W. (2012). Physiology of epithelial chloride and fluid secretion. *Cold Spring Harbor perspectives in medicine*, *2*(6), a009563.

- Gonzalez-Silva, C., Vera, J., Bono, M. R., González-Billault, C., Baxter, B., Hansen, A., ... & Bacigalupo, J. (2013). Ca 2+-activated Cl− channels of the ClCa family express in the cilia of a subset of rat olfactory sensory neurons. *PloS one*, *8*(7), e69295.

- Hartzell, C., Putzier, I., & Arreola, J. (2005). Calcium-activated chloride channels. *Annu. Rev. Physiol.*, *67*, 719-758.

- Hoffmann, E. K., Holm, N. B., & Lambert, I. H. (2014). Functions of volume-sensitive and calcium-activated chloride channels. *IUBMB life*, *66*(4), 257-267.

- Huang, F., Wong, X., & Jan, L. Y. (2012). International Union of Basic and Clinical Pharmacology. LXXXV: calcium-activated chloride channels. *Pharmacological reviews*, *64*(1), 1-15.

- Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.*, *116*, 770-808.

- Jentsch, T. J. (2008). CLC chloride channels and transporters: from genes to protein structure, pathology and physiology. *Critical reviews in biochemistry and molecular biology*, *43*(1), 3-36.

- Jentsch, T. J., & Günther, W. (1997). Chloride channels: an emerging molecular picture. *Bioessays*, *19*(2), 117-126.

- Jentsch, T. J., Neagoe, I., & Scheel, O. (2005). CLC chloride channels and transporters. *Current opinion in neurobiology*, *15*(3), 319-325.

- Jentsch, T. J., Stein, V., Weinreich, F., & Zdebik, A. A. (2002). Molecular structure and physiological function of chloride channels. *Physiological reviews*, *82*(2), 503-568.

- Kim, M. J., Cheng, G., & Agrawal, D. K. (2004). Cl–channels are expressed in human normal monocytes: a functional role in migration, adhesion and volume change. *Clinical & Experimental Immunology*, *138*(3), 453-459.

- Li, M., Wang, Q., Lin, W., & Wang, B. (2009). Regulation of ovarian cancer cell adhesion and invasion by chloride channels. *International Journal of Gynecological Cancer*, *19*(4), 526-530.

- Mao, J., Chen, L., Xu, B., Wang, L., Wang, W., Li, M., ... & Jacob, T. J. (2009). Volume-activated chloride channels contribute to cell-cycle-dependent regulation of HeLa cell migration. *Biochemical pharmacology*, *77*(2), 159-168.

- Nilius, B., & Droogmans, G. (2003). Amazing chloride channels: an overview. *Acta Physiologica*, *177*(2), 119-147.

- O'Rourke, B. (2007). Mitochondrial ion channels. *Annu. Rev. Physiol.*, *69*, 19-49.

- Okada, Y., Shimizu, T., Maeno, E., Tanabe, S., Wang, X., & Takahashi, N. (2006). Volume-sensitive chloride channels involved in apoptotic volume decrease and cell death. *The Journal of membrane biology*, *209*(1), 21-29.

- Peng, C. K., Havlin, S., Stanley, H. E., & Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *5*(1), 82-87.

- Peretti, M., Angelini, M., Savalli, N., Florio, T., Yuspa, S. H., & Mazzanti, M. (2015). Chloride channels in cancer: focus on chloride intracellular channel 1 and 4 (CLIC1 AND CLIC4) proteins in tumor development and as novel therapeutic targets. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1848*(10), 2523-2531.

- Pifferi, S., Cenedese, V., & Menini, A. (2012). Anoctamin 2/TMEM16B: a calcium-activated chloride channel in olfactory transduction. *Experimental physiology*, *97*(2), 193-199.

- Planells-Cases, R., & Jentsch, T. J. (2009). Chloride channelopathies. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1792*(3), 173-189.

- Stanley, H. E., Afanasyev, V., Amaral, L. A. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., ... & Prince, P. A. (1996). Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. *Physica A: Statistical Mechanics and its Applications*, *224*(1-2), 302-321.

- Stauber, T., & Jentsch, T. J. (2007). Chloride in vesicular trafficking and function. *Annual review of physiology*, *75*, 453-477.

- Suzuki, M., Morita, T., & Iwamoto, T. (2006). Diversity of Cl− channels. *Cellular and molecular life sciences*, *63*(1), 12.

- Tang, C. Y., & Chen, T. Y. (2011). Physiology and pathophysiology of CLC-1: mechanisms of a chloride channel disease, myotonia. *BioMed Research International*, *2011*.

- Verkman, A. S., & Galietta, L. J. (2009). Chloride channels as drug targets. *Nature reviews Drug discovery*, *8*(2), 153-171.

- Voglis, G., & Tavernarakis, N. (2006). The role of synaptic ion channels in synaptic plasticity. *EMBO reports*, *7*(11), 1104-1110.

Research n°2: Intracellular dynamics of the adenylate energy system

- Abrusci, P., Chiarelli, L. R., Galizzi, A., Fermo, E., Bianchi, P., Zanella, A., & Valentini, G. (2007). Erythrocyte adenylate kinase deficiency: characterization of recombinant mutant forms and relationship with nonspherocytic hemolytic anemia. *Experimental hematology*, *35*(8), 1182-1189.

- Alberty, R. A., & Goldberg, R. N. (1992). Standard thermodynamic formation properties for the adenosine 5'-triphosphate series. *Biochemistry*, *31*(43), 10610-10615.

- Ataullakhanov, F. I., & Vitvitsky, V. M. (2002). What determines the intracellular ATP concentration. *Bioscience reports*, *22*(5-6), 501-511.

- Atkinson, D. E., & Walton, G. M. (1967). Adenosine triphosphate conservation in metabolic regulation rat liver citrate cleavage enzyme. *Journal of Biological Chemistry*, *242*(13), 3239-3241.

- Ball, W. J., & Atkinson, D. E. (1975). Adenylate energy charge in Saccharomyces cerevisiae during starvation. *Journal of Bacteriology*, *121*(3), 975-982.

- Blangy, D., Buc, H., & Monod, J. (1968). Kinetics of the allosteric interactions of phosphofructokinase from Escherichia coli. *Journal of molecular biology*, *31*(1), 13-35.

- Boender, L. G., Almering, M. J., Dijk, M., van Maris, A. J., de Winde, J. H., Pronk, J. T., & Daran-Lapujade, P. (2011). Extreme calorie restriction and energy source starvation in Saccharomyces cerevisiae represent distinct physiological states. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, *1813*(12), 2133-2144.

- Bønsdorff, T., Gautier, M., Farstad, W., Rønningen, K., Lingaas, F., & Olsaker, I. (2004). Mapping of the bovine genes of the de novo AMP synthesis pathway1. *Animal genetics*, *35*(6), 438-444.

° 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*(3), 189-194.

° Burnstock, G. (2012). Discovery of purinergic signalling, the initial resistance and current explosion of interest. *British journal of pharmacology*, *167*(2), 238-255.

° Chapman, A. G., & Atkinson, D. E. (1977). Adenine nucleotide concentrations and turnover rates. Their correlation with biological activity in bacteria and yeast. *Advances in microbial physiology*, *15*, 253-306.

° Chapman, A. G., Fall, L., & Atkinson, D. E. (1971). Adenylate energy charge in Escherichia coli during growth and starvation. *Journal of Bacteriology*, *108*(3), 1072-1086.

° 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. *Biomedical Chromatography*, *21*(6), 628-634.

° Ching, K. K. (1972). Content of adenosine phosphates and adenylate energy charge in germinating ponderosa pine seeds. *Plant physiology*, *50*(5), 536-540.

° Curien, G., Bastien, O., Robert-Genthon, M., Cornish-Bowden, A., Cárdenas, M. L., & Dumas, R. (2009). Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters. *Molecular systems biology*, *5*(1), 271.

° De la Fuente, I. M. (1999b). Diversity of temporal self-organized behaviors in a biochemical system. *BioSystems*, *50*(2), 83-97.

° De la Fuente, I. M. (2010b). Quantitative analysis of cellular metabolic dissipative, self-organized structures. *International journal of molecular sciences*, *11*(9), 3540-3599.

° De la Fuente, I. M. (2014). Metabolic dissipative structures. In *Systems Biology of Metabolic and Signaling Networks* (pp. 179-211). Springer Berlin Heidelberg.

° De la Fuente, I. M., & Cortes, J. M. (2012). Quantitative analysis of the effective functional structure in yeast glycolysis. *PLoS One*, *7*(2), e30162.

° De la Fuente, I. M., Benítez, N., Santamaría, A., Aguirregabiria, J. M., & Veguillas, J. (1999a). Persistence in metabolic nets. *Bulletin of mathematical biology*, *61*(3), 573-595.

° De la Fuente, I. M., Cortes, J. M., Pelta, D. A., & Veguillas, J. (2013). Attractor metabolic networks. *PLoS One*, *8*(3), e58284.

° De la Fuente, I. M., Cortes, J. M., Perez-Pinilla, M. B., Ruiz-Rodriguez, 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*(11), e27224.

° De la Fuente, I. M., Martinez, L., & Veguillas, J. (1996b). Intermittency route to chaos in a biochemical system. *Biosystems*, *39*(2), 87-92.

° De la Fuente, I. M., Martínez, L., Aguirregabiria, J. M., & Veguillas, J. (1998). Coexistence of multiple periodic and chaotic regimes in biochemical oscillations with phase shifts. *Acta Biotheoretica*, *46*(1), 37-51.

° De la Fuente, I. M., Martínez, L., Pérez-Samartín, A. L., Ormaetxea, L., Amezaga, C., & Vera-López, A. (2008). Global self-organization of the cellular metabolic structure. *Plos One*, *3*(8), e3100.

° De la Fuente, I. M., Martinez, L., Veguillas, J., & Aguirregabiria, J. M. (1996a). Quasiperiodicity route to chaos in a biochemical system. *Biophysical journal*, *71*(5), 2375-2379.

° De la Fuente, I. M., 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.

° De la Fuente, I. M., Vadillo, F., Pérez-Samartín, A. L., Pérez-Pinilla, M. B., Bidaurrazaga, J., & Vera-López, A. (2010a). Global self-regulation of the cellular metabolic structure. *PLoS One*, *5*(3), e9484.

° Edwards, J. M., Roberts, T. H., & Atwell, B. J. (2012). Quantifying ATP turnover in anoxic coleoptiles of rice (Oryza sativa) demonstrates preferential allocation of energy to protein synthesis. *Journal of experimental botany*, *63*(12), 4389-4402.

° Ellis, R. J. (2001). Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Current opinion in structural biology*, *11*(1), 114-119.

° Erlinge, D. (2010). Purinergic and pyriminergic activation of the endothelium in regulation of tissue perfusion. In *Extracellular ATP and Adenosine as Regulators of Endothelial Cell Function* (pp. 1-13). Springer Netherlands.

° Falzoni, S., Donvito, G., & Di Virgilio, F. (2013). Detecting adenosine triphosphate in the pericellular space. *Interface Focus*, *3*(3), 20120101.

° Forsyth, A. M., Wan, J., Owrutsky, P. D., Abkarian, M., & Stone, H. A. (2011). Multiscale approach to link red blood cell dynamics, shear viscosity, and ATP release. *Proceedings of the National Academy of Sciences*, *108*(27), 10986-10991.

° Getty-Kaushik, L., Richard, A. M. T., & Corkey, B. E. (2005). Free fatty acid regulation of glucose-dependent intrinsic oscillatory lipolysis in perifused isolated rat adipocytes. *Diabetes*, *54*(3), 629-637.

° Goldbeter, A. (1974). Modulation of the adenylate energy charge by sustained metabolic oscillations. *FEBS letters*, *43*(3), 327-330.

° Goldbeter, A., & Lefever, R. (1972). Dissipative structures for an allosteric model: application to glycolytic oscillations. *Biophysical Journal*, *12*(10), 1302-1315.

- Goldbeter, A., & Prigogine, I. (1990). *Rythmes et chaos: dans les systèmes biochimiques et cellulaires*. Masson.
- Hagen, J. (2006) Industrial catalysis: A practical approach. Weinheim, Germany: Wiley-VCH.
- Hardie, D. G. (2011). Signal transduction: how cells sense energy. *Nature*, *472*(7342), 176-177.
- Holz, G. G., Heart, E., & Leech, C. A. (2008). Synchronizing Ca2+ and cAMP oscillations in pancreatic β-cells: a role for glucose metabolism and GLP-1 receptors? Focus on "Regulation of cAMP dynamics by Ca2+ and G protein-coupled receptors in the pancreatic β-cell: a computational approach". *American Journal of Physiology-Cell Physiology*, *294*(1), C4-C6.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*(6804), 651-654.
- Jin, J., Dong, W., & Guarino, L. A. (1998). The LEF-4 subunit of baculovirus RNA polymerase has RNA 5′-triphosphatase and ATPase activities. *Journal of virology*, *72*(12), 10011-10019.
- Knowles, J. R. (1980). Enzyme-catalyzed phosphoryl transfer reactions. *Annual review of biochemistry*, *49*(1), 877-919.
- Lim, E. L., Hollingsworth, K. G., Thelwall, P. E., & 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 in biomedicine*, *23*(8), 952-957.
- Lloyd, D., & Murray, D. B. (2005). Ultradian metronome: timekeeper for orchestration of cellular coherence. *Trends in biochemical sciences*, *30*(7), 373-377.
- Lloyd, D., & Murray, D. B. (2006). The temporal architecture of eukaryotic growth. *FEBS letters*, *580*(12), 2830-2835.
- Marquez, S., Crespo, P., Carlini, V., Garbarino-Pico, E., Baler, R., Caputto, B. L., & Guido, M. E. (2004). The metabolism of phospholipids oscillates rhythmically in cultures of fibroblasts and is regulated by the clock protein PERIOD 1. *The FASEB journal*, *18*(3), 519-521.
- Moran, L. A., Horton, R. A., Scrimgeour, G., Perry, M. (2011) Principles of biochemistry (5th Edition). New Jersey: Prentice Hall.
- Mori, Y., Matsumoto, K., Ueda, T., & Kobatake, Y. (1986). Spatio-temporal organization of intracellular ATP content and oscillation patterns in response to blue light byPhysarum polycephalum. *Protoplasma*, *135*(1), 31-37.
- Murray, D. B., Beckmann, M., & Kitano, H. (2007). Regulation of yeast oscillatory dynamics. *Proceedings of the National Academy of Sciences*, *104*(7), 2241-2246.

- Nath, S., & Jain, S. (2000). Kinetic modeling of ATP synthesis by ATP synthase and its mechanistic implications. *Biochemical and biophysical research communications*, *272*(3), 629-633.

- Nelson, D. L., Lehninger, A. L., & Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

- Nenninger, A., Mastroianni, G., & Mullineaux, C. W. (2010). Size dependence of protein diffusion in the cytoplasm of Escherichia coli. *Journal of bacteriology*, *192*(18), 4535-4540.

- Özalp, V. C., Pedersen, T. R., Nielsen, L. J., & Olsen, L. F. (2010). Time-resolved measurements of intracellular ATP in the yeast Saccharomyces cerevisiae using a new type of nanobiosensor. *Journal of Biological Chemistry*, *285*(48), 37579-37588.

- Petty, H. R., & Kindzelskii, A. L. (2001). Dissipative metabolic patterns respond during neutrophil transmembrane signaling. *Proceedings of the National Academy of Sciences*, *98*(6), 3145-3149.

- Privalle, L. S., & Burris, R. H. (1983). Adenine nucleotide levels in and nitrogen fixation by the cyanobacterium Anabaena sp. strain 7120. *Journal of bacteriology*, *154*(1), 351-355.

- Rakotonirainy, M. S., & 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*(3), 182-186.

- Rapoport, T. A., Heinrich, R., & Rapoport, S. M. (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. *Biochemical Journal*, *154*(2), 449-469.

- Reich, J. G., & Sel'Kov, E. E. (1974). Mathematical analysis of metabolic networks. FEBS letters, 40, S112-S118.

- Rengan, R., & Omann, G. M. (1999). Regulation of oscillations in filamentous actin content in polymorphonuclear leukocytes stimulated with leukotriene B4 and platelet-activating factor. *Biochemical and biophysical research communications*, *262*(2), 479-486.

- Richard, P., Teusink, B., Hemker, M. B., Van Dam, K. A. R. E. L., & Westerhoff, H. V. (1996). Sustained oscillations in free-energy state and hexose phosphates in yeast. *Yeast*, *12*(8), 731-740.

- Rosenspire, A. J., Kindzelskii, A. L., & Petty, H. R. (2001). Pulsed DC electric fields couple to natural NAD (P) H oscillations in HT-1080 fibrosarcoma cells. *Journal of cell science*, *114*(8), 1515-1520.

- Sear, R. P. (2005). The cytoplasm of living cells: a functional mixture of thousands of components. *Journal of Physics: Condensed Matter*, *17*(45), S3587.

- Sel'kov, E. E. (1968). Self-oscillations in glycolysis. 1. A simple kinetic model. Eur. J. Biochem, 4, 79-86.

- Sel'kov, E. E. (1975). Stabilization of energy charge, generation of oscillations and multiple steady states in energy metabolism as a result of purely stoichiometric regulation. European Journal of Biochemistry, 59(1), 151-157.

- Shankaran, H., Ippolito, D. L., Chrisler, W. B., Resat, H., Bollinger, N., Opresko, L. K., & Wiley, H. S. (2009). Rapid and sustained nuclear–cytoplasmic ERK oscillations induced by epidermal growth factor. *Molecular systems biology*, *5*(1), 332.

- Sheng, X. R., Li, X., & Pan, X. M. (1999). An iso-random Bi Bi mechanism for adenylate kinase. *Journal of Biological Chemistry*, *274*(32), 22238-22242.

- Soga, N., Kinosita, K., Yoshida, M., & Suzuki, T. (2011). Efficient ATP synthesis by thermophilic Bacillus FoF1-ATP synthase. *FEBS journal*, *278*(15), 2647-2654.

- Steigmiller, S., Turina, P., & Gräber, P. (2008). The thermodynamic H+/ATP ratios of the H+-ATPsynthases from chloroplasts and Escherichia coli. *Proceedings of the National Academy of Sciences*, *105*(10), 3745-3750.

- Suska, M., & Skotnicka, E. (2009). Changes in Adenylate Nucleotides Concentration and Na. *Veterinary medicine international*, *2010*.

- Swedes, J. S., Sedo, R. J., & Atkinson, D. E. (1975). Relation of growth and protein synthesis to the adenylate energy charge in an adenine-requiring mutant of Escherichia coli. *Journal of Biological Chemistry*, *250*(17), 6930-6938.

- Tanaka, K., Gilroy, S., Jones, A. M., & Stacey, G. (2010). Extracellular ATP signaling in plants. *Trends in cell biology*, *20*(10), 601-608.

- Thuma, E., Schirmer, R. H., & Schirmer, I. (1972). Preparation and characterization of a crystalline human ATP: AMP phosphotransferase. *Biochimica et Biophysica Acta (BBA)-Enzymology*, *268*(1), 81-91.

- 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. *Experimental cell research*, *169*(1), 191-201.

- Valero, E., Varón, R., & García-Carmona, F. (2006). A kinetic study of a ternary cycle between adenine nucleotides. *FEBS Journal*, *273*(15), 3598-3613.

- Versées, W., & Steyaert, J. (2003). Catalysis by nucleoside hydrolases. *Current opinion in structural biology*, *13*(6), 731-738.

- Walker-Simmons, M. A. R. Y., & Atkinson, D. E. (1977). Functional capacities and the adenylate energy charge in Escherichia coli under conditions of nutritional stress. *Journal of bacteriology*, *130*(2), 676-683.

- 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*(3), 707-716.

° Ytting, C. K., Fuglsang, A. T., Hiltunen, J. K., Kastaniotis, A. J., Özalp, V. C., Nielsen, L. J., & Olsen, L. F. (2012). Measurements of intracellular ATP provide new insight into the regulation of glycolysis in the yeast Saccharomyces cerevisiae. *Integrative Biology*, *4*(1), 99-107.

° Zuo, P. (2007). *Modeling the airway surface liquid regulation in human lungs*. ProQues.

Research n°3: Vaccine design though combinatorial methods

- Allegrini, P., Buiatti, M., Grigolini, P., & West, B. J. (1998). Fractional Brownian motion as a nonstationary process: An alternative paradigm for DNA sequences. *Physical Review E*, *57*(4), 4558.
- 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. *Journal of Biological Physics*, *30*(1), 33-81.
- De la Fuente, I. M., Martínez, L., Benitez, N., Veguillas, J., & Aguirregabiria, J. M. (1998). Persistent behavior in a phase-shift sequence of periodical biochemical oscillations. *Bulletin of mathematical biology*, *60*(4), 689-702.
- De la Fuente, I. M., 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., ... & Hahn, B. H. (2007). Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nature medicine*, *13*(1), 100-106.
- GenBank (2013). http://www.ncbi.nlm.nih.gov/genbank/ (Online; Accessed 21 Sept 2013)
- Giles, B. M., & Ross, T. M. (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. *Biochemical Oscillations and Cellular Rhythms, by Albert Goldbeter, Foreword by MJ Berridge, Cambridge, UK: Cambridge University Press, 1997*.

- Henry-Labordere, A. (1969). Record balancing problem-a dynamic programming solution of a generalized traveling salesman problem. *Revue Francaise D Informatique De Recherche Operationnelle*, *3*(NB 2), 43.
- HIV Molecular Immunology Database (2013). http://www.hiv.lanl.gov/content/immunology/ (Online; Accessed 21 Sept 2013)
- Ibm, ILOG CPLEX Optimization Studio (2013). http://www-03.ibm.com/software/products/us/en/ibmilogcpleoptistud/ (Online; Accessed 21 Sept 2013)
- Java (2013). http://www.java.com (Online; Accessed 21 Sept 2013)
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & 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. (2006). Using "epitomes" to model genetic diversity: Rational design of HIV vaccine cocktails. *Advances in Neural Information Processing Systems*, *18*, 587.
- Jones, N. C., & Pevzner, P. (2004). *An introduction to bioinformatics algorithms*. MIT press.
- Kazachenko, V. N., Astashev, M. E., & Grinevich, A. A. (2007). Multifractal analysis of K+ channel activity. *Biochemistry (Moscow) Supplement Series A: Membrane and Cell Biology*, *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, A. K., Yan, J., ... & Le Gall, S. (2013). HIV-1 p24 gag 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, September). Computability of models for sequence assembly. In *International Workshop on Algorithms in Bioinformatics* (pp. 289-301). Springer Berlin Heidelberg.
- Miller, C. E., Tucker, A. W., & Zemlin, R. A. (1960). Integer programming formulation of traveling salesman problems. *Journal of the ACM (JACM)*, *7*(4), 326-329.
- Nickle, D. C., Rolland, M., Jensen, M. A., Pond, S. L. K., Deng, W., Seligman, M., ... & Jojic, N. (2007). Coping with viral diversity in HIV vaccine design. *PLoS Comput Biol*, *3*(4), e75.
- O'Neill, E., Kuo, L. S., Krisko, J. F., Tomchick, D. R., Garcia, J. V., & Foster, J. L. (2006). Dynamic evolution of the human immunodeficiency virus type 1 pathogenic factor, Nef. *Journal of virology*, *80*(3), 1311-1320.
- Saksena, J. P. (1970) Mathematical model of scheduling clients through welfare agencies. CORS J 8:185–200

° Srivastava, S. S., Kumar, S., Garg, R. C., & Sen, P. (1969). Generalized traveling salesman problem through n sets of nodes. *CORS journal*, *7*, 97-101.

° Tarhio, J., & Ukkonen, E. (1988). A greedy approximation algorithm for constructing shortest common superstrings. *Theoretical computer science*, *57*(1), 131-145.

° Toussaint, N. C., Dönnes, 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.

Research n°4: Preterm labor prediction by autoregressive models

- Alvarez, H., & Caldeyro-Barcia, R. (1954). The normal and abnormal contractile waves of the uterus during labour. *Gynecologic and Obstetric Investigation*, *138*(2), 190-212.
- Berghella, V., Baxter, J. K., & Hendrix, N. W. (2009). Cervical assessment by ultrasound for preventing preterm delivery. *The Cochrane Library*.
- Caldeyro-Barcia, R., Pose, S. V., & Alvarez, H. (1957). Uterine contractility in polyhydramnios and the effects of withdrawal of the excess of amniotic fluid. *American journal of obstetrics and gynecology*, *73*(6), 1238-1254.
- Copper, R. L., Goldenberg, R. L., Das, A., Elder, N., Swain, M., Norman, G., ... & Jones, P. (1996). The preterm prediction study: Maternal stress is associated with spontaneous preterm birth at less than thirty-five weeks' gestation. *American journal of obstetrics and gynecology*, *175*(5), 1286-1292.
- Esplin, M. S., O'Brien, E., Fraser, A., Kerber, R. A., Clark, E., Simonsen, S. E., ... & Varner, M. W. (2008). Estimating recurrence of spontaneous preterm delivery. *Obstetrics & Gynecology*, *112*(3), 516-523.
- Georgiou, H. M., Di Quinzio, M. K., Permezel, M., & Brennecke, S. P. (2015). Predicting preterm labour: current status and future prospects. *Disease markers*, *2015*.
- Goldenberg, R. L., Iams, J. D., Mercer, B. M., Meis, P. J., Moawad, A. H., Copper, R. L., ... & Miodovnik, M. (1998). 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*, *88*(2), 233-238.
- Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm birth. *The lancet*, *371*(9606), 75-84.

- Honest, H., Forbes, C. A., Durée, K. H., Norman, G., Duffy, S. B., Tsourapas, A., ... & Khan, K. S. (2009). Screening to prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with economic modelling.

- Iams, J. D., Goldenberg, R. L., Meis, P. J., Mercer, B. M., Moawad, A., Das, A., ... & Roberts, J. M. (1996). The length of the cervix and the risk of spontaneous premature delivery. *New England Journal of Medicine*, *334*(9), 567-573.

- Lammers, W. J. (2013). The electrical activities of the uterus during pregnancy. *Reproductive Sciences*, *20*(2), 182-189.

- Leitich, H., Egarter, C., Kaider, A., Hohlagschwandtner, M., Berghammer, P., & Husslein, P. (1999). Cervicovaginal fetal fibronectin as a marker for preterm delivery: a meta-analysis. *American journal of obstetrics and gynecology*, *180*(5), 1169-1176.

- Malaina, I., Matorras, R., Fernandez-LLebrez, L., Bringas, C., Aranburu, L., Gonzalez, L., Arana, I. & De La Fuente, I.M. (2016). Precocious diagnosis of preterm labor immediacy by Autoregressive Integrated Moving Average Models. *IWBBIO Proceedings 2016*, 263-274.

- Newman, R. B., Goldenberg, R. L., Iams, J. D., Meis, P. J., Mercer, B. M., Moawad, A. H., ... & Thurnau, G. R. (2008). Preterm prediction study: comparison of the cervical score and Bishop score for prediction of spontaneous preterm delivery. *Obstetrics and gynecology*, *112*(3), 508.

- Paternoster, M. D., Muresan, D., Vitulo, A., Serena, A., Battagliarin, G., Dell'Avanzo, M., & Nicolini, U. (2007). Cervical phIGFBP-1 in the evaluation of the risk of preterm delivery. *Acta obstetricia et gynecologica Scandinavica*, *86*(2), 151-155.

- Revah, A., Sue-A-Quan, A., & Hannah, M. (1997). Fetal fibronectin as a predictor of preterm birth: a systematic review of the literature. *American Journal of Obstetrics and Gynecology*, *176*(1), S53.

- Smith, R., Imtiaz, M., Banney, D., Paul, J. W., & Young, R. C. (2015). Why the heart is like an orchestra and the uterus is like a soccer crowd. *American journal of obstetrics and gynecology*, *213*(2), 181-185.

- Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical length measurement for prediction of preterm birth in women with threatened preterm labor: a meta-analysis. *Ultrasound in Obstetrics & Gynecology*, *35*(1), 54-64.

- Stout, M. J., & Cahill, A. G. (2011). Electronic fetal monitoring: past, present, and future. *Clinics in perinatology*, *38*(1), 127-142.

- Tong, W. C., Choi, C. Y., Kharche, S., Holden, A. V., Zhang, H., & Taggart, M. J. (2011). Correction: A Computational Model of the Ionic Currents, Ca2+ Dynamics and Action Potentials Underlying Contraction of Isolated Uterine Smooth Muscle. *PloS one*, *6*(10).

° Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a. *Quetelet*, *10*, 113-121.

° Zhang, M., Tidwell, V., La Rosa, P. S., Wilson, J. D., Eswaran, H., & Nehorai, A. (2016). Modeling magnetomyograms of uterine contractions during pregnancy using a multiscale forward electromagnetic approach. *PloS one*, *11*(3), e0152421.

Annex: Mathematical applications to biomedicine through history, a brief summary

- Alvargonzález, D. (2011). Multidisciplinarity, interdisciplinarity, transdisciplinarity, and the sciences. *International Studies in the Philosophy of Science*, *25*(4), 387-403.
- Arazi A, Pendergraft W, Ribeiro R, Perelson A, Hacohen N. Human systems immunology: hypothesis-based modeling and unbiased data-driven approaches. Semin. Immunol. 2013; 25:193–200. [PubMed: 23375135]
- Bellomo, N., Angelis, E., & Preziosi, L. (2004). Multiscale modeling and mathematical problems related to tumor evolution and medical therapy. *Journal of Theoretical Medicine,* 5(2), 111-136.
- Berger, D. (1999). A brief history of medical diagnosis and the birth of the clinical laboratory. Part 1—Ancient times through the 19th century. *MLO Med Lab Obs*, *31*(7), 28-30.
- Bernoulli, D. (1738). Hydrodynamica. *Dulsecker. Consultable en ligne http://imgbase-scd-ulp. u-strasbg. fr/displayimage. php*.
- Borowitz, M. J., Craig, F. E., DiGiuseppe, J. A., Illingworth, A. J., Rosse, W., Sutherland, D. R., ... & Richards, S. J. (2010). Guidelines for the diagnosis and monitoring of paroxysmal nocturnal hemoglobinuria and related disorders by flow cytometry. *Cytometry Part B: Clinical Cytometry*, *78*(4), 211-230.
- Bozic, I., Reiter, J. G., Allen, B., Antal, T., Chatterjee, K., Shah, P., ... & Lipson, E. J. (2013). Evolutionary dynamics of cancer in response to targeted combination therapy. *Elife*, *2*, e00747.
- Bronowski, J. (1978). *The common sense of science*. Harvard University Press.
- Brown, R. (1833). XXXV. On the Organs and Mode of Fecundation in Orchideæ and Asclepiadeæ. *Transactions of the Linnean Society of London*, *16*(3), 685-738.

- Burdo, T. H., Lo, J., Abbara, S., Wei, J., DeLelys, M. E., Preffer, F., ... & Grinspoon, S. (2011). Soluble CD163, a novel marker of activated macrophages, is elevated and associated with noncalcified coronary plaque in HIV-infected patients. *Journal of Infectious Diseases*, *204*(8), 1227-1236.

- Burton, A. C. (1966). Rate of growth of solid tumours as a problem of diffusion. *Growth* 30, 157–176.

- Byrne, H. M. (1999). Using mathematics to study solid tumour growth, in Proceedings of the 9th General Meetings of European Women in Mathematics, pp. 81–107

- Caldeyro-Barcia, R., & Poseiro, J. J. (1960). PHYSIOLOGY OF THE UTERINE CONTRACTION. *Clinical Obstetrics and Gynecology*, *3*(2), 386-410.

- Chen, C. C., DaPonte, J. S., & Fox, M. D. (1989). Fractal feature analysis and classification in medical imaging. *IEEE transactions on medical imaging*, *8*(2), 133-142.

- Choi, B. C., & Pak, A. W. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and investigative medicine*, *29*(6), 351.

- Chuck, A. S., Clarke, M. F., & Palsson, B. O. (1996). Retroviral infection is limited by Brownian motion. *Human gene therapy*, *7*(13), 1527-1534.

- Cusanus, N. (1450). Idiota de Staticis Experimentis, Dialogus. *Codex Cusanus*, *1456*(64).

- De la Fuente, I. M., Cortés, J. M., Valero, E., Desroches, M., Rodrigues, S., Malaina, I., & Martínez, L. (2014). On the dynamics of the adenylate energy system: homeorhesis vs homeostasis. *PloS one*, *9*(10), e108676.

- De la Fuente, I. M., Malaina, I., Pérez-Samartín, A., Boyano, M. D., Pérez-Yarza, G., Bringas, C., ... & Martínez, L. (2017). Dynamic properties of calcium-activated chloride currents in Xenopus laevis oocytes. Scientific Reports, 7.

- Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der physik*, *322*(8), 549-560.

- El-Sahwi, S., Gaafar, A. A., & Toppozada, H. K. (1967). A new unit for evaluation of uterine activity. *American journal of obstetrics and gynecology*, *98*(7), 900-903.

- Eynard, H. G., Soria, E. A., Cuestas, E., Rovasio, R. A., & Eynard, A. R. (2009). Assessment of colorectal cancer prognosis through nuclear morphometry. *Journal of Surgical Research*, *154*(2), 345-348.

- Fick, A. (1855). Ueber diffusion. *Annalen der Physik*, *170*(1), 59-86.

- Fick, A. (1870). Uber die messung des Blutquantums in den Hertzventrikeln. Sitz der Physik. *Med Ges Wurzburg*, *16*.
- Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Royal Society of Edinburgh*, *52*, 399-433.
- Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3-32.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.
- Galilei, G. (1623). Il Saggiatore, nel quale con bilancia esquisita e giusta. *Rome: Mascardi*
- Galton, F. (1874). *Notes and queries on anthropology: for the use of travellers and residents in uncivilized lands*. E. Stanford.
- Galton, F. (1875). IV. Statistics by intercomparison, with remarks on the law of frequency of error. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *49*(322), 33-46.
- Galton, F. (1877). Typical laws of heredity. *Nature, 15*(388, 389, 390): 492-495,512-514,532-533.
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246-263.
- Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, *45*(273-279), 135-145.
- Galton, F. (1889). Natural Inheritance. Macmillan, London.
- Grief, A. D., & Richardson, G. (2005). Mathematical modelling of magnetically targeted drug delivery. *Journal of Magnetism and Magnetic Materials*, *293*(1), 455-463.
- Hales, S. (1733). Statistical essays: Concerning haemastaticks; or, an account of some hydraulick and hydrostatical experiments made on the blood and blood-vessels of animals. *Published by W Innys and R Manby, London*, *1733*.
- Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology* (Vol. 239). Oxford: Oxford university press.
- Heiny, A. T., Miotto, O., Srinivasan, K. N., Khan, A. M., Zhang, G. L., Brusic, V., ... & August, J. T. (2007). Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. *PloS one*, *2*(11), e1190.
- Holland, J. H. (1975). Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. *Ann Arbor, MI: University of Michigan Press*.

- Huhn, E. A., Lobmaier, S., Fischer, T., Schneider, R., Bauer, A., Schneider, K. T., & Schmidt, G. (2011). New computerized fetal heart rate analysis for surveillance of intrauterine growth restriction. *Prenatal diagnosis*, *31*(5), 509-514.

- International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931-945.

- Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., ... & Lesnick, T. G. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, *12*(2), 207-216.

- Jojic, V., Shay, T., Sylvia, K., Zuk, O., Sun, X., Kang, J., ... & Immunological Genome Project Consortium. (2013). Identification of transcriptional regulators in the mouse immune system. *Nature immunology*, *14*(6), 633-643.

- Jorgenson, L. A., Newsome, W. T., Anderson, D. J., Bargmann, C. I., Brown, E. N., Deisseroth, K., ... & Marder, E. (2015). The BRAIN Initiative: developing technology to catalyse neuroscience discovery. Phil. Trans. R. Soc. B, 370(1668), 20140164.

- Khan, A. M., Heiny, A. T., Lee, K. X., Srinivasan, K. N., Tan, T. W., August, J. T., & Brusic, V. (2006). Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus. *BMC bioinformatics*, *7*(5), S4.

- Kidd, B. A., Peters, L. A., Schadt, E. E., & Dudley, J. T. (2014). Unifying immunology with informatics and multiscale biology. *Nature immunology*, *15*(2), 118-127.

- Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic acids research*, *32*(suppl 2), W526-W531.

- Kitano, H. (2000). Perspectives on systems biology. *New Generation Computing*, *18*(3), 199-216.

- Kitano, H. (2002). Systems biology: a brief overview. *Science*, *295*(5560), 1662-1664.

- Kolmogorov, A. N. (1958). Kibernetika. *inBol'shaia Sovetskaia entsiklopediia*, 324-328.

- Krogh, A. (1919). The number and distribution of capillaries in muscles with calculations of the oxygen pressure head necessary for supplying the tissue. *J. Physiol,* 52, 409–415

- Kuenen, M. P., Herold, I. H., Korsten, H. H., de la Rosette, J. J., Wijkstra, H., & Mischi, M. (2014). Maximum-likelihood estimation for indicator dilution analysis. *IEEE Transactions on Biomedical Engineering*, *61*(3), 821-831.

- Kunz-Schughart, L. A., Kreutz, M., & Knuechel, R. (1998). Multicellular spheroids: a three-dimensional in vitro culture system to study tumour biology. *International journal of experimental pathology*, *79*(1), 1-23.

° Lancaster, H. O. (1956). Some geographical aspects of the mortality from melanoma in Europeans. *The Medical Journal of Australia*, *43*(26), 1082-1087.

° Lancaster, H. O. (1994). *Quantitative methods in biological and medical sciences*. New York etc: Springer.

° Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes* (No. 1). F. Didot.

° Liotta, L. A., J. Kleinerman & G. M. Saidel (1974). Quantitative relationships of intravascular tumor cells, tumor vessels, and pulmonary metastases following tumor implantation. *Cancer Res,* 34, 997–1004.

° Malaina, I., Matorras, R., Martínez, L., Fernandez-LLebrez, L., Bringas, C., Aranburu, L. & De La Fuente, I.M. (2016). Montevideo Units Vs Autoregressive Models on Preterm Labor Detection. *ITISE Proceedings 2016*, 799-807.

° Martínez, L., Milanič, M., Legarreta, L., Medvedev, P., Malaina, I., & Ildefonso, M. (2015). A combinatorial approach to the design of vaccines. Journal of mathematical biology, 70(6), 1327-1358.

° Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.

° Moore, N. (1908). *The History of the Study of Medicine in the British Isles: The Fitz-Patrick Lectures for 1905-6, Delivered Before the Royal College of Physicians of London*. Clarendon Press.

° Namimoto, T., Yamashita, Y., Sumi, S., Tang, Y., & Takahashi, M. (1997). Focal liver masses: characterization with diffusion-weighted echo-planar MR imaging. *Radiology*, *204*(3), 739-744.

° Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., ... & Frampton, G. M. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, *144*(2), 296-309.

° Onuk, A. E., Akcakaya, M., Bardhan, J., Erdogmus, D., Brooks, D. H., & Makowski, L. (2015). Constrained Maximum Likelihood Estimation of the Abundances of Protein Conformation in a Heterogeneous Structural Ensemble from Small Angle X-ray Scattering Intensity Measurements. *Biophysical Journal*, *108*(2), 210a.

° Peters, H. O., Mendoza, M. G., Capina, R. E., Luo, M., Mao, X., Gubbins, M., ... & Wachihi, C. (2008). An integrative bioinformatic approach for studying escape mutations in human immunodeficiency virus type 1 gag in the Pumwani Sex Worker Cohort. *Journal of virology*, *82*(4), 1980-1992.

° Peters, J. M., & Ansari, M. Q. (2011). Multiparameter flow cytometry in the diagnosis and management of acute leukemia. *Archives of pathology & laboratory medicine*, *135*(1), 44-54.

- Plackett, R. L. (1988). Data analysis before 1750. *International Statistical Review/Revue Internationale de Statistique*, 181-195.
- Pott, P. (1775). *Chirurgical observations relative to the cataract the polypus of the nose, the cancer of the scrotum, the different kinds of ruptures, and the mortification of the toes and feet, by Percivall Pott, FRS*. TJ Carnegy.
- Poulton, B. C., & West, M. A. (1993). Effective multidisciplinary teamwork in primary health care. *Journal of advanced nursing*, *18*(6), 918-925.
- Pradeu, T., Jaeger, S., & Vivier, E. (2013). The speed of change: towards a discontinuity theory of immunity?. *Nature Reviews Immunology*, *13*(10), 764-769.
- Richardson, L. F. (1926). Atmospheric diffusion shown on a distance-neighbour graph. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *110*(756), 709-737.
- Ross, R. (1915). Some a priori pathometric equations. *British medical journal*, *1*(2830), 546.
- Ross, R. (1916). An application of the theory of probabilities to the study of a priori pathometry. Part I. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, *92*(638), 204-230.
- Schmidt, J. C. (2008). Towards a philosophy of interdisciplinarity. *Poiesis & Praxis*, *5*(1), 53-69.
- Sharp, P., & Hockfield, S. (2017). Convergence: The future of health. Science, 355(6325), 589-589.
- Shih, S. R., Chang, Y. C., Li, H. Y., Liau, J. Y., Lee, C. Y., Chen, C. M., & Chang, T. C. (2013). Preoperative prediction of papillary thyroid carcinoma prognosis with the assistance of computerized morphometry of cytology samples obtained by fine-needle aspiration: Preliminary report. *Head & neck*, *35*(1), 28-34.
- Sigler, L. (2003). *Fibonacci's Liber abaci: a translation into modern English of Leonardo Pisano's Book of calculation*. Springer Science & Business Media.
- Signorini, M. G., Magenes, G., Cerutti, S., & Arduini, D. (2003). Linear and nonlinear parameters for the analysisof fetal heart rate signal from cardiotocographic recordings. *IEEE Transactions on Biomedical Engineering*, *50*(3), 365-374.
- Sprindzuk, M., Dmitruk, A., Kovalev, V., Bogush, A., Tuzikov, A., Liakhovski, V., & Fridman, M. (2011). Computer-aided image processing of angiogenic histological samples in ovarian cancer. *Journal of clinical medicine research*, *1*(5), 249-261.
- Stokols, D., Harvey, R., Gress, J., Fuqua, J., & Phillips, K. (2005). In vivo studies of transdisciplinary scientific collaboration: lessons learned and implications for active living research. *American journal of preventive medicine*, *28*(2), 202-213.

- Stout, M. J., & Cahill, A. G. (2011). Electronic fetal monitoring: past, present, and future. *Clinics in perinatology*, *38*(1), 127-142.
- Thiele, T. N. (1880). Om Anvendelse af mindste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfældige Fejlkilder giver Fejlene en 'systematisk'Karakter. *Det Kongelige Danske Videnskabernes Selskabs Skrifter-Naturvidenskabelig og Mathematisk Afdeling*, 381-408.
- Thiele, T. N. (1889). *Forelaesninger over almindelig lagttagelseslaere: sandsynlighedsregning og mindste Kvadraters methode*. I kommision hos CA Reitzel.
- Tong, J. C., & Ren, E. C. (2009). Immunoinformatics: current trends and future directions. *Drug discovery today*, *14*(13), 684-689.
- Tong, J. C., Tan, T. W., & Ranganathan, S. (2007). Methods and protocols for prediction of immunogenic epitopes. *Briefings in Bioinformatics*, *8*(2), 96-108.
- Tripp, S., & Grueber, M. (2011). Economic impact of the human genome project. *Battelle Memorial Institute*, 4-7.
- Van Diest, P. J., Kayser, K., Meijer, G. A., & Baak, J. P. (1995). Syntactic structure analysis. *Pathologica*, *87*(3), 255-262.
- Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a. *Quetelet*, *10*, 113-121.
- Verhulst, P. F. (1845). Nouveaux memoires de l'Academie royale des sciences et belles-lettres de Bruxelles, 18, 1.
- Wang, L., Valderramos, S. G., Wu, A., Ouyang, S., Li, C., Brasil, P., ... & Aliyari, R. (2016). From mosquitos to humans: genetic evolution of Zika virus. *Cell host & microbe*, *19*(5), 561-565.
- Wang, Z., Butner, J. D., Kerketta, R., Cristini, V., & Deisboeck, T. S. (2015, February). Simulating cancer growth with multiscale agent-based modeling. In *Seminars in cancer biology* (Vol. 30, pp. 70-78). Academic Press.
- Watson, J. D., & Crick, F. H. C. (1953). Uma Estrutura para o Ácido Desoxirribonucléico. *Nature*, *171*(4356), 737-738.
- Weldon, W. F. R. (1892). Certain correlated variations in Crangon vulgaris. *Proceedings of the Royal Society of London*, *51*(308-314), 1-21.
- Werner, H. M., Mills, G. B., & Ram, P. T. (2014). Cancer systems biology: a peek into the future of patient care?. *Nature reviews Clinical oncology*, *11*(3), 167-176.
- Wiener, N. (1948). *Cybernetics* (p. 112). Paris: Hermann.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, *1*(6), 80-83.

° Wolkenhauer, O., Auffray, C., Jaster, R., Steinhoff, G., & Dammann, O. (2013). The road from systems biology to systems medicine. *Pediatric research*, *73*(4-2), 502-507.

° Wynder, E. L., Lemon, F. R., & Bross, I. J. (1959). Cancer and coronary artery disease among Seventh-Day Adventists. *Cancer*, *12*(5), 1016-28.

° Xiao, Y., Miao, H., Tang, S., & Wu, H. (2013). Modeling antiretroviral drug responses for HIV-1 infected patients using differential equation models. *Advanced drug delivery reviews*, *65*(7), 940-953.

° Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., ... & Velculescu, V. E. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, *467*(7319), 1114-1117.

° Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, *9*(1), 40.