| Volumen 40 | Número 1 | **ENERO-ABRIL 2019**

UNA PUBLICACIÓN DE LA
SOMIB
Sociedad Mexicana
de Ingeniería Biomédica

REVISTA MEXICANA DE Ingeniería Biomédica

- **Image Registration Measures and Chronic Osteoarthritis Knee Pain Prediction: Data from the Osteoarthritis Initiative**
- **Prototipo de Silla de Ruedas Dirigida Usando Parpadeos**
- **Reducción del Riesgo en Equipos Biomédicos y en Instalaciones Eléctricas de Entornos Clínicos**
- **ABPSE: Alineador de ADN Basado en Paralelismo a Nivel de Bit y la Estrategia Siembra y Extiende**
- **Preparación de un Adhesivo Sensible a la Presión (PSA) con la Incorporación de Nanopartículas de ZnO. Estudio de sus Propiedades Fisicoquímicas y Antimicrobianas**

Edición Especial

Artículos de Investigación en la Especialidad *Biología Teórica y Sintética*

http://www.rmib.mx

ISSN 2395-9126

SOMIB
Sociedad Mexicana
de Ingeniería Biomédica

Sociedad Mexicana de Ingeniería Biomédica

La Mesa Directiva de la Sociedad Mexicana de Ingeniería Biomédica hace una extensa invitación a las personas interesadas en participar, colaborar y pertenecer como Socio Activo de la SOMIB. La SOMIB reúne a profesionistas que se desarrollan en áreas de Ingeniería Biomédica, principalmente ingenieros biomédicos, así como otros profesionistas afines con el desarrollo de tecnología para la salud.

Membresía Estudiante
$850.00 ***PESOS MXN***
15% de descuento para grupos de 5 o más personas.

Membresía Profesional
$1,450.00 ***PESOS MXN***
15% de descuento para grupos de 5 o más personas.

Membresía Institucional
$11,600.00 ***PESOS MXN***
No aplica descuento.

Membresía Empresarial
$15,500.00 ***PESOS MXN***
No aplica descuento.

EL PAGO CUBRE UN AÑO DE CUOTA. EN CASO DE REQUERIR FACTURA
FAVOR DE SOLICITARLA, ADJUNTANDO COMPROBANTE DE PAGO Y ESPECIFICANDO CONCEPTO,
AL CORREO ELECTRÓNICO: ***facturación@somib.org.mx***

REVISTA MEXICANA DE Ingeniería Biomédica

AUTORES

Los trabajos a publicar en la RMIB, deben ser originales, inéditos y de excelencia. Los costos de publicación para autores son los siguientes:

NO SOCIOS:	***$4,060.00*** ***PESOS MXN*** *(INCLUYE I.V.A.)*
SOCIOS:	***$1,276.00*** ***PESOS MXN*** *(INCLUYE I.V.A.)*

PUBLICIDAD

A las empresas e instituciones interesadas en publicitar su marca o productos en la RMIB, los costos por número son los siguientes:

MEDIA PLANA:	***$4,999.00*** ***PESOS MXN*** *(INCLUYE I.V.A.)*
UNA PLANA:	***$6,799.00*** ***PESOS MXN*** *(INCLUYE I.V.A.)*
CONTRAPORTADA:	***$7,799.00*** ***PESOS MXN*** *(INCLUYE I.V.A.)*
FORROS INTERIORES:	***$7,799.00*** ***PESOS MXN*** *(INCLUYE I.V.A.)*

DESCUENTO DEL **20%** AL CONTRATAR PUBLICIDAD EN DOS O MÁS NÚMEROS.

Para ser socio

- Presentar el formato de inscripción.
- Realizar el pago de derechos, de acuerdo a la categoría.
- Enviar formato de inscripción, currículum y comprobante de pago a: *socios@somib.org.mx*.
- Se emitirá carta de aceptación y constancia de membresía por parte de la mesa directiva (aprobada la solicitud).
- Para mayor información sobre beneficios, ingresar a: *www.somib.org.mx*.

Datos bancarios

- **Beneficiario:** *Sociedad Mexicana de Ingeniería Biomédica A. C.*
- **Banco:** *Scotiabank*
- **Referencia:** *1000000333*
- **Cuenta:** *11006665861*
- **CLABE Interbancaria:** *044770110066658614*

La inserción de la publicidad será publicada en el libro electrónico y en el área de patrocinios en el sitio Web de la revista (RMIB), disponible en:

http://rmib.mx

REVISTA MEXICANA DE
Ingeniería Biomédica

Índices
La Revista Mexicana de Ingeniería Biomédica aparece en los siguientes índices científicos:
Sistema de Clasificación de Revistas Científicas y Tecnologías del CONACYT - Q4, SCOPUS, SciELO, REDALyC, EBSCO, LATINDEX, Medigraphic Literatura Biomedica, Sociedad Iberoamericana de Información Científica - SIIC.

www.rmib.mx
ISSN 2395-9126

Editor Técnico y en Internet
Enrique Ban Sánchez

Sociedad Mexicana de Ingeniería Biomédica
Plaza Buenavista #2, Col. Buenavista, Del. Cuauhtémoc, C.P. 06350, Ciudad de México, México, (555) 574-4505

SOMIB
Sociedad Mexicana
de Ingeniería Biomédica

Afiliada a:
International Federation of Medical and Biological Engineering (IFMB-IUPSM-ICSU)
Federación de Sociedades Científicas de México, A.C. (FESOCIME)
Consejo Regional de Ingeniería Biomédica para América Latina (CORAL)

SOMIB
Plaza Buenavista #2, Col. Buenavista Del. Cuauhtémoc, C.P. 06350 Ciudad de México, México (555) 574-4505
www.somib.org.mx

REVISTA MEXICANA DE INGENIERÍA BIOMÉDICA, Vol. 40, No. 1, Enero-Abril 2019, es una publicación cuatrimestral editada por la Sociedad Mexicana de Ingeniería Biomédica A.C., Plaza Buenavista #2, Col. Buenavista, Del. Cuauhtémoc, Ciudad de México, 06350, (555) 574-4505, www.somib.org.mx, rmib.somib@gmail.com. Editor responsable: Nelly Gordillo Castillo. Reserva de Derechos al Uso Exclusivo No. 04-2015-041310063800-203, ISSN (impreso) 0188-9532; ISSN (electrónico) 2395-9126, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este número: Lic. Enrique Federico Ban Sánchez, Plaza Buenavista #2, Col. Buenavista, Del. Cuauhtémoc, Ciudad de México, 06350, (555) 574-4505, fecha de última modificación, 15 de diciembre de 2016.

Disponible en línea:
www.rmib.mx

CONTENIDO
CONTENTS

El contenido aparece según su publicación en el portal de la revista.

ARTÍCULO DE INVESTIGACIÓN

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-13

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.1

E-LOCATION ID: **e201812**

Image Registration Measures and Chronic Osteoarthritis Knee Pain Prediction: Data from the Osteoarthritis Initiative

Métricas de Registro de Imágenes y Predicción de Dolor de Rodilla por Osteoartritis Crónica: Datos de la *Osteoarthritis Initiative*

J. I. Galván-Tejada, C. E. Galván Tejada, F. E. López-Monteagudo, O. Alonso-González, A. Moreno-Báez, J. M. Celaya-Padilla, L. A. Zanella-Calzada

Universidad Autónoma de Zacatecas

ABSTRACT

Osteoarthritis (OA) is the most common type of arthritis, is a growing disease in the industrialized world. OA is an incapacitate disease that affects more than 1 in 10 adults over 60 years old. X-ray medical imaging is a primary diagnose technique used on staging OA that the expert reads and quantify the stage of the disease. Some Computer-Aided Diagnosis (CADx) efforts to automate the OA detection have been made to aid the radiologist in the detection and control, nevertheless, the pain inherits to the disease progression is left behind. In this research, it's proposed a CADx system that quantify the bilateral similarity of the patient's knees to correlate the degree of asymmetry with the pain development. Firstly, the knee images were aligned using a B-spline image registration algorithm, then, a set of similarity measures were quantified, lastly, using this measures it's proposed a multivariate model to predict the pain development up to 48 months. The methodology was validated on a cohort of 131 patients from the Osteoarthritis Initiative (OAI) database. Results suggest that mutual information can be associated with K&L OAI scores, and Multivariate models predicted knee chronic pain with: AUC 0.756, 0.704, 0.713 at baseline, one year, and two years' follow-up.

KEYWORDS: Osteoarthritis Initiative; biomarker; KellgrenLawrence grade; knee registration; pain prediction

RESUMEN

La osteoartritis (OA) es el tipo de artritis más común. OA es una enfermedad limitante que afecta a 1 de 10 adultos con 60 años o más. Las imágenes de rayos-x son una técnica de diagnóstico primario que permite conocer el estado de OA, las cuales el experto lee y cuantifica así la etapa de la enfermedad. El Diagnóstico Asistido por Computadora (CADx, por sus siglas en inglés) ha buscado automatizar el diagnóstico de OA para ayudar al radiólogo en la detección y control; sin embargo, el dolor provocado por la progresión de la enfermedad es dejado atrás. En este trabajo se propone un sistema de CADx que cuantifica la similitud bilateral de las rodillas de los pacientes, con el fin de correlacionar el grado de asimetría con el dolor. Inicialmente, las imágenes de las rodillas fueron alineadas usando el algoritmo B-spline para su registro, después, un conjunto de métricas estándar fue cuantificado; finalmente, con estas métricas se propone un modelo multivariado para predecir el dolor de rodilla desarrollado en 48 meses. La metodología fue validada con 131 pacientes obtenidos de la base de datos de la Osteoarthritis Initiative (OAI). Los resultados sugieren que las métricas pueden ser asociadas con los puntajes de KellgrenLawrence; además, los modelos predicen significativamente el dolor crónico de rodilla con: AUC 0.756, 0.704 y 0.7113, al inicio, un año y dos años después, respectivamente.

PALABRAS CLAVE: *Osteoarthritis Initiative*; biomarcador; grado KellgrenLawrence; predicción de dolor

Correspondencia

DESTINATARIO: **Carlos Eric Galván Tejada**
INSTITUCIÓN: **Universidad Autónoma de Zacatecas**
DIRECCIÓN: **Jardín Juárez #147, Col. Centro, C. P. 98000, Zacatecas, Zacatecas, México**
CORREO ELECTRÓNICO: **ericgalvan@uaz.edu.mx**

Fecha de recepción:

12 de enero de 2018

Fecha de aceptación:

9 de octubre de 2018

INTRODUCTION

There are over 200 different types of arthritis [1]. Two of the most common types are osteoarthritis and rheumatoid arthritis [2]. Nowadays, Osteoarthritis (OA) is the most common representative of arthritis, and a growing disease in the industrialized world. Lifestyle and habits appear to be the cause of increasing cases of OA [3-5]. This disabling disease conducts poor quality of life to patients, becoming the quotidian activities into painful tasks. This disorder affects at least 1 in 10 adults advanced in over 60 years, in the United States, and is classified between the principal causes of medical attention requests [6-8].

Due to its simplicity and broad base deployment, X-ray medical imaging is a primary diagnose technique used on staging OA [9]. Expert radiologists evaluate radiological evidence of x-ray images using several radiological methods to establish this stage, some bony changes such like the emergence of osteophytes, anatomical changes or joint space narrowing (JSN) are the main features observed to perform this task [9]. This radiological features have not been fully studied in association with the most common symptoms; pain and stiffness [10,11]. Correlate radiological evidence, and a subjective late onset symptom as pain, is one of the biggest challenges in OA.

Associations create information between the disorder and its behavior looking for treatments or therapies for OA. The Osteoarthritis Initiative (OAI) effort are bringing information that will allow to comprend the disease behavior. OAI has collected a big quantity of clinical data from OA patients, subjects with probably risk, and control subjects under validated tests and standardized image assessment procedures.

Early diagnosis is key to treat the symptoms and the treat the advance of the disease. Looking for a better explanation of the pain as a complex symptom OA researchers has developed different clinical tests as *KOOS* and *WOMAC* [12-15], and some atlas based on image evaluation [16]. There are methods such as *KellgrenLawrence* (*K & L*) or the *OARSI* grading scale [17] that are part of the radiological evaluation on images, these methods depend on trained radiologist and human criteria to determine stage and a path of action [18, 19].

Using bilateral x-ray images of knees from the open databases OAI, the objective of this search is to correlate measures obtained in an automated way with knee chronic pain as the principal symptom, through computational algorithms. Through recent image registration approaches [20], three well known measures of similarity between knees are obtained: mutual information [21], correlation [22], and mean squared error [22]. In a previous effort [23], these measures were explored as a tool for association with *K & L*, one of the most used OA grading tools. In a preliminary work, a small group of patients was used to explore the association between the *K & L* scale and the error metrics between the recorded images, that work suggested that there could be a relationship that is dealt with more broadly here [24].

Being chronic pain a late onset symptom, the use of said measures as a risk factor may help to develop a rapid diagnosis and a better treatment option. In this work, the association between the automated measures and chronic knee pain is studied. Three different time points are explored, the time of baseline visit, a year after enrollment and two years after enrollment. The three measures in multivariate predictive models are explored, and the most significate variable in a univariate predictive model is studied.

The results suggest a close relationship between measures from image registration and chronic pain in all time points analyzed. All multivariate models using all measures were predictive. Mutual information as a univariate model obtained a better performance pre-

dicting pain in two studied time points. Also, mutual information maintains the relationship whit *K & L* in all time points studied.

The main objective of this work is to have a first contact with the use of registration tools for the early detection of osteoarthritis of the knee, as already analyzed in previous works, the radiological evidence evaluated by experts, determine that small changes in the structure of the bone, are risk factors that can be associated with the development of chronic knee pain. That is why the measure of asymmetry becomes relevant, since as it has been reported, the disease appears with greater intensity in one of the patient's knees. The hypothesis is that the greater the error in the registry, the greater the deformation in one of the knees.

This paper is organized as follows; after Introduction, the Materials and Methods are explained, in Image segmentation subsection, the process of the automated segmentation of the knees is presented, in Image registration subsection, the registration of the left knee into the right knee is explained, in Metric quantification subsection, the equations for the measures extraction are exposed. In the Results and Discussion section, the ROC curves with the numerical results are presented, finally, the Conclusion.

MATERIALS AND METHODS

On this study population: Data used in the preparation of this article were obtained from the Osteoarthritis Initiative (OAI) database, which is available for public access at *http://www.oai.ucsf.edu/datarelease/*.

Being a pain prediction study, the chronic pain was defined as the variable to look at. A cohort of 131 patients is used in this study; all patients should have the baseline radiological study, and complete chronic pain information. The selection criteria for the cohort were:

Control patients were selected according to the criteria of:

1. No clinical symptoms of knee pain, from baseline to 60-month follow-up,
2. No symptomatic clinical OAI data, from baseline to 60-month follow-up;
3. No analgesic NSAID intake, from baseline to 60-month follow-up.

Patients were selected according to the criteria of:

1. No clinical symptoms of knee pain at the initial visit;
2. No analgesic NSAID intake at the reference visit;
3. Clinical symptomatic manifestation of chronic pain in the right knee, at any time after the reference visit until the 60-month follow-up.

Demographic information and statistical details of the cohort are presented as follows: the total of patients (F) were 131 (78), with an age range (S.D.) of 45-79 (10.41), an average height of 1678.15 (86.10) and an average BMI (S.D) of 29.48 (5.11). From the total of patients, 38 (15) were controls, with an age range of 45-78 (10.9), an average height of 1711.59 (82.71) and an average BMI of 28.19 (3.78). The rest of the patients, 93 (63), were cases, with an age range of 45-79 (10.29), an average height of 1663.43 (83.22) and an average BMI of 30.04 (5.55).

FIGURE 1. **Left and right knees input image, ROI is delimited by a yellow rectangle.**

In this analysis, images and databases from OAI used are: OAI is a multi-center, longitudinal, prospective observational study of knee OA. The OAI will establish and maintain a natural history database for osteoarthritis that will include clinical evaluation data, radiological (x-ray and magnetic resonance) images, and a bio-specimen repository from 4796 men and women aged between 45 and 79 years old, enrolled between February 2004 and May 2006").

1. Bilateral knee x-ray images (contains bilateral fixed-flexion knee radiographs).
2. JointSx00 ver 0.2.2 (contain questionnaire results regarding arthritis symptoms in the knee; arthritis-related joint function and disability; and general health-related function and disability in the baseline).
3. JointSx01 ver 1.2.1 (contain questionnaire results regarding arthritis symptoms in the knee; arthritis-related joint function and disability; and general health-related function and disability in the 12-month visit).
4. JointSx02 ver 2.2.2 (contain questionnaire results regarding arthritis symptoms in the knee; arthritis-related joint function and disability; and general health-related function and disability in the 24-month visit).
5. kXR_SQ_BU00_SAS ver 0.7 (contains central longitudinal readings of serial knee X-rays for tibiofemoral radiographic OA in the baseline).
6. kXR_SQ_BU01_SAS ver 1.7 (contains central longitudinal readings of serial knee X-rays for tibiofemoral radiographic OA in the 12-month visit).
7. kXR_SQ_BU03_SAS ver 3.6 (contains central longitudinal readings of serial knee X-rays for tibiofemoral radiographic OA in the 24-month visit).
8. "Right knee symptom status" (combines past thirty days and twleve months, used in OAI definition of sympthomatic knee OA).
9. All *K & L* scores were assessed by OAI Boston University radiologist group using the standard atlas for OA [25, 26].

In the bilateral images used for this work, left and right knees are presented side by side. Per the high dynamism of the knees, a direct comparison between left and right knees images cannot be realized truthful. Therefore, before the analysis an alignment of the images must be performed.

The principal steps that were done in methodology:

1. The images of the patient's knees are segmented to delete undesired information,
2. The left knee is aligned/registered to the right knee,
3. An evaluation of the similarity measures is computed looking for the relationship between the degree of similarity in both knees and the disease phases.

Prior to image registration process, each image was manually preprocessed. By the registration process, in each image the region of interest (ROI) was attached to avoid regions without meaningful information. In Figure 1 from the input image the ROI is delimited by a yellow rectangle.

In the process of ROI adjustment in each of the x-ray images, was generated an individual image for rights and lefts knees, afterwards, each image of the left knee was reflected around the vertical axis to allow the image registration procedure, finally, each pixel of the image is submitted to a logarithmic transformation to improve the low intensity pixels [27, 28].

Image segmentation

The background noise and artifacts of the x-ray images were eliminated by a method of automated segmentation. By this method is created a segmenta-

tion mask, then, the mask dismissed those pixels under five level deviations based on the noise level of the images. The segmentation in based to the mask abstracts the knee bones structure and dismisses the background. Finally, the region that was the largest connected on the image was subjected to a hole filling algorithm, based on dilatation and erosion, according to the morphological functions to assure a solid surface abstraction, using a 3x3 supporting región, as defined in Equation 1:

$$S_j^i(x,y) = (I_j^i(x,y) \oplus B(x,y)) \ominus B(x,y) \qquad (1)$$

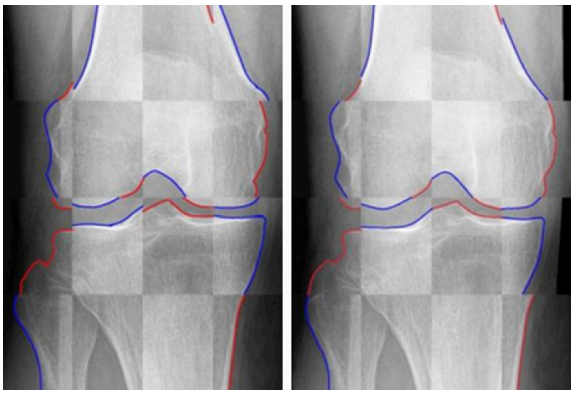
Where S_j^i (x, y) and I_j^i (x, y)represent the segmented and raw images for the *ith* view, and the *jth* side, left or right and, $\oplus$ and $\ominus$ are the grayscale dilatation and erosion morpholoical operations, respectevely, and $B(x, y)$ is a 3x3 structural element.

Image registration

The procedure of the left knee image registration into the right knee was performed based on the segmentation mask. Initially, the left knee image was reflected and then co-registered with the image of the right knee corresponding. An algorithm of B-Spline multi resolution had the purpose of optimizing the Mattes mutual data measures in the bilateral image registration [29, 30]. Then, a deformable B-Spline transform was used, this process based its function in the transformation of an image adjusting control points of a net in base on a similarity measure maximization, this method usually avoids local minimal in the parameter search space and decreases computational time [31, 32].

By the multi resolution method used, all images were registered in the lowest resolution. In the near steps, the transformation parameters are scaled to the higher resolutions and is calculated again the parameter optimization. Fort he 2D images that are involved in this study, B-splines can be modeled by the tensor product of the 1D cubic B-splines. A 2D rigid transformation can be represented as in Equation 2, where r_p represents the x and y coordinates of the *pth* pixel and $d(r_p)$ the deformation it suffers.

$$T(r_p) = r_p + d(r_p) \qquad (2)$$



a) b)

FIGURE 2. A) Checkerboard image of the unregistered right to left knee, B) Checkerboard image of the registered right to left knee. Red (left knee) and blue (right knee) lines where draw over the edge of the knee images to facilitate the graphical comparison.

The 2D deformation was modeled using the tensor product of β, represented as $d(r_p) = \langle d^x(r_p), d^y(r_p)\rangle$ using the tensor product of β, the nth-order B-splines, as follows in Equation 3:

$$d^q(r_p) = \sum_{i=j}^{n-1} c_{i,j}^q \beta\left(\frac{x}{m_x} - i\right)\beta\left(\frac{y}{m_y} - j\right) \qquad (3)$$

Where $d^q(r_p)$ represents the deformation of the *pth* pixel in the *qth* axis plane (x or y), $c = c_{i,j}^q$ is the deformation coefficient for the *qth* plane, and m_q is the knot spacing in the *qth* direction. The deformation coefficients were estimated by maximizing the similarity metric ψ, according to Equation 4:

$$\hat{c} = argmax_q \psi(S_{right}^i(x,y), S_{left}^i(T(x,y;c))) \qquad (4)$$

Then, the registration algorithm returned a transformation file, T(x,y). This file has the purpose of finding a point in the left image according to each point (x,y)

in the right image [33]. Figure 2 shows the checkerboard of the registered image output.

It is important to mention that Equation 4 was minimized using the *RegularStepGradientDescentOptimizer* function, from the ITK library [34]. This function basically refers to the gradient descent minimizing method.

Metric quantification

Three meaningful measurements were calculated to set the relationship between the registered image and the target image. These measurements were mean squared error (MSE), correlation coefficient and mutual information [35]. These parameters are widespread utilized to compare two different images [36].

The first parameter, MSE, is calculated according to the Equation 5, which assumes that the images are the same at registration, therefore, it is implicitly assumed that do not exist differences between intensity levels. The MSE parameter is sensitive to outliers, that is to say, a small group of voxels characterized by having high differences on the intensity levels. In this equation, F is the model image, R is the objective image into the image and F^T is the transformed image after the registration.

$$MSE = \frac{1}{XY}\sum_{i=1}^{X}\sum_{j=1}^{Y}[F^T(i,j) - R(i,j)]^2 \qquad (5)$$

In Equation 6 is calculated the Pearson's correlation coefficient (CC), which is a parameter that measures the linear dependence between two different variables on images. R_m is the mean of the pixel R in the domain (R, F) and F_m^T is the mean of F^T in the domain (R, F).

$$CC = \frac{\sum_i (R_i - R_m)(F_i{}^T - F_m{}^T)}{\sqrt{\sum_i \left(R_i - R_m\right)^2}\sqrt{\sum_i \left(F_i{}^T - F_m{}^T\right)^2}} \qquad (6)$$

In Equation 7 the mutual information (MI) is calculated, which obtains a parameter of probabilistic dependence between two different intensity distributions. In this study is obtained the Shannon-Wiener entropy measure H by this equation, where $H(A)$ and $H(B)$ represent the entropies of the images A and B, and $H(A, B)$ represent the joint entropy.

$$MI(A,B) = H(A) + H(B) - H(A,B) \qquad (7)$$

In Equation 8 is calculated $H(A, B)$, which measures how much uncertainty there is in the two random variables A and B taken together; where $p(a, b)$ denotes the probability mass function considering two random variables jointly distributed.

$$H(A,B) = -\sum_{a,b} p_{AB}(a,b)\log p_{AB}(a,b) \qquad (8)$$

Statistical Analysis

All image registration measures were correlated with the *K & L* score. For data analysis, three multivariate searches were performed using the knee chronic pain as an outcome variable. The time points evaluated for chronic pain, in the 48-month, in each search were: baseline visit (T0), one year after baseline (T1), and two years after the baseline visit (T2).

To evaluate the predictive individual measures performance, a logistic regression was performed according to the binary outcome variable, which is represented as No pain = 0 and Chronic pain = 1. Using the image registration measures individual variable, a logistic regression was performed using T0, T1, and T2 as an outcome.

They were developed general linear models, which were analyzed later. Also, they were calculated the odds ratios, and the area under the receiver operating characteristic (ROC) curve (AUC) for each model. Leave one out cross validation (LOOCV) was performed. The parameter calculated, ROC, is a function which generates a curve and is interpreted as a graph-

ical representation of the sensitivity (taking values from 0 to 1) and the specificity (taking values from 1 to 0) for a binary classifier or model system as the discrimination threshold is varied, looking for the quantity of true negatives and true positives, in order to validate the model accuracy.

For the subsequent statistical analyzes was used the free statistical software R and some of its packages [37].

RESULTS AND DISCUSSION

Experimental results of the mutual information, correlation, and mean squared error are presented in Figure 3 for T0, T1, and T2.

In Figure 4, ROC curves of mutual information, correlation and mean squared error by time points T0, T1 and T2, are displayed. Each multivariate model predictive performance is presented for T0, T1, and T2. In Figure 5 ROC curves of multivariate models are displayed.

A chronic pain association study using image registration measures is presented for the first time. By analyzing bilateral knee x-ray images the results can be used to generate an automated knee evaluation model for knee OA. Some other studies use image registration in OA, but not in X-ray, and not for the symptom association [38-40].

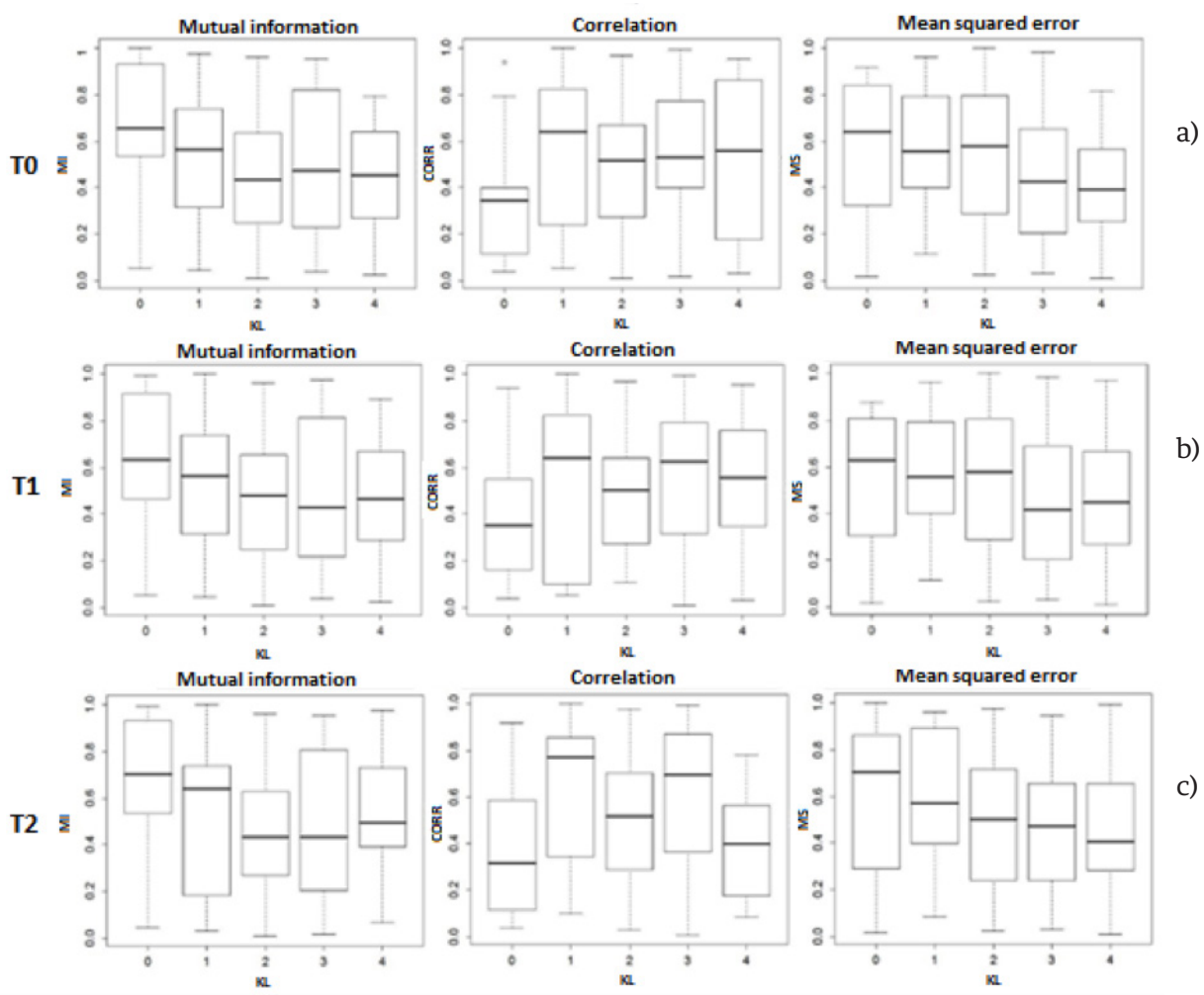


FIGURE 3. **A) Mutual information, correlation and mean squared error in T0, B) Mutual information, correlation and mean squared error in T1, C) Mutual information, correlation and mean squared error in T2.**

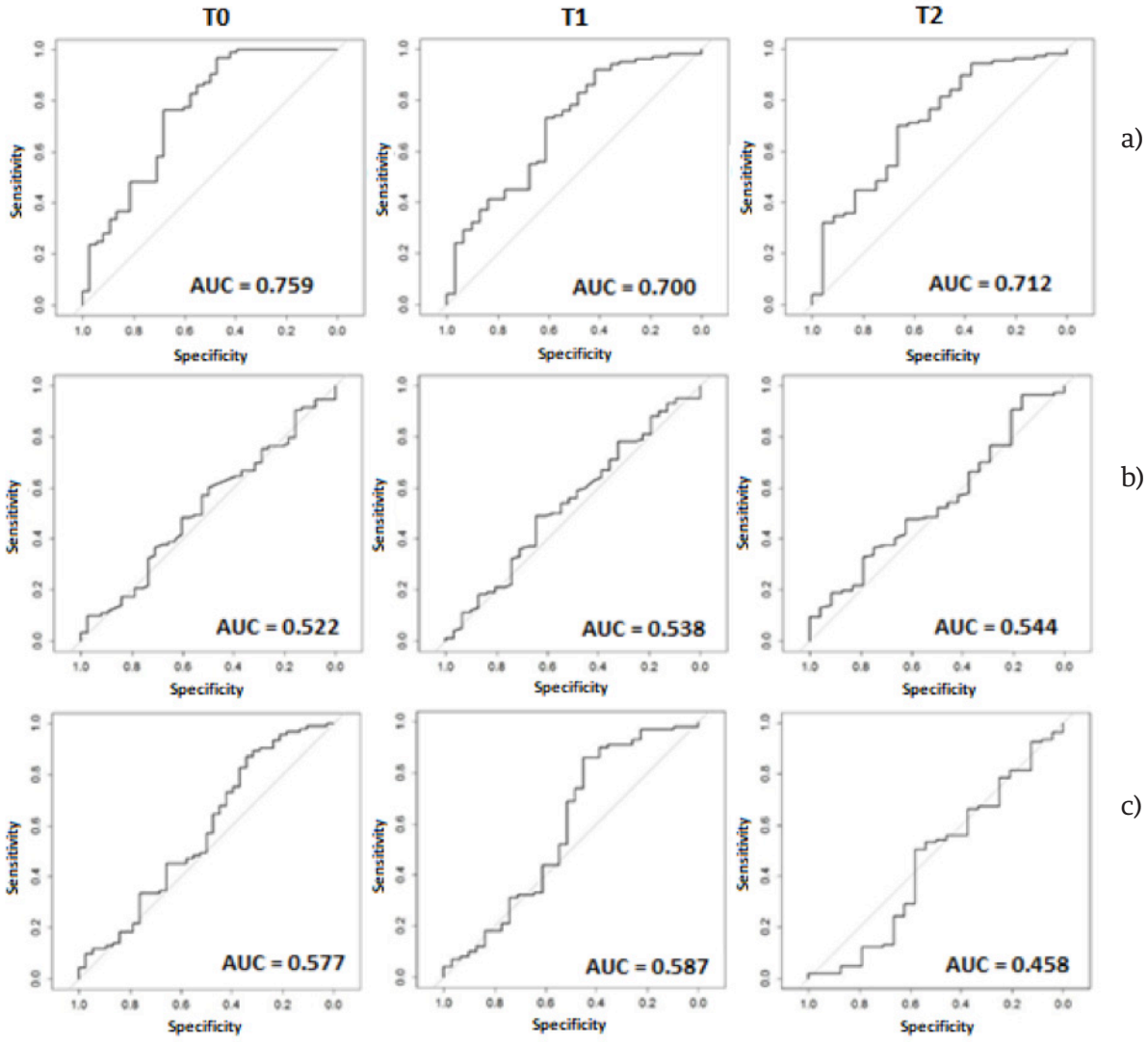
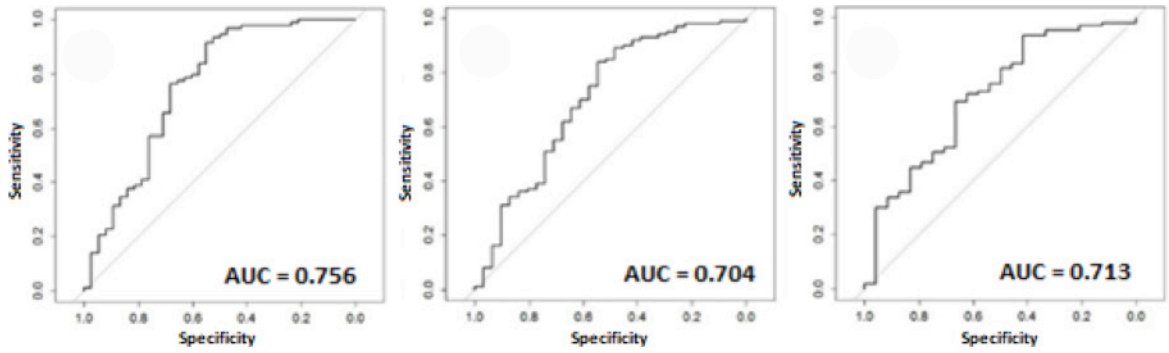


FIGURE 4. **A) Mutual information ROC for T0, T1 and T2. B) Correlation ROC for T0, T1 and T2. C) Mean squared error ROC for T0, T1 and T2.**

The mutual information shows a behavior downward with respect to the increase in the *K & L*, in this experiment the evaluation of *K & L* was not balanced so apparently, the correlation and the mean square error appear not to associate with *K & L* as shown in Figure 4 at T0, T1 and T2.

After evaluating the predictive performance measures as individual variables, it is evident as is shown in Figure 5, that the correlation and the mean square error have no predictive power for themselves. However, mutual information has better performance than univariate models in T0 and T2, and their behavior alone is very similar to the multivariate model in T1.

After evaluating the three multivariate models, we can say that there is a close association between the measures obtained automatically and chronic pain presented in the three time points observed. The predictive power of the models presented in Figure 5 is very similar but superior to the models obtained by radiological information measured by expert radiologists [41].



a) b) c)

FIGURE 4. Multivariate models ROC curves A) T0, B) T1, and C) T2.

Multivariate models have acceptable performance in their predictability based on the AUC, and then only the model for T1 has a higher performance univariate model, and is very similar.

There are limitations in the study. First, patient selection criteria make the study population decline as no loss of information in databases. Also, pain is a subjective outcome that changes from person to person, and its mechanism is not fully studied.

Given these limitations, we cannot generalize the findings and the external validation of the results is required to assess the clinical applicability of the models.

On the other hand, it is important to mention that in this work it was proposed the local image registration method because, according to the work of Celaya et al. [42], where a comparison for bilateral registration mammography was performed, comparing different types of registration in two images that theoretically should present the same characteristics but, due to the biological phenomena, they presented heterogenous tissue, demonstrating that for this type of problem, which is very similar with the problem presented here, the results obtained using this method were the most robust, which allows to confirm the hypothesis that the greater the error in the registry, the greater the deformation in one of the knees.

The results suggest that measures of asymmetry obtained from the image registration show a close relationship with chronic pain development in patients with OA. Also, they suggest that the use of measures obtained automatically, can lead to the development of a tool to set the stage of the disease in which there is no human intervention. Also, suggest that the mutual information can be used individually as a risk factor for the future development of chronic pain as a symptom of OA.

CONCLUSIONS

Of research results we can conclude that the use of image registration in x-ray images has potential for the development of automated setting step in which a patient is OA tools. Also, the degree of asymmetry between the knees, can lead to early diagnosis and thus obtain a better prognosis for the patient on the progression of the disease and its symptoms.

Due to the widespread use of x-rays, it opens a great opportunity to get a support system for decision making for the radiologist, even in places where there is no access to advanced medical services.

The public health systems would benefit of an automated system, as in developing countries, where the number of trained radiologists is limited, and the workload to which they are exposed is too much.

A system that allows the radiologist to decrease the workload can lead to better diagnosis of patients and thus controlling disease progression in patients with signs of OA. The use of computational tools in medical science has a big impact, since these tools allow us to handle a large amount of information accurately and get results that are not possible with traditional methods.

As future work is to expand the number of subjects, and if possible, get data immunological bases from other studies to corroborate the results obtained in this work.

ACKNOWLEDGEMENTS

This work was partially supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT), by Grant 16864 Ciencia Básica from CONACYT, Bioinformatics work group from Tecnológico de Monterrey.

J. I. G.T. thanks to PROMEP for partially support his doctoral studies, the second author wants to thank the CONACYT, for the support under grant "CONACYT Cátedra 129 - Convocatoria 2016".

"The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners".

REFERENCES

[1] Beasley J. Osteoarthritis and rheumatoid arthritis: conservative therapeutic management. Journal of hand 255 therapy. 2012;25(2):163-172. DOI: 10.1016/j.jht.2011.11.001

[2] Pearson MJ, Jones SW. Long non-coding RNAs in the regulation of inflammatory pathways in rheumatoid arthritis and osteoarthritis. Arthritis & Rheumatology. 2016. DOI: 10.1002/art.39759

[3] Jones G, Cooley HM, Stankovich JM. A cross sectional study of the association between sex, smoking, and other lifestyle factors and osteoarthritis of the hand. The Journal of rheumatology. 2002;29(8):1719-1724.

[4] Vrezas I, Elsner G, Bolm-Audorff U, Abolmaali N, Seidler A. Case-control study of knee osteoarthritis and lifestyle factors considering their interaction with physical workload. International archives of occupational and environmental health. 2010;83(3):291-300. DOI: 10.1007/s00420-010-0536-0

[5] O'Reilly S, Doherty M. Lifestyle changes in the management of osteoarthritis. Best Practice & Research Clinical Rheumatology. 2001;15(4):559-568. DOI: 10.1053/berh.2001.0173

[6] Arden N, Nevitt MC. Osteoarthritis: epidemiology. Best practice & research Clinical rheumatology. 2006;20(1):3-25. DOI: 10.1016/j.berh.2005.09.007

[7] Agaliotis M, Fransen M, Bridgett L, Nairn L, Votrubec M, Jan S, et al. Risk factors associated with reduced work productivity among people with chronic knee pain. Osteoarthritis and Cartilage. 2013;21(9):1160-1169. DOI: 10.1016/j.joca.2013.07.005

[8] White DK, Tudor-Locke C, Felson DT, Gross KD, Niu J, Nevitt M, et al. Do radiographic disease and pain account for why people with or at high risk of knee osteoarthritis do not meet physical activity guidelines? Arthritis & Rheumatism. 2013;65(1):139-147.

[9] Hayashi D, Guermazi A, Roemer FW. Radiography and computed tomography imaging of osteoarthritis. Oxford Textbook of Osteoarthritis and Crystal Arthropathy. 2016. DOI: 10.1002/art.37748

[10] Stutman D, Beck TJ, Carrino JA, Bingham CO. Talbot phase-contrast x-ray imaging for the small joints of the hand. Physics in medicine and biology. 2011;56(17):5697.

[11] Kornaat PR, Bloem JL, Ceulemans RY, Riyazi N, Rosendaal FR, Nelissen RG, et al. Osteoarthritis of the Knee: Association between Clinical Features and MR Imaging Findings 1. Radiology. 2006;239(3):811-817. DOI: 10.1148/radiol.2393050253

[12] Roos EM, Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis. Health and quality of life outcomes. 2003;1(1):64. DOI: 10.1186/1477-7525-1-64

[13] Roos EM, Roos HP, Lohmander LS. WOMAC Osteoarthritis Index—additional dimensions for use in subjects with post-traumatic osteoarthritis of the knee. Osteoarthritis and Cartilage. 1999;7(2):216-221. DOI: 10.1053/joca.1998.0153

[14] Yang KA, Raijmakers N, Verbout A, Dhert W, Saris D. Validation of the short-form WOMAC function scale for the evaluation of osteoarthritis of the knee. Bone & Joint Journal. 2007;89(1):50-56. DOI: 10.1302/0301-620X.89B1.17790

[15] Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. The Journal of rheumatology. 288 1988;15(12):1833-1840.

[16] Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, et al. Development of criteria for the classification and reporting of osteoarthritis: classification of osteoarthritis of the knee. Arthritis & Rheumatism. 1986;29(8):1039-1049. DOI: 10.1002/art.1780290816

[17] Duryea J, Li J, Peterfy C, Gordon C, Genant H. Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. Medical physics. 2000;27(3):580-591. DOI: 10.1118/1.598897

[18] Pathria M, Sartoris D, Resnick D. Osteoarthritis of the facet joints: accuracy of oblique radiographic assessment. Radiology. 1987;164(1):227-230. DOI: 10.1148/radiology.164.1.3588910

[19] Kornaat PR, Ceulemans RY, Kroon HM, Riyazi N, Kloppenburg M, et al. MRI assessment of knee osteoarthritis: Knee Osteoarthritis Scoring System (KOSS)—inter-observer and intra-observer reproducibility of a compartment-based scoring system. Skeletal radiology. 2005;34(2):95-102. DOI: 10.1007/s00256-004-0828-0

[20] Zitova B, Flusser J. Image registration methods: a survey. Image and vision computing. 2003;21(11):977-1000. DOI: 10.1016/S0262-8856(03)00137-9

[21] Pluim JP, Maintz JA, Viergever MA. Mutual-information-based registration of medical images: a survey. IEEE transactions on medical imaging. 2003;22(8):986-1004. DOI: 10.1109/TMI.2003.815867

[22] Brown LG. A survey of image registration techniques. ACMcomputing surveys (CSUR). 1992;24(4):325-376. DOI: 10.1145/146370.146374

[23] Galván-Tejada JI, Celaya-Padilla JM, Treviño V, Tamez-Peña JG. Knee osteoarthritis image registration: data from the Osteoarthritis Initiative. In: SPIE Medical Imaging. International Society for Optics and Photonics; 306 2015. p. 94143C-94143C. DOI: 10.1117/12.2082426

[24] Galván-Tejada J.I., Galván-Tejada C.E., Celaya-Padilla J.M., Delgado-Contreras J.R., Cervantes D., Ortiz M. (2016) Automated Image Registration for Knee Pain Prediction in Osteoarthritis: Data from the OAI. In: Martínez-Trinidad J., Carrasco-Ochoa J., Ayala Ramirez V., Olvera-López J., Jiang X. (eds) Pattern Recognition. MCPR 2016. Lecture Notes in Computer Science, vol 9703. Springer, Cham. DOI: 10.1007/978-3-319-39393-3_33

[25] Kellgren JH, Jeffrey MR, Ball J. The epidemiology of chronic rheumatism: a symposium. vol. 2. FA Davis Company; 1963.

[26] Kellgren J, Lawrence J. Radiological assessment of osteo-arthrosis. Annals of the rheumatic diseases. 1957;16(4):494.

[27] Jourlin M, Pinoli JC. A model for logarithmic image processing. Journal of microscopy. 1988;149(1):21-35. DOI: 10.1111/j.1365-2818.1988.tb04559.x

[28] Jourlin M, Pinoli JC. Image dynamic range enhancement and stabilization in the context of the logarithmic image processing model. Signal processing. 1995;41(2):225-237. DOI: 10.1016/0165-1684(94)00102-6

[29] Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review. 2001;5(1):3-55. DOI: 10.1145/584091.584093

[30] Celaya-Padilaa JM, Rodriguez-Rojas J, Trevino V, Tamez-Pena JG. Local image registration a comparison for bilateral registration mammography. In: IX International Seminar on Medical Information Processing and Analysis. International Society for Optics and Photonics; 2013. p. 892210-892210. DOI: 10.1117/12.2035516

[31] Illescas MB, Menéndez CL, Rodríguez MR, Quintero RF. Nuevos criterios ASAS para el diagnóstico de espondiloartritis. Diagnóstico de sacroileítis por resonancia magnética. Radiología. 2014;56(1):7-15. DOI: 10.1016/j.rx.2013.05.004

[32] Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE transactions on medical imaging. 1999;18(8):712-721. DOI: 10.1109/42.796284

[33] Celaya-Padilla J,Martinez-Torteya A, Rodriguez-Rojas J, Galvan-Tejada J, Treviño V, Tamez-Peña J. Bilateral image subtraction and multivariate models for the automated triaging of screening mammograms. BioMed research international. 2015;2015. DOI: 10.1155/2015/231656

[34] Ibanesz, Luis, et al. The ITK software guide. 2005.

[35] Crum WR, Hartkens T, Hill D. Non-rigid image registration: theory and practice. The British Journal of Radiology. 2014. DOI: 10.1259/bjr/25329214

[36] Guo Y, Suri J, Sivaramakrishna R. Image registration for breast imaging: a review. In: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. IEEE;330 2006. p. 3379-3382. DOI: 10.1109/IEMBS.2005.1617202

[37] Team RC. The R project for statistical computing. Available at: *www R-project org/ Accessed October.332 2014;31:2014*.

[38] Rogers BP, Haughton VM, Arfanakis K, Meyerand ME. Application of image registration to measurement of intervertebral rotation in the lumbar spine. Magnetic resonance in medicine. 2002;48(6):1072-1075. DOI: 10.1002/mrm.10319

[39] Lynch JA, Peterfy CG, White DL, Hawkins RA, Genant HK. MRI-SPECT image registration using multiple MR pulse sequences to examine osteoarthritis of the knee. In: Medical Imaging'99. International Society for Optics and Photonics; 1999. p. 68-77. DOI: 10.1117/12.348626

[40] Bron EE, van Tiel J, Smit H, Poot DH, Niessen WJ, Krestin GP, et al. Image registration improves human knee cartilage T1 mapping with delayed gadolinium-enhanced MRI of cartilage (dGEMRIC). European radiology. 2013;23(1):246-252. DOI: 10.1007/s00330-012-2590-3

[41] Galván-Tejada JI, Celaya-Padilla JM, Treviño V, Tamez-Peña JG. Multivariate radiological-based models for the prediction of future knee pain: Data from the OAI. Computational and mathematical methods in medicine. 2015;2015. DOI: 10.1155/2015/794141

[42] Celaya Padilla, José M., et al. Local image registration a comparison for bilateral registration mammography. En IX International Seminar on Medical Information Processing and Analysis. International Society for Optics and Photonics, 2013; 8922: 892210. DOI:10.1117/12.2035516

ARTÍCULO DE INVESTIGACIÓN

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-13

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.2

E-LOCATION ID: **e201823**

Prototipo de Silla de Ruedas Dirigida Usando Parpadeos

A Wheel Chair Prototype Moved by means of Eye Blinks

M. M. Morín-Castillo, A. Santillán-Guzmán, S. L. Sainos-González, J. J. Oliveros-Oliveros

Benemérita Universidad Autónoma de Puebla

RESUMEN

El presente trabajo describe un prototipo de una silla de ruedas que es dirigido hacia enfrente y hacia atrás usando 2 o 3 parpadeos, respectivamente, y es detenido cuando se alcanzan ciertos niveles de atención. El objetivo principal es que las personas que tienen discapacidad motora en sus extremidades puedan usarlo para desplazarse y les brinde autonomía. Para captar la señal de los parpadeos, se utilizó la diadema *MindWave Mobile* de *Neurosky*. Se implementó un circuito electrónico en conjunto con *Arduino* que permite complementar la ejecución del accionamiento del prototipo. El prototipo se probó con 10 personas cuyas edades oscilan entre 20 y 35 años. Los resultados muestran que, en un 80% de los casos, el prototipo se mueve correctamente. La gran ventaja del presente trabajo es que la *interfaz cerebro-computadora* con la que cuenta este prototipo no requiere entrenamiento previo del sistema, por lo cual, puede ser usado por cualquier persona. Además, su costo es más accesible comparado con otros dispositivos para el mismo fin.

PALABRAS CLAVE: interfaz cerebro-computadora; prototipo; parpadeos

ABSTRACT

The present work describes a prototype of a wheel chair directed by means of eye blinks, which can be moved forwards, and backwards using 2 or 3 eye blinks, respectively, and stopped when a certain attention level is met. The main objective of this work is to help people, who have motor disabilities on their arms and legs, move and have autonomy. In order to register the eye blinking signals, the *MindWave Mobile* device from *Neurosky* was used. Moreover, an electronic circuit in combination with *Arduino* has been used to make the prototype work. This prototype has been tested in 10 healthy people from 20 to 35 years old. According to the results, in 80% of the cases the prototype worked correctly. The main advantage of the present work is that the *brain-computer interface*, which is part of the prototype, does not require training, and hence, it could be used by most of the people. Moreover, its cost is less than similar devices.

KEYWORDS: brain-computer interface; prototype; eye-blinks

Correspondencia

DESTINATARIO: **María Monserrat Morín Castillo**

INSTITUCIÓN: **Benemérita Universidad Autónoma de Puebla**

DIRECCIÓN: **Calle 4 Sur #104, Col. Centro Histórico, C.P. 72000, Puebla, Puebla,México**

CORREO ELECTRÓNICO: **maria.morin@correo.buap.mx**

Fecha de recepción:

5 de junio de 2018

Fecha de aceptación:

26 de octubre de 2018

INTRODUCCIÓN

La electroencefalografía (EEG) es una técnica no invasiva para el estudio de la actividad eléctrica del cerebro, tanto en sujetos sanos, como en pacientes que padecen epilepsia, Parkinson u otras enfermedades [1, 2]. Es bien sabido que el cerebro consiste en dos hemisferios, los cuales se subdividen en cuatro lóbulos (Frontal, Parietal, Temporal, Occipital) y la corteza cerebral, la cual es una capa de tejido gris donde se producen señales espontáneas, conocidas como *ritmos*. Éstos se encuentran ubicados en ciertas bandas de frecuencia, y dependiendo de la actividad que se esté realizando pueden ser registrados, a través de sistemas EEG [1-4]. Para referirse a los ritmos se usan letras griegas; siendo los más comunes *delta, theta, alpha* y *beta*. Cuando se registran las señales cerebrales éstas son acompañadas por distorsiones fisiológicas (producidas por el sujeto/paciente en estudio) o técnicas (producidas por el sistema de medición). Entre las fisiológicas están: Movimientos oculares, parpadeos, movimientos musculares, latido cardiaco, entre otras. En cuanto a las distorsiones técnicas, la más común es la que produce la línea de alimentación, que puede ser de 50 o 60 Hz, dependiendo del área geográfica en donde se realice la medición [5-7]. Generalmente, las señales contaminadas con distorsiones son filtradas usando filtros pasa-bajas o pasa-bandas, o mediante técnicas más complejas como la de *análisis de componentes independientes* (*ICA*, por sus siglas en inglés) para obtener una señal limpia a ser analizada posteriormente, como se describe en [6, 8-10]. En cualquier caso, la idea es obtener una señal cerebral limpia para continuar con su análisis y procesamiento. En el presente trabajo, sin embargo, se utilizan cierto tipo de distorsiones fisiológicas (parpadeos), las cuales son usadas como información de entrada para mover el prototipo de silla de ruedas, en forma remota. Para ello se ha desarrollado un circuito que permite la interfaz entre los parpadeos y el prototipo de silla de ruedas para que, dependiendo del número de parpadeos, éste se pueda desplazar hacia adelante o atrás; además, se utilizan los niveles de atención para detener la silla. La importancia de desarrollar este tipo de dispositivos es que son de gran ayuda a las personas que padecen una discapacidad motriz, en ambas extremidades, y que solamente son capaces de mover los músculos faciales y los ojos. Las personas con este problema, tienen una fuerte dependencia de las personas más cercanas a ellos, y por ende, tienen poca autonomía.

Existen en el mercado sillas de ruedas acompañadas con *exoesqueletos* que ayudan a tener una mejor calidad de vida de las personas con discapacidad motora. Un ejemplo de estos es ROKI, el cual es un aparato que se coloca en las piernas de las personas con discapacidad, dándoles soporte para levantarse y volver a caminar. De esta manera sirve como terapia de rehabilitación. Tiene un peso de 13 kg, aproximadamente (detalles de su funcionamiento se muestra en la publicación de Milenio [11]). Este tipo de dispositivos resultan útiles para personas con problemas motores en sus extremidades inferiores, pero que no presentan disfuncionalidad en sus miembros superiores, pudiendo así cargar ellos mismos el exoesqueleto y reincorporarse a su vida cotidiana después de unos días de terapia. No se requiere ningún entrenamiento del sistema de control; sin embargo, su costo resulta poco accesible para quienes han perdido movilidad en sus miembros inferiores.

Como se mencionó anteriormente, el presente trabajo presenta un prototipo de silla de ruedas que tiene como finalidad ayudar a personas con discapacidad en miembros superiores e inferiores basándose en una *interfaz cerebro-computadora* (*Brain-Computer Interface*, *BCI*, en inglés) con la cual se puede dirigir el prototipo de silla de ruedas usando el movimiento de los ojos (parpadeos).

Generalmente, las interfaces cerebro computadora usan señales evocadas en las que una vez que se hace una extracción de propiedades de las señales, se hace

una clasificación de las mismas y posteriormente mediante algoritmos específicos se toma una decisión para poder controlar algún dispositivo. Todo ello conlleva un entrenamiento del sujeto/paciente en estudio para que los algoritmos se adapten a sus propias señales. Esto requiere de tiempo y de ajuste de parámetros para que el sistema BCI se adapte a cada sujeto/paciente, como en Wolpan y otros [12, 13]. En Pantech solutions [14], se describen algunos proyectos en donde la diadema *MindWave Mobile de Neurosky* [15], la cual cuenta con un electrodo frontal y uno de referencia y misma que se usa en el presente trabajo, es utilizada para controlar diversos dispositivos a través de las señales de atención y meditación que el electrodo capta. Una aplicación más usando esta diadema se presenta en Calderón Martínez et at [16], en donde este tipo de *BCIs* no requieren de entrenamiento previo de los sujetos que la usen, lo cual representa una ventaja en tiempo y complejidad. Sin embargo, en este caso, para mover al carrito hacia la derecha, éste debe moverse primero a la izquierda con dos parpadeos, luego otros dos parpadeos regresan las llantas en posición recta y finalmente con otros dos parpadeos se mueve a la derecha. Esto podría ser una desventaja si se quiere que el carrito se desplace hacia la derecha en cierto momento y éste no puede ser dirigido inmediatamente sino hasta que la secuencia antes mencionada sea ejecutada.

En el trabajo de Yan Zhi [17], se realiza la caracterización de la señal de parpadeo para usarla como una toma decisión, para posteriormente usar el módulo *TGAM* del dispositivo *Neurosky* y realizar el desplazamiento del carro. En el presente trabajo, para evitar el entrenamiento del sistema y contrarrestando la desventaja antes mencionada, se usan los parpadeos para dirigir un prototipo de silla de ruedas hacia enfrente y atrás, y los niveles de atención para detenerlo, teniendo en cuenta un determinado número de paquetes de datos, enviados a través de la diadema usada. De esta manera, los movimientos se realizan casi en forma inmediata y son independientes uno de otro.

En la siguiente sección se presenta la caracterización de las señales usadas para la BCI propuesta. Posteriormente, se describe el sistema como tal, así como los resultados obtenidos y la discusión de los mismos. Finalmente, se enuncian las conclusiones y lo que se puede hacer en un futuro.

Movimientos oculares verticales

Los movimientos oculares verticales son movimientos naturales en las personas (conocidos también como parpadeos), los cuales se producen cuando se abren y cierran los ojos y son registrados por el EEG. Como se mencionó anteriormente, este tipo de movimientos pertenecen a la categoría de artefactos o distorsiones fisiológicas y se registran tanto en personas sanas como en personas enfermas, como muestran Tatum IV et al en [5] e Ille en [6]. Los parpadeos pueden ser fácilmente reconocidos debido a que tienen gran amplitud y se observan en el EEG principalmente en los electrodos frontales. Este tipo de distorsiones se reflejan en las bandas *delta* y *theta*, y, generalmente, son eliminadas, ya que se consideran señales de interferencia que afectan la medición. Cabe mencionar que existen diferentes métodos para eliminar este tipo de distorsiones, como por ejemplo aplicando una combinación de *ICA* con un filtro de tipo *Wiener* como lo describen Heute y Santillán en [18]. Sin embargo, en este trabajo, son los parpadeos las señales que se usarán para desplazar el prototipo de silla de ruedas. De esta manera, la persona que desplazará el prototipo usando la diadema, no requiere de entrenamiento para poder mandar alguna instrucción a cualquier dispositivo externo, como se hace comúnmente en las interfaces cerebro-computadora. Esto representa una ventaja en cuanto al tiempo de uso y control del dispositivo. Además, los parpadeos entre los diferentes sujetos no presentan cambios sustanciales, por lo que cualquier persona que presente alguna discapacidad de movimiento de sus extremidades puede usarlo y con el uso de sus ojos puede lograr mover o controlar algún dispositivo en forma remota.

Caracterización de las señales

Como primer paso, previo al desarrollo del algoritmo para mover el prototipo de silla de ruedas, se caracterizaron las señales provenientes de la diadema *MindWave Mobile* de *Neurosky*. Dicha diadema consta de un electrodo frontal y uno de referencia (posicionado en el lóbulo de la oreja). El electrodo es capaz de registrar las ondas cerebrales *delta, theta, alpha, beta* y *gama* (3-100 Hz), con una frecuencia de muestreo de 512 Hz. Además, registra los niveles de meditación y atención, y cuenta con un módulo *Bluetooth* para lograr la comunicación con la computadora o con algún otro dispositivo.

Para conocer las características de las señales que proporciona la diadema de *Neurosky*, se realizaron pruebas con 10 sujetos (4 mujeres, 6 hombres) de un rango de edad de 20 a 35 años, como se observa en la Tabla 1. Cabe mencionar que todos los sujetos de prueba otorgaron su consentimiento informado para poder realizar las pruebas.

Siguiendo el mismo protocolo, a cada sujeto se le pidió que realizara lo siguiente:

- Ponerse en posición cómoda y relajada.
- Realizar parpadeos normales.
- Realizar parpadeos provocados o forzados (considerando parpadeos rápidos y parpadeos con más duración de lo usual).
- Realizar movimientos de ojos a la izquierda y derecha.
- Realizar movimientos de cabeza (izquierda, derecha, arriba, abajo).
- Realizar gestos (fruncir el ceño, levantar cejas, fruncir nariz y fingir sonrisa).

Los parpadeos normales son los que se realizan de manera natural, con una amplitud promedio de 800x10-6 V y una duración de aproximadamente 0.3s. Por otro lado, los parpadeos provocados o forzados son aquéllos que se realizan voluntariamente cerrando los ojos con mayor fuerza y duración que los parpadeos normales.

Se utilizó el software *Open Vibe* [19] para el registro de las señales obtenidas con la diadema MindWave Mobile de *Neurosky*. Para analizar los datos adquiridos se utilizó el lenguaje de programación *MATLAB*, con el cual se caracterizaron las señales registradas (amplitud y duración de los parpadeos).

TABLA 1. Registro de sujetos para caracterizar la señal.

REGISTRO DE SUJETOS PARA CARACTERIZAR LA SEÑAL			
EDAD	#SUJETOS	#MUJER	#HOMBRE
20 años	1	0	1
23 años	2	1	1
25 años	1	0	1
27 años	1	1	0
29 años	2	1	1
31 años	1	0	1
33 años	1	0	1
35 años	1	1	0

Se realizaron pruebas como las de movimiento de cabeza, movimientos horizontales de ojos y gestos, con la finalidad de caracterizarlas y evitar una falsa detección de parpadeos forzados. Dichos movimientos son los que más pueden incidir o afectar la medición.

Asimismo, para caracterizar los niveles de atención de los sujetos de estudio, se les pidió concentrarse en algo, mirar un punto fijo y hacer respiraciones profundas.

En la Figura 1 se muestra un ejemplo de las señales registradas para su caracterización. Como se observa en la Figura 1a, el parpadeo normal, mostrado entre los segundos 4 y 6, tiene una duración aproximada de 0.2s

y una amplitud de 800x10^{-6}V. Los parpadeos forzados, mostrados en la Figura 1b, se presentan en los primeros 2 segundos, entre los segundos 6 y 10 y entre los segundos 14 y 16. En los tres casos, la duración promedio es de hasta 0.4s, y su amplitud es de hasta 2000x10^{-6}V. La Figura 1c presenta movimientos horizontales de ojos en los primero dos segundos, entre los segundos 10 y 12 y entre los segundos 14 y 16. Su morfología es parecida a los parpadeos normales, pero de menor amplitud (aproximadamente 600x10^{-6}V) y mayor duración (0.3s). En la Figura 1d se grafica el movimiento de cabeza, el cual se observa entre los segundos 2 y 4, teniendo una amplitud mayor de poco más de 900x10^{-6}V y una duración de 0.6s. Finalmente, en la Figura 1e se presentan los gestos producidos por el sujeto en estudio. En este caso, los gestos son observados a lo largo de los 16s, con una amplitud de hasta 2000x10^{-6}V. Dadas todas estas características, se puede observar que los parpadeos forzados tienen una mayor amplitud comparada con los parpadeos normales, movimientos horizontales de ojos y movimientos de cabeza; y pueden tener una amplitud similar con los gestos. Sin embargo, dada la duración de dos o tres parpadeos forzados (mismos que son usados para mover el prototipo hacia adelante o hacia atrás), es distinta que la de los gestos, por lo que con ello se evita tener una falsa detección.

La tarjeta Arduino cuenta con ciertos protocolos para la lectura de las señales EEG provenientes de la diadema *MindWave Mobile* de *Neurosky*, el cual las clasifica de la siguiente manera:

- 1x10-6-50x10-6V: Niveles normales.
- 50x10-6-80x10-6V: Niveles de Atención o Meditación.
- 80x10-6- 240x10-6V: Parpadeos.
- 240x10-6-255x10-6V: No hace contacto la diadema.

Se analizaron y compararon las mediciones de los parpadeos en *Arduino* con los datos recopilados con *Open Vibe*. En la Figura 2a se muestra la señal del electrodo. Entre los segundos 15 y 20, así como entre los segundos 30 y 35, se observan parpadeos forzados. Cuando la señal del electrodo muestra un valor de hasta 2000x10-6V, la señal de detección de parpadeos forzados (Figura 2b), alcanza un valor de hasta 240x10-6V, que, de acuerdo a la clasificación anterior corresponde a un parpadeo. La señal correspondiente a la detección de los parpadeos es la que se considera para el algoritmo propuesto y descrito en la siguiente sección.

La duración de los parpadeos, por su parte, está dada por una función de *Arduino* establecida por *Neurosky*, la cual muestra el tiempo de recepción de datos. En este caso, se hace uso de ella para determinar la duración de un parpadeo forzado. Se realizaron las pruebas mencionadas anteriormente (parpadeos normales, forzados, movimientos horizontales de ojos, movimientos de cabeza y gestos) a los 10 sujetos y se obtuvo el promedio de amplitud y duración entre todos ellos.

A su vez se corroboró la presencia de parpadeos normales y forzados mediante el cálculo de la varianza y desviación estándar de las señales.

De esta forma se pudo determinar la amplitud y duración promedio de los parpadeos forzados, tomando en cuenta las amplitudes y duración que provee *Arduino*. Analizando los datos se concluyó que un parpadeo forzado válido tiene una amplitud de 85x10^{-6}V a 240x10^{-6}V y una duración promedio de 0.4s (un parpadeo) a 1.2s (3 parpadeos).

De acuerdo a la señal del electrodo de la Figura 2a, entre los segundos 15 y 20 se muestran dos parpadeos forzados. En la Figura 2b, en esos mismos segundos se observan 4 líneas, indicando que se han detectado 4 parpadeos. Sin embargo, solamente 2 de las 4 líneas que se muestran son considerados como parpadeos forzados válidos de acuerdo a la clasificación previamente mencionada.

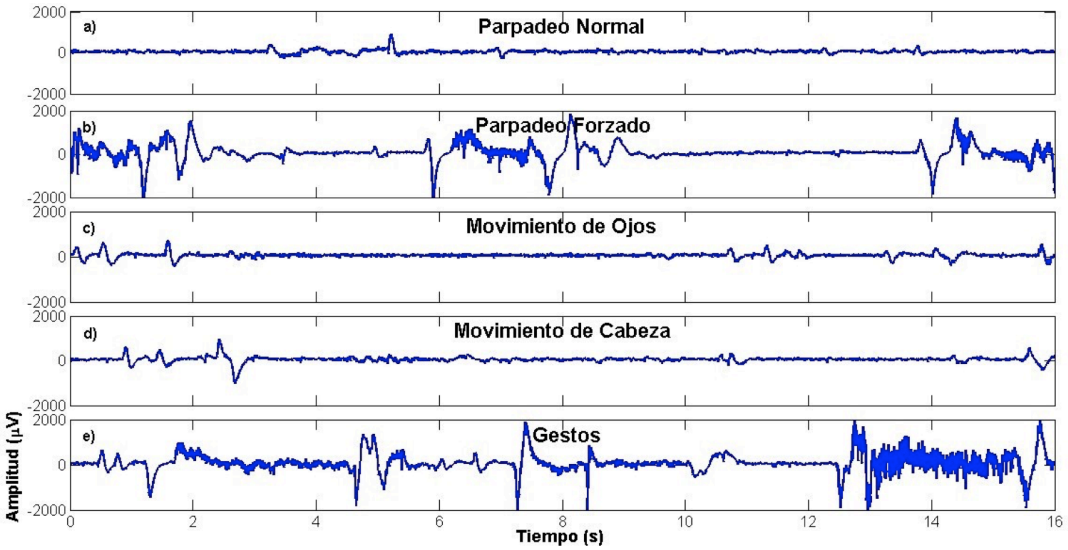


FIGURA 1. Señales obtenidas de un sujeto con la diadema MindWave Mobile de Neurosky: a) Parpadeo normal; b) Parpadeo forzado; c) Movimientos horizontales de ojos; d) Movimientos de cabeza; e) Gestos.

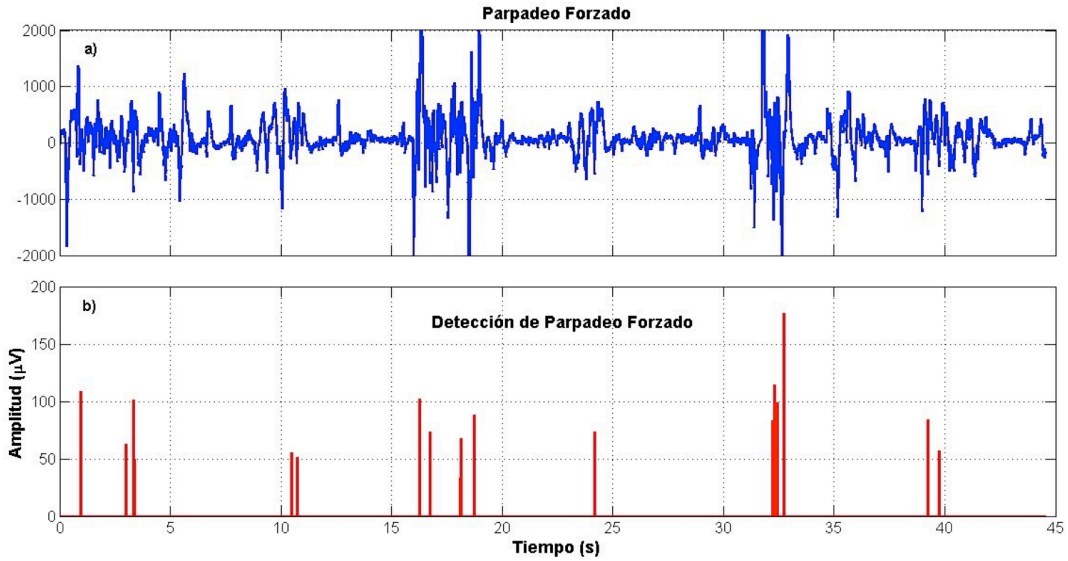


FIGURA 2. Señal de un sujeto obtenida con la diadema MindWave Mobile de Neurosky: a) Parpadeo forzado; b) Detección de parpadeo forzado.

Descripción del sistema

Una vez caracterizadas las señales a utilizar, éstas se usarán para desplazar el prototipo hacia adelante o hacia atrás. Además, se usarán los niveles de atención para detener el prototipo. La Figura 3 muestra el diagrama a bloques que representa gráficamente la secuencia de cómo está conformado el sistema o la interfaz cerebro-computadora.

Primeramente, se utiliza la diadema *MindWave Mobile* de *Neurosky* para hacer el registro de la actividad eléctrica del cerebro. Como se mencionó anteriormente, con dicha diadema es posible detectar también los parpadeos. Para la adquisición de los datos, se usó una tarjeta *Arduino Uno* de la marca *ATmega328* descrita en [20], que es una tarjeta microcontrolador con un hardware y software flexible, de lenguaje C para su programación

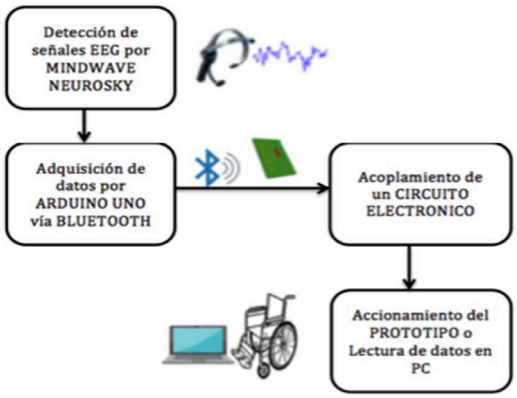


FIGURA 3. **Diagrama a bloques del sistema.**

(*Arduino Programming Language* y *Arduino Development Environment*). Para el presente trabajo se usó un programa base que *Neurosky* proporciona de su dispositivo *MindWave* en *Arduino Uno*, el cual se basa en usar los niveles de atención registrados con la diadema para encender una serie de 10 LEDs. Se instrumentó un circuito electrónico, el cual se muestra en la Figura 4, que incluye un módulo bluetooth *HC-05* para la recepción de datos, un dispositivo *L293D*, el cual es un puente H dual para el doble giro del motor y cuatro LEDS indicadores, los cuales se encenderán dependiendo del número de parpadeos forzados que se contabilicen o si los niveles de atención alcanzan cierto umbral. El prototipo se accionará de la siguiente manera:

- 1 parpadeo forzado válido, hay una detección de un parpadeo válido e inicia el incremento del número de paquetes (LED amarillo encendido);
- 2 parpadeos forzados válidos, prototipo avanza hacia adelante (LED verde encendido);
- 3 parpadeos forzados válidos, prototipo retrocede (LED rojo encendido);
- Niveles de atención superior a 50x10-6V, el prototipo se detiene (LED blanco encendido, LEDs restantes, apagados).

El motor utilizado para mover el prototipo, fue un *DC R140*, el cual cuenta con las siguientes características:

- Tensión: 1.5 ~ 3V.
- Velocidad (1.5V): 5700 / min.
- Velocidad (2V): 8200 / min.
- Velocidad (3V): 12400 / min.
- Conmutación: Con cepillo.
- Torque: 0.07kg / cm.
- Consumo con Carga: 0.7A.
- Consumo sin Carga: 0.370A.

Se ocupó un juego de 5 engranes para dar una mejor potencia en el desplazamiento del prototipo, como se ve en la Figura 5 y como se describe a continuación:

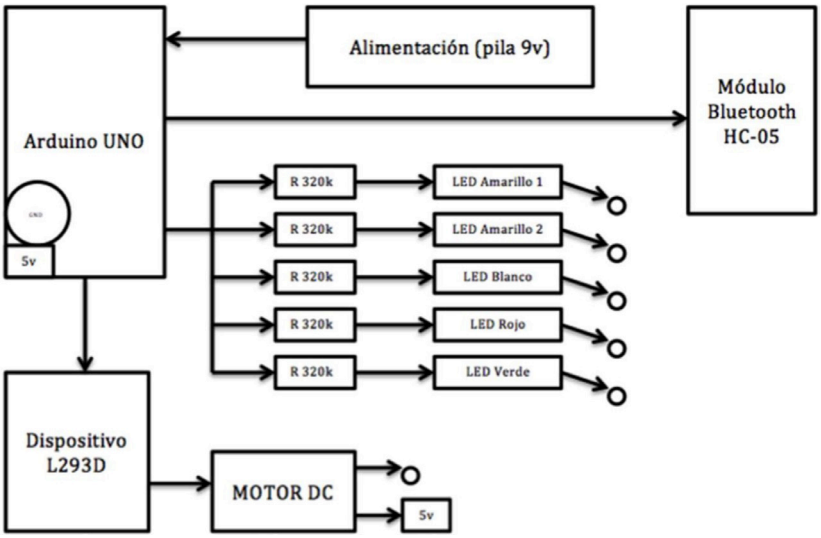


FIGURA 4. **Circuito electrónico.**

- 1 engrane de piñón (0.025m de diámetro);
- 4 de engranaje recto con piñón (0.025m de diámetro y base 0.015m);
- 2 ejes.

Todo se montó a una base de madera con 12 tornillos, más un tornillo con tuerca con el cual se sujeta a la parte inferior del prototipo.

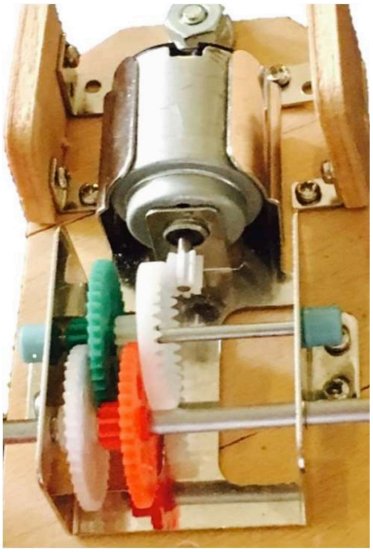


FIGURA 5. Montaje del sistema mecánico: Motor y engranes.

Dentro de los protocolos dictados por el sistema *Thinkgear* [21], el cual es el microchip de *Neurosky* implementado en la Diadema utilizada, se adquiere la información del sensor de la diadema por un flujo de paquetes, los cuales van de 4 a 173 bytes. El tiempo estimado de la entrega de cada paquete de datos es aproximadamente menos de $7\text{-}8\text{x}10^{-3}$s. En *Arduino*, es válido un paquete de datos que tiene un valor de 170 bytes con una longitud de carga útil mayor a 32 bytes para iniciar la adquisición.

Usando la tarjeta *Arduino*, se obtiene la señal del electrodo (SE) global, atención y meditación, las cuales se tomaron en cuenta para la ejecución del prototipo. Se contabilizó el número de paquetes para llevar a cabo las diversas acciones del prototipo: Hacia enfrente, hacia atrás, o detención. Se determinó que el número de paquetes que mejor funciona para el accionamiento del prototipo es de 20.

El sistema de accionamiento del prototipo funciona de la siguiente manera: Una vez colocada la diadema *MindWave*, así como el electrodo en la parte frontal, se hace la conexión con el *Arduino* y la computadora (*Arduino*-diadema) a través de *bluetooth* y se comienzan a recolectar datos. Como se mencionó, para mover el prototipo se necesitan 2 o 3 parpadeos forzados. Éstos deben estar en un rango de 85 a $240\text{x}10^{-6}$V de amplitud y tener una duración de entre 0.4 a 1.2s (dependiendo el número de parpadeos). Si ambas condiciones se cumplen, inicia el conteo de parpadeos y paquetes. Con el primer parpadeo válido, el LED amarillo se enciende indicando que el número de paquetes comienza a incrementarse. Este estado permanece así hasta que el número de paquetes es igual a 20, en donde se hace un conteo nuevamente del número de parpadeos válidos. Si el total de éstos es igual a 2, entonces el prototipo avanza hacia enfrente y el LED indicador verde enciende. Si son tres los parpadeos válidos, el LED indicador rojo enciende y el prototipo avanza hacia atrás. Cuando el número de paquetes es mayor a 20, todos los contadores se reinician (contador de parpadeos y de paquetes). El número de paquetes vuelve a incrementarse cuando las condiciones de amplitud y duración se cumplen, repitiéndose el ciclo hasta que la diadema es desconectada. Los niveles de atención se validan cada 20 paquetes, justo cuando se reinician los contadores. Si los niveles de atención están entre 50 y $100\text{x}10^{-6}$V, el prototipo hará alto. De lo contrario, el prototipo continuará con su ejecución de acuerdo al número de parpadeos previamente determinado.

RESULTADOS

La Figura 6 muestra las dimensiones del prototipo de silla de ruedas propuesto en este trabajo. Tiene un peso de 1.785kg. El material empleado para la parte externa que cubre a los circuitos es acrílico.

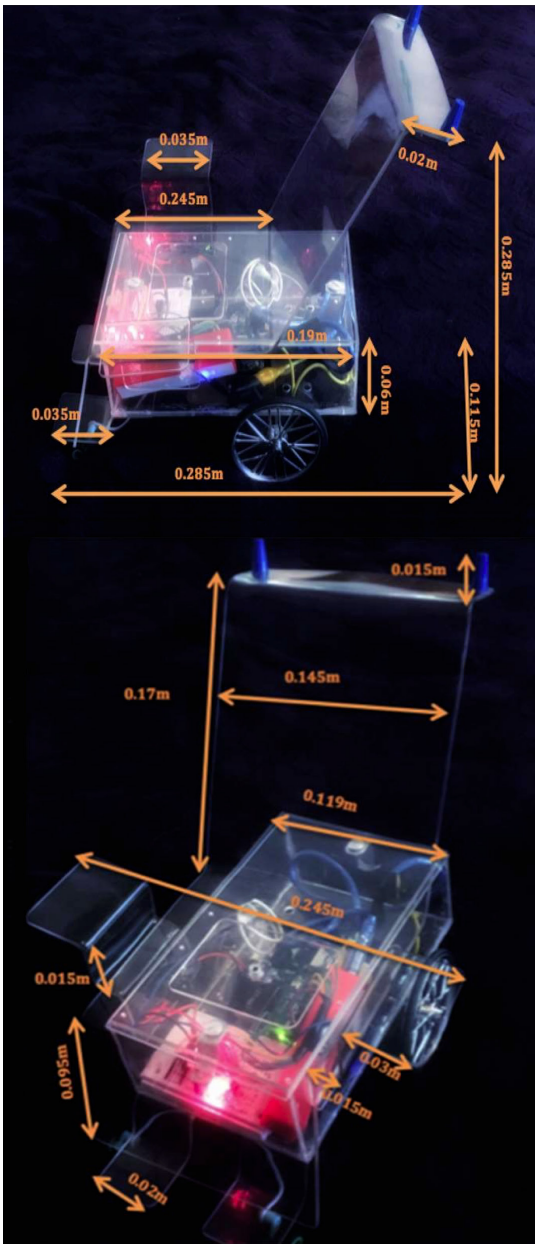


FIGURA 6. **Prototipo con el montaje del circuito electrónico y sus dimensiones.**

En la Tabla 2 se presenta un ejemplo del funcionamiento del sistema, mostrando los resultados obtenidos en la consola de Arduino. Este ejemplo corresponde a la prueba de funcionamiento del prototipo sobre un sujeto cuya edad es de 26 años. En dicha Tabla se observa que el número de parpadeos comienza a incrementarse cuando la amplitud de SE y la duración cumplen las condiciones mencionadas anteriormente, teniendo así el primer parpadeo válido, y por lo tanto el accionamiento del sistema, es decir, el inicio del incremento del número de paquetes, indicado por el encendido del LED amarillo, sombreado en la tabla con color gris. Este estado permanece así durante 20 paquetes. Cuando el número de paquetes es 20, se hace un conteo del número de parpadeos. En este caso, solo es un parpadeo, por lo que no pasa nada (quinta fila de la Tabla 2). Se reinicia el sistema de contadores y el LED amarillo se apaga. El número de paquetes se incrementa nuevamente, se enciende el LED amarillo indicando que se inicia el conteo de los 20 paquetes y cuando son 20 se vuelve hacer el conteo. Ahora el número de parpadeos es dos, por lo que el prototipo avanza hacia enfrente y se enciende el LED verde (sombreado en color verde claro en la Tabla 2). Este estado permanece así hasta que el siguiente bloque de 20 paquetes hace el conteo del número de parpadeos (sombreado en color verde fuerte en la Tabla 2). En este ejemplo, el número de parpadeos es tres, por lo que el LED rojo enciende y el prototipo avanza hacia atrás (sombreado en color rosa claro en la Tabla 2). Todo esto ocurre siempre y cuando en el reinicio de los contadores los niveles de atención estén por debajo de $50x10^{-6}$V. De lo contrario, el prototipo se parará y el LED blanco encenderá, como se presenta en la última línea de la Tabla 2.

DISCUSIÓN

Como se mencionó anteriormente, para mover el prototipo se necesitan 2 o 3 parpadeos forzados. Éstos deben estar en un rango de 85 a $240x10^{-6}$V de amplitud y tener una duración de entre 0.4 a 1.2s (dependiendo el número de parpadeos). Si ambas condiciones se cumplen, se comienza el conteo del número de parpadeos y el número de paquetes.

Al realizar diferentes pruebas del algoritmo con el prototipo de la silla de ruedas, las respuestas fueron satisfactorias un 80%. Se observó que no todas las personas solían elevar sus niveles de atención de la misma manera, esto con el fin de detener el prototipo.

TABLA 2. Ejemplo de ejecución del sistema para mover el prototipo de silla de ruedas.

No. Paquetes	SE (µV)	Duración	Parpadeos válidos	Color LEDs	Nivel Atención (µV)
0	63	52	0		10
1	127	430	1	A	5
2	54	200	1	A	6
…	…	…	…	…	…
19	0	0	1	A	45
0	0	0	0		0
…	…	…	…	…	…
4	179	427	1	A	34
5	54	1	1	A	26
6	146	802	2		43
…	…	…	…	…	…
19	15	5	2	A,V	7
0	0	0	0	V	20
…	…	…	…	A, V	…
3	200	457	1	A,V	21
4	169	946	2	A,V	40
…	…	…	…	…	…
7	15	64	2	A,V	0
18	198	1071	3	A,V	6
19	3	12	3	A, R	2
0	0	0	0	B	87

Un factor que influyó y que probablemente no se tuvo en cuenta en el momento, fue el estrés o nervios del sujeto al manipular el prototipo. Esto ocasionó que en algunos casos se elevara su atención y se detuviera el prototipo sin que se le indicara al realizar las pruebas.

No se requirió de tiempo de entrenamiento, para que las personas pudieran manipularlo, tan solo fue suficiente las indicaciones de colocárselo y hacer parpadeos forzados.

CONCLUSIONES

En el presente trabajo se presentó un prototipo a escala de silla de ruedas que es desplazado hacia enfrente y atrás de acuerdo a cierto número de parpadeos forzados válidos que un sujeto hace usando la diadema *MindWave Mobile* de *NeuroSky*. Para detectar el número de parpadeos forzados, se tomó en cuenta su amplitud, su duración y el número de paquetes en los que se hace el registro. Los niveles de atención se usaron para detener el prototipo. Todo el sistema se implementó en *Arduino* y se usó *bluetooth* para la comunicación entre la diadema y el sistema mismo.

Este prototipo puede ser robustecido usando otras métricas, tales como la *varianza*, la *curtosis*, el *exponente de Hurst*, que son útiles también en la detección de parpadeos. Asimismo se pretende modificar la forma de detención del prototipo, de tal manera que no sea tan subjetivo.

Además, como siguiente paso, se pretende implementar un algoritmo que permita movimientos a la derecha e izquierda, y a distintas velocidades, así como la implementación de un sistema de control que permita subir y bajar rampas.

Una vez implementado todo lo anterior, se llevará a cabo una etapa de potencia para poder realizar todo el procesamiento en una silla de ruedas real. De esta manera, se cubrirán las necesidades de pacientes que padecen alguna discapacidad en sus extremidades.

REFERENCIAS

[1] Sanei S, Chambers J.A. EEG Signal Processing. John Wiley &Sons, England, 2007.

[2] Santillán Guzmán A. Digital enhancement of EEG/MEG signals, PhD. dissertation, Christian-Albrechts Universität zu Kiel, Germany, 2013.

[3] Hämäläinen M, Hari R, Knuutila R. J, Lounasmaa O. V. Magnetoencephalography-theory, instrumentation, and applications to nonivasive studies of the working human brain, Reviews of Modern Physics, 1993 April; 65(2):413-497. DOI: 10.1103/revmodphys.65.413

[4] Daube J.R, Rubin D.I. Clinical 588 Neurophysiology. Oxford University Press, 2009.

[5] Tatum IV W.O, Husain A.M, Benbadis S.R, Kaplan P.W. Handbook of EEG Interpretation. Demos Medical Publishing, LLC, USA, 2007.

[6] Ille N. Artifact correction in continuous recordings of the electro- and magnetoencephalogram by spatial filtering, PhD dissertation, Mannheim University, 2001.

[7] Santillán Guzmán A, Heute U, Stephani U, Muhle H, Siniatchkin M, Galka A. Hybrid filter for removing power-supply artifacts from EEG signals, in 10th IASTED Conf. BioMed Eng., Innsbruck, IASTED, Acta Press, 2013, pp. 41-45. DOI: 10.2316/P.2013.791-022

[8] Common P. Independent component analysis, a new concept? Signal Processing, 1994; 36:287-314. https://doi.org/10.1016/0165-1684(94)90029-9

[9] Vigário R, Särelä J, Jousmäki V, Hämäläinen M, Oja E. Independent component approach to the analysis of EEG and MEG recordings. IEEE Transaction on Biomed. Eng. 2000 May; 47(5):589-593. DOI: 10.1109/10.841330

[10] Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. Neural Networks. 2000; 13:411-430. DOI: 10.1016/S0893-6080(00)00026-5

[11] Personas en silla de ruedas pueden volver a caminar con ROKI. Milenio. [Internet]. 2017 May. Available from: *http://www.milenio.com/region/exoesqueleto_roki-personas_discapacidad-Universidad_Panamericana_0_765523513.html*

[12] Wolpaw J.R, Birbaumerc N, McFarland D.J, Pfurtschellere G, Vaughana T.M. Brain-computer interfaces for communication and control. Clinical Neurophysiology. 2002; 113:767-791. DOI: https://doi.org/10.1016/S1388-2457(02)00057-3

[13] Mak J. N, Wolpaw J. R. Clinical Applications of Brain-Computer Interfaces: Current State and Future Prospects. IEEE Rev Biomed Eng. 2009; 2:187-199. DOI: 10.1109/RBME.2009.2035356

[14] Pantech solutions (Proyectos BCI). [Internet]. 2017 Jul. Available from: *https://www.pantechsolutions.net/brain-computer interface*

[15] Diadema Neurosky MindWave. [Internet]. 2017 May. Available from: *http://neurosky.com/*

[16] Calderón Martínez D. Procesamiento de ondas cerebrales con microprocesador ARM para control de coche teledirigido. Tesis de Licenciatura, Universidad de Sevilla, 2016.

[17] Zhi -Ming Y., Xiao-. Long W., Meng W. Jun W., Research and implementation of adaptive control method base don EEG. Conference on Applied Mechanics, Electronics and Mechatronics Engineering (AMEME 2016)

[18] Heute U, Santillán Guzmán A. Removing "Cleaned" Eyeblinking artifacts from EEG Measurements. in SPIN 2014, New Delhi, India, February 2014, pp. 576-580. DOI: 10.1109/SPIN.2014.6777020

[19] OpenVibe. [Internet]. 2017 May. Available from: *http://openvibe.inria.fr/*

[20] AT328 Datasheet. [Internet]. 2017 May. Available from: *http://www.alldatasheet.com/view.jsp?Searchword=ATMEGA328 &sField=4*

[21] Thinkgear NeuroSky, protocol. [Internet]. 2017 May. Available from: *http://developer.neurosky.com/docs/doku.php?id*

ARTÍCULO DE INVESTIGACIÓN

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-13

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.3

E-LOCATION ID: **e201822**

Reducción del Riesgo en Equipos Biomédicos y en Instalaciones Eléctricas de Entornos Clínicos

Risk Reduction in Electrical Networks and Safety of Biomedical Equipment in Clinical Settings

M. Arregui[1], *N. Alfaro*[1], *M. Baldizzoni*[2], *I. Wald*[2], *R. Gambogi*[2], *A. Ferreira*[2], *F. Simini*[1]

[1]Núcleo de Ingeniería Biomédica - Universidad de la República

[2]Institución Fondo Nacional de Recursos

RESUMEN

Se analizan 112 auditorías de instalaciones eléctricas y seguridad de equipos biomédicos en 78 Institutos de Medicina Altamente Especializada (IMAE) del Uruguay, realizadas a lo largo de 14 años, clasificando el nivel de riesgo y de cumplimiento de normas desde el punto de vista de Ingeniería Clínica. Cada visita incluye una encuesta al personal encargado de mantener y gestionar la infraestructura eléctrica y el equipamiento biomédico, que abarca el estado de mantenimiento, el control y la documentación de las instalaciones eléctricas y del equipamiento biomédico. Se evalúa el riesgo con un puntaje de 0 a 4. En 2004-2007 el 74% de los IMAE tenía irregularidades en la instalación eléctrica, gestión de equipamiento, control de calidad o documentación. Además, un 15% de los que tenían problemas, tenía en particular equipamiento indicado como "equipo peligroso". En los períodos siguientes esta proporción baja paulatinamente hasta 0% en 2016-2017. No obstante, continúa existiendo un déficit en la gestión del equipamiento y en la documentación formal. El aporte de la Universidad en el seguimiento técnico de los IMAE se ha materializado en una mejora en materia de seguridad.

PALABRAS CLAVE: Evaluación de Riesgo; Equipos Biomédicos; Instalación Eléctrica; Ingeniería Clínica; Mantenimiento

ABSTRACT

112 field inspections to 78 high technology medical centers (IMAE is the Spanish acronym) over 14 years are analyzed. All visits were evaluated as to Clinical Engineering good practices and were assigned a risk level. All audits included a questionaire to maintenance management personnel on electrical network operation as well as on biomedical equipment follow-up and documentation from acquisition to disposal. Risk is assigned a level 0 to 4 at each visit. In 2004-2007, 74% of IMAEs had safety problems in one or more of electrical network, maintenance management or documentation, and 15% of the IMAEs with safety problems had one piece of equipment described as simply "dangerous". Electrical safety problems were eventually reduced to 0% in 2016-2017, probably as a consequence of regular audit and counseling by this University Clinical Engineering Program.

KEYWORDS: Risk Evaluation; Biomedical Equipment; Electrical Network; Clinical Engineering; Maintenance

Correspondencia

DESTINATARIO: **Martín Arregui**

INSTITUCIÓN: **Núcleo de Ingeniería Biomédica - Universidad de la República**

DIRECCIÓN: **Sala 2, Piso 15, Hospital Clínicas, Av. Italia S/N, Montevideo, Uruguay**

CORREO ELECTRÓNICO: **marregui@fing.edu.uy**

Fecha de recepción:

12 de junio de 2018

Fecha de aceptación:

7 de noviembre de 2018

INTRODUCCIÓN

La Medicina del siglo XXI está en constante evolución e incorpora procedimientos nuevos con elevada frecuencia. La complejidad de los instrumentos de diagnóstico y tratamiento aumenta en relación directa con las nuevas tecnologías empleadas por la Ingeniería Biomédica en su diseño: materiales, electrónica miniaturizada, informática y telecomunicaciones [1].

Los objetivos aceptados colectivamente de mejora de la calidad de vida de todos los ciudadanos, la decisión de construir sociedades "inclusivas" que incorporen ciudadanos diferentes por habilidades, origen, decisiones y definiciones personales, junto a la creciente complejidad de los problemas que la tecnología está llamada a resolver, conlleva a que sea necesario un abordaje interdisciplinario que caracteriza nuestra época [2]. El Uruguay ha optado tempranamente por reconocer el alto costo y la alta complejidad de ciertos procedimientos médicos que no pueden ser resueltos por los prestadores privados pre-pagos, y tampoco por los prestadores integrales (estatales y privados) que componen el Sistema Nacional Integrado de Salud (SNIS) [3]. Es así que en 1980 se crearon los Institutos de Medicina Altamente Especializada (IMAE), financiados por un seguro universal denominado Fondo Nacional de Recursos (FNR), que funciona como Persona Pública no Estatal [4]. En 2007, una vez instaurado el SNIS, el FNR pasó a formar parte del modelo de financiamiento del SNIS, como seguro obligatorio complementario al Seguro Nacional de Salud, confirmando la cobertura financiera universal para los procedimientos de medicina altamente especializada y los medicamentos de alto costo. Los procedimientos de alto costo y de alta complejidad siguen siendo realizados por los IMAE, que son Centros privados o Servicios dentro de los prestadores integrales, ya sea estatales o privados. Los IMAE reciben el pago del FNR según aranceles negociados y acordados entre el Poder Ejecutivo y los IMAE. Cada procedimiento o medicación de alto costo es previamente autorizado por el FNR para cada paciente. La distribución del gasto del FNR se ha mantenido en proporciones similares de 80% para procedimientos médicos, 17% para medicamentos y 3% de gastos de administración.

Los actos médicos financiados en 2017 por el FNR [5] son: el tratamiento sustitutivo de la función renal (diálisis), los trasplantes renales, cardíacos, de hígado y pulmón, los trasplantes de precursores hematopoyéticos, varios procedimientos cardiológicos (cateterismos, angioplastias y cirugía cardíaca), tratamiento del gran quemado, el implante de marcapasos y de cardiodesfibriladores, el reemplazo de cadera y rodilla, los implantes cocleares y los medicamentos de alto precio. El FNR actualiza anualmente esta lista de prestaciones que financia, de acuerdo con un procedimiento establecido por ley, que incluye la evaluación previa por una Comisión Técnico Asesora. En 2007 el presupuesto del FNR era de 100 millones de dólares (30 dólares por habitante del país y por año) [6] y una década después fue de 201 millones dólares (2017), o sea 61 dólares por habitante y por año.

Para mantener la calidad de los procesos asistenciales de la salud, el FNR tiene desde 2004 un Programa de Evaluación y Seguimiento de los IMAE. Este Programa, cuyos procedimientos están normalizados [7], incluye la evaluación de la seguridad eléctrica que es objeto de la presente publicación.

El equipamiento biomédico debe garantizar que los procedimientos se realicen con seguridad y un mínimo de riesgo tanto para pacientes como para operadores. Un alto porcentaje de los eventos adversos que ocurren en la atención sanitaria está relacionado con el uso de tecnologías biomédicas [8]. Por tal motivo es importante conocer el estado de las instalaciones eléctricas, las características del equipamiento así como el nivel de entrenamiento y capacitación del personal que los usa. Para auditar estos aspectos, el FNR y la Universidad de la República acordaron en 2004 [9] que el Núcleo de

Ingeniería Biomédica (NIB) tomara parte activa en el Programa. El NIB evalúa a) el cumplimiento de normas de seguridad de equipos, b) la seguridad de instalaciones eléctricas y c) la gestión del mantenimiento.

La estrategia conjunta es la de verificar el cumplimiento de normas de seguridad y de buenas prácticas, involucrando a las partes interesadas en la incorporación de modalidades de gestión, adquisición de instrumental biomédico y capacitación de personal que lo usa. Este enfoque, propuesto en 2004, es retomado oficialmente por el Ministerio de Salud Pública en 2017, haciéndose eco de una normativa MERCOSUR [10]. Esta normativa está generalizando los criterios de garantía de calidad a todos los centros de salud, ampliando el conjunto inicialmente limitado a los IMAE.

Nos proponemos analizar la tarea compartida a lo largo de 14 años de seguimiento de los IMAE (2004 a 2017) en relación a la gestión de mantenimiento de los equipos biomédicos y al estado de las instalaciones eléctricas.

METODOLOGÍA

Visita Interdisciplinaria

El FNR y el NIB conformaron un equipo interdisciplinario (donde participa personal de la salud de diversas especialidades e ingenieros de formación biomédica o en ingeniería clínica) con el fin de controlar la calidad de los servicios brindados por los IMAE. El equipo planificó las visitas interdisciplinarias FNR-NIB durante las cuales verificó la información en registros y documentos. En particular, los ingenieros del NIB entrevistaron a los responsables de mantener, administrar y controlar las instalaciones eléctricas y los equipos electrónicos biomédicos. Luego de la entrevista inicial, recorrieron las áreas de interés tomando notas y fotografías que evidencien lo declarado por los entrevistados y recogen la documentación que describe la gestión del equipamiento y de las instalaciones eléctricas.

El Ministerio de Salud Pública (MSP) ejerce su función general de inspección de los centros médicos del país, de los cuales los IMAE son un subconjunto de alta especialización. Al igual que todos los centros médicos del país, los IMAE son sujetos a la inspección del MSP que, por acuerdo, está incluida en la visita interdisciplinaria FNR-NIB. Se relevaron por lo tanto datos de orden general destinados al MSP, según el detalle de la Tabla 1. Estos datos constituyen un subconjunto de los datos referentes al aseguramiento de la calidad que son objeto de este estudio y que son relevados en los IMAE. Estos datos están agrupados en categorías o dimensiones presentadas en la Tabla 2.

El relevamiento realizado en cada visita a los IMAE se refiere al grado de cumplimiento de estándares en sus aspectos de seguridad. Las dimensiones contempladas en las auditorías (Tabla 2) abarcan desde la verificación de las instalaciones eléctricas hasta las políticas de redundancia y de incorporación tecnológica, de acuerdo al siguiente detalle:

1. Instalación eléctrica segura
2. Especificaciones de compra y adquisición
3. Equipamiento biomédico controlado
4. Estado de funcionamiento de los equipos
5. Operación por personal certificado
6. Trazabilidad del uso de equipos biomédicos
7. Continuidad del servicio
8. Revisión formal de procedimientos

Método de análisis

Durante la visita se realiza una encuesta e inspección que abarca distintos items referidos al estado actual, mantenimiento, control y documentación tanto de las instalaciones eléctricas como del equipamiento electrónico biomédico del IMAE. Se contemplan además aspectos de gestión y buenas prácticas del servicio. Esta información es concentrada en una planilla electrónica mediante una puntuación en cada item que refleja la gravedad de la carencia, es decir la mayor

TABLA 1. Aspectos de Gestión de Mantenimiento de Equipos Biomédicos para MSP

Descripción	Cumple	No cumple
Cuenta con un procedimiento documentado que asegure que los equipos incorporados y utilizados sean adecuados a las necesidades de los estudios o tratamientos realizados.		
Cuenta con una colección de registros asociado a cada equipo, que refleje las entradas y salidas de servicio y las intervenciones preventivas y correctivas.		
Cuenta con mecanismos para evitar que se utilicen equipos que requieran revisión, calibración, mantenimiento preventivo o correctivo.		
Cuenta con un plan de mantenimiento u otro mecanismo que asegure que las intervenciones preventivas, calibraciones o revisiones se realizan antes que el uso del equipo represente un riesgo.		
Las fechas de las intervenciones de calibración o mantenimiento preventivo planificadas deben estar visibles para los operadores del equipo.		
Si la responsabilidad sobre el mantenimiento es transferida a un tercero, debe haber un contrato formal.		
Cuenta con manuales operativos de equipos en español.		
La instalación eléctrica cumple las recomendaciones de UTE.		
Las instalaciones eléctricas son adecuadas para el equipamiento instalado.		
El estado de las instalaciones y el funcionamiento de las protecciones son verificados periódicamente.		
Se registran los eventos asociados a la instalación de los equipos y en especial las situaciones inusuales o inesperadas, las sospechas de mal funcionamiento y las intervenciones sobre la instalación.		
La instalación eléctrica cuenta con un sistema de energía alternativo que permite no interrumpir procedimientos y no derivar pacientes a otros centros.		
El equipamiento es incorporado teniendo en cuenta el concepto de redundancia para asegurar la continuidad del servicio.		
La empresa de mantenimiento cuenta con certificación y registro en MSP.		
El IMAE cuenta con plan de contingencia frente a incendios.		
El IMAE cuenta con habilitación de bomberos para sus locales.		

Nota.- Estos aspectos de gestión agrupados constituyen la mayoría de las DIMENSIONES del relevamiento.

puntuación refleja la mayor gravedad. La gravedad de los problemas es evaluada según la escala de la Tabla 3. En cada dimensión de la inspección se anotan las desviaciones de lo esperado en forma numérica, simplificada. Un valor nulo es indicador de cumplimiento, "2" es "fuera de norma" y el "1" indica una situación intermedia en el rubro considerado.

Para que el resultado global de la visita sea fácilmente interpretable y se jerarquicen los elementos graves, ideamos un sistema de puntaje de gravedad decreciente, desde la peligrosidad de una instalación deficiente hasta detalles de procedimientos sobre la gestión de los equipos [11].

La gravedad creciente de riesgo detectado puede ser resumida con un simple puntaje de la situación de cada IMAE en una determinada visita. Así, los IMAE con puntaje cero (0) no sugieren mayores observaciones en cuanto al manejo de su seguridad, de las instalaciones eléctricas y de los equipos biomédicos. Un puntaje de 1 indica que se debe mejorar la documentación de los equipos. Un puntaje de 2 significa que se encontraron defectos en la instalación eléctrica, como malas puestas a tierra o aislamientos precarios, entre otros. El puntaje 3 significa que la gestión del equipamiento es precaria o nula, por ejemplo, cuando no hay planes de mantenimiento y verificaciones documentadas y fiscalizadas por parte de la institución. Finalmente, la

TABLA 2. Dimensiones a Evaluar en Cada Institución Médica Visitada

1	**Instalaciones eléctricas en condiciones adecuadas**	La organización debe asegurar que las instalaciones eléctricas donde se conecta el equipamiento, son adecuadas a las necesidades.
2	**Equipamiento adecuado**	La organización debe asegurar que los equipos utilizados para realizar estudios y tratamientos sean adecuados a las necesidades y que se mantienen tales durante toda su vida útil.
3	**Equipamiento controlado**	La organización debe asegurar que los equipos utilizados para realizar estudios y tratamientos se mantienen controlados durante toda su vida útil.
4	**Equipamiento en correcto estado de funcionamiento**	La organización debe asegurar que los equipos utilizados para realizar estudios y tratamientos funcionan correctamente.
5	**Equipamiento operado correctamente**	La organización debe asegurar que los equipos utilizados para realizar estudios y tratamientos son operados correctamente.
6	**Trazabilidad del uso de equipamiento**	La organización debe asegurar que existe trazabilidad entre los estudios y tratamientos realizados con cada equipo, el estado de funcionamiento del equipo y calificación del operador.
7	**Continuidad del servicio**	La organización debe asegurar la continuidad de las prestaciones durante todo el período de servicio.
8	**Revisión formal de procedimientos**	La organización debe asegurar que el proceso con el que gestiona el uso de los equipos, se mantiene actualizado a lo largo de todo el período de servicio.

TABLA 3. Puntaje Global Conforme a la Gravedad del Riesgo

	Gravedad del riesgo
Documentación incompleta	1
Defectos en la instalación eléctrica	2
Gestión de equipos precaria o nula	3
Equipo peligroso	4

situación que hemos calificado como más grave -puntaje 4- es cuando se encuentra en la institución el uso de equipo inadecuado o en malas condiciones que representa un riesgo inminente para la seguridad de pacientes y operadores.

Los puntajes globales de peligrosidad fueron asignados a juicio del auditor, en reunión posterior de redacción del informe (2004-2007) o a posteriori, releyendo los informes (2008-2012). Desde 2013, los puntajes globales (0 a 4) de cada IMAE en cada visita son resultado de una combinación de los puntajes asignados en cada sub dimensión de las dimensiones indicadas en la Tabla 2 (cada sub-dimensión recibe un puntaje de 0 a 2 de acuerdo al grado de cumplimiento). Para poder estudiar en forma conjunta los resultados de las visitas de 2008 a 2017, las definiciones de gravedad de la Tabla 3 son compatibles con los puntajes de los primeros años (2004 a 2007) [11]. Las visitas interdisciplinarias son realizadas a los IMAE agrupados en tipos de procedimiento médico según el detalle de la Tabla 4.

Analizamos informes emitidos de 2004 a 2017 y estudiamos las recomendaciones contenidas en cada uno y en secuencia por cada IMAE. Si no hay observaciones, la gravedad es indicada como cero (0). El no cumplimiento de menor gravedad es la falta de documentación sobre la gestión del mantenimiento y tiene "gravedad de riesgo=1". La mayor gravedad (4) es registrada cuando hay equipos peligrosos con un elevado riesgo de causar daño. Esta atribución progresiva de gravedad, en la cual se registra solamente el nivel más

TABLA 4. IMAE Agrupados por Tipo y Visitas Interdisciplinarias 2004-2017

Tipo de IMAE	Acrónimo de IMAE	2004 - 2007	2008-2011	2012-2015	2016-2017	2004-2017
Sala Blanca: artroplastia cadera y rodilla	ACR	1	4	3	7	15
Block Quirúrgico y Centro de materiales	BLQ	0	7	11	0	18
Centro de Tratamiento Intensivo (CTI), polivalentes, cardiológico y de quemados	CTI	1	6	0	0	7
Diálisis renal	DIA	10	4	3	7	24
Hemodinamia y Angioplastia	HYA	17	4	6	2	29
Marcapasos y cardiodesfibriladores	MYC	0	0	1	10	11
Transplante renal	TR	0	0	3	0	3
Litotricia	LITO	5	0	0	0	5
TOTAL	**---**	**34**	**25**	**27**	**26**	**112**

Nota.- IMAE, Institutos de Medicina Altamente Especializada.

TABLA 5. Evolución del Riesgo en Centros Médicos IMAE

Periodo	2004 - 2007	2008 - 2011	2012 - 2015	2016 - 2017
IMAE visitados	34	25	27	26
Sin observaciones	26%	16%	4%	4%
Documentación incompleta	23%	4%	15%	19%
Defectos en la instalación eléctrica	23%	8%	0%	23%
Gestión de equipos precaria o nula	17%	12%	74%	54%
Equipo peligroso	11%	60%	7%	0%
TOTAL	100%	100%	100%	100%

Nota.- Para el estudio se asignaron los siguientes puntajes: Sin observaciones = 0, Documentación incompleta = 1, Defectos en la instalación eléctrica = 2, Gestión de equipamiento biomédico precaria o nula = 3, Equipo peligroso = 4. Se anota el puntaje más alto para cada IMAE en un dado momento.

alto encontrado, responde a la necesidad de subsanar lo urgente, antes que nada, dándole la máxima prioridad, en la óptica de reducir riesgos. En consecuencia, la indicación de una gravedad elevada hace que se espere encontrar en las visitas sucesivas un efecto de reducción de riesgo, atribuible al seguimiento del grupo interdisciplinario. Por ejemplo, un IMAE con instalación eléctrica defectuosa y equipos en condiciones peligrosas tiene una gravedad de "4", al igual que otro IMAE que "solamente" tiene equipos peligrosos. Un IMAE que tiene una buena instalación eléctrica y ningún equipo en estado de peligro inminente, pero fallas en la gestión de su mantenimiento, tiene una gravedad de "3". De igual manera se combinan las situaciones en los aspectos de la Tabla 3, en una única "gravedad" que refleja lo urgente e impostergable.

La gravedad del riesgo es tenida en cuenta de acuerdo a las recomendaciones de la OMS [12, 13] y en particular se hace hincapié en los incumplimientos de la normativa vigente en Uruguay, principalmente la que se refiere a la instalación eléctrica [14-16] emanada de la Empresa Estatal de generación y distribución de energía eléctrica UTE.

RESULTADOS

El ritmo de visita anuales (Tabla 5) se ha mantenido constante con una visita a un IMAE al azar cada dos meses durante 12 años hasta 2015, año en que el ritmo de visitas aumenta a una visita mensual.

En la Figura 1 se resume el nivel de riesgo de los IMAE agrupados por tipo. El nivel de riesgo bajo (por debajo de 1) de las salas blancas en 2008-2015 contrasta con la situación de los blocks quirúrgicos, hemodinamia y diálisis, que en promedio mejoraron su situación recién en 2016-2017. Las salas blancas, por otra parte, aparecen con niveles aumentados recientemente, cuyas razones serán comentadas en la sección de "Discusión".

En la Figura 2, en el periodo 2004 a 2007 se ve que las observaciones confirmaron las limitaciones que existían en los IMAE. La primera columna de la Tabla 5 puede ser tomada como línea de base del análisis siguiente. El 74% de los IMAE auditados tenían aspectos a mejorar en lo que refiere a la instalación eléctrica, gestión de equipamiento, control de calidad o documentación [11]. El análisis de 2007 indicaba que el 11% tenía algún equipamiento "peligroso". Se trata de una situación que podía llevar a accidentes, que sucedieron por cierto antes de 2004 y que fueron el disparador para poner en práctica el Convenio FNR-UR informado en el presente artículo. Es de notar que las 34 visitas de este período abarcaron fundamentalmente centros de diálisis, salas de hemodinamia/angioplastia y litotricia.

En el periodo 2008 a 2011 la composición de IMAE visitados incluye centros de mayor complejidad como salas blancas y centros de medicina intensiva (CTI), lo que tuvo como consecuencia que el porcentaje de IMAE con equipos en situación de "peligro" aumentara al 60%.

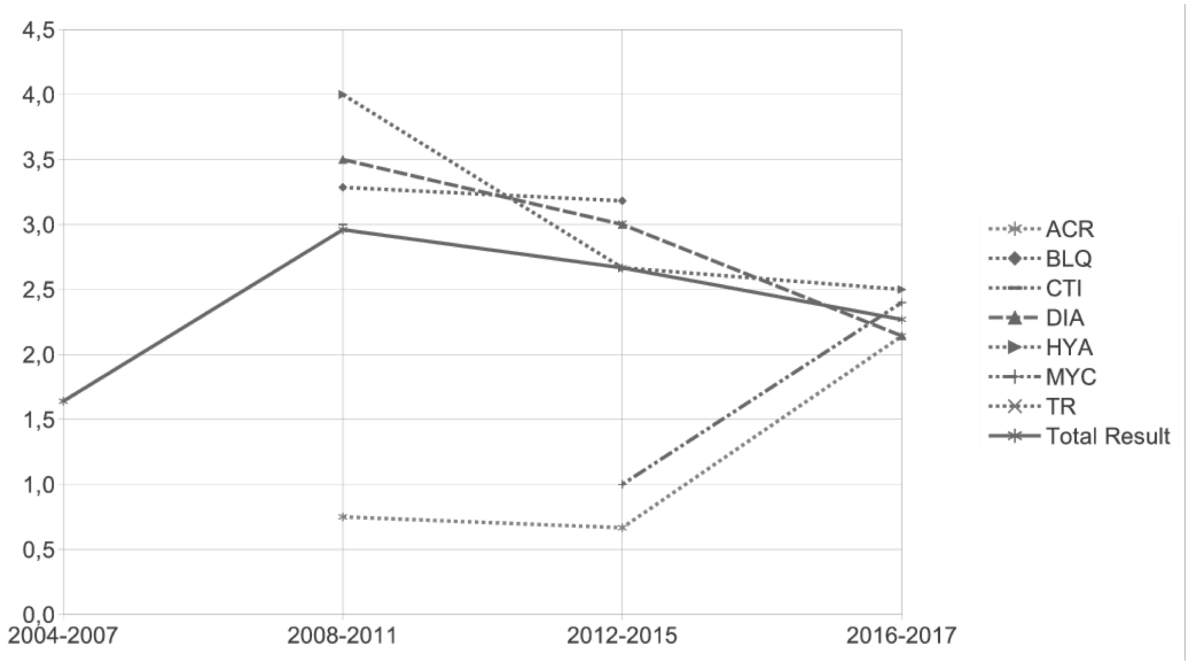


FIGURA 1. Gravedad de situación en cuatro períodos 2004-2016 por tipo de IMAE (siglas en Tabla 4). No se distinguen el riesgo en cada IMAE en el primer período.

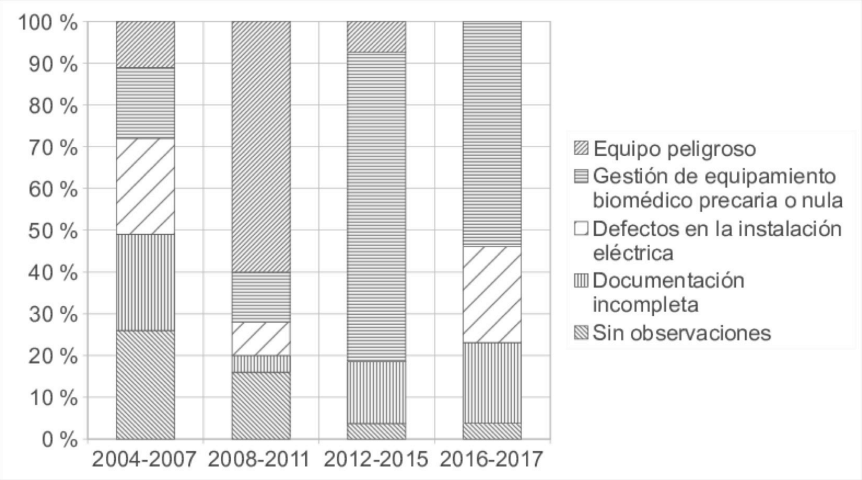


FIGURA 2. Evolución de nivel de riesgo en IMAE por tipo. La composición de niveles de riesgo ha variado en tres lustros: equipos peligrosos dieron lugar a carencias de gestión y documentación. Aparece la necesidad de mejoras de alto nivel en las instalaciones eléctricas.

En 2012-2015, las instituciones que presentaban equipos peligrosos descendieron drásticamente de 60% a 7%. Sin embargo, 74% de los IMAE presentaron déficit en la gestión del equipamiento.

En 2016-2017, no se encontraron equipos inadecuados ni en situación de "peligro". Se observa una mejora en la gestión de equipamiento que pasa de 74% deficitaria en 2012-2015, a 54%. No obstante, sigue siendo el factor dominante.

DISCUSIÓN Y CONCLUSIÓN

En la situación actual de los IMAE del país, cabe destacar que no se han encontrado equipos que presenten peligro para pacientes y operarios. La situación ha ido mejorando desde los primeros 2 períodos de visitas (2004 a 2011), donde se encontraban todavía instituciones que empleaban equipos inapropiados, en mal estado o con deficiente aislación eléctrica, condiciones que generan riesgo de electrocución o de otro daño. Es posible inferir, de los datos que aportamos aquí, que las instituciones en su conjunto han respondido a la asesoría FNR-NIB y han corregido sus prácticas fundamentales, mejorando la seguridad eléctrica.

La buena gestión del equipamiento biomédico es una situación altamente deseable y sin embargo la Tabla 5 refleja un aparente empeoramiento desde 2004 hasta 2017. Esto es el resultado de un corrimiento de los problemas hacia la gestión, una vez que se fue solucionando el elevado riesgo de instalaciones y equipos. En 2008-2011 había 60% de centros con equipos peligrosos, reducidos a 7% en 2012-2015, a la vez que los problemas de gestión rudimentaria crecieron de 12% a 74% en universos similares de 25 y 27 IMAE respectivamente. La modalidad de registro en los niveles de 0 a 4 era tal que con un equipo "peligroso" no se notaran problemas menores como una "mala gestión de equipamientos", como se mencionó en la Metodología. Esto lleva a que, una vez resueltos los problemas de peligrosidad eléctrica, aparecieran los problemas de gestión, dando un aparente aumento de su peso en la problemática global. En ese momento (2013), la herramienta de registro de las visitas interdisciplinarias fue mejorada con una planilla dividida en dimensiones y sub dimensiones, inspiradas del manejo de la calidad total [17]. Por lo tanto, existió un control de calidad de los equipos ejercido con mayor rigurosidad y formalismo en la línea sugerida por las normas ISO 9001-

2015. Se verificó el cumplimiento de las recomendaciones de fábrica, el cumplimento de la documentación formal de procedimientos en la actividad diaria, el cumplimiento de los planes de mantenimiento y de la trazabilidad del personal, de los estudios y del equipamiento. Esto llevó a que se calificara negativamente en el último período (2016-2017) un 54% de las instituciones con déficit en control de calidad de los equipos.

Típicamente, en 2017 se debe insistir aún en la implementación de la identificación única de cada equipo, en la visibilidad de las fechas de vencimiento de los mantenimientos y en la puesta en práctica de la historia clínica de cada equipo [12] así como también se debe exigir mayor rigurosidad en la implementación y en la documentación del plan de mantenimiento.

El estado de las instalaciones eléctricas es también un aspecto crítico, que llevó 14 años de seguimiento para lograr el marcado decaimiento de la peligrosidad de las instalaciones que se deduce de la Tabla 5. Esto responde a la atención de un estado de emergencia en el que muchas de las instituciones se encontraban en 2004-2011 y que fue señalado por la ejecución del plan de las visitas interdisciplinarias FNR-NIB iniciado por el Dr. Alvaro Haretche desde 2004. En esa época se puntuaban como defectos en la instalación los cables colgando fuera de sus canalizaciones, la ausencia de protecciones básicas como llaves diferenciales o termo-magnéticas intempestivamente suprimidas por un mal mantenimiento, toma corrientes flojos y fuera de su cubículo con partes electrizadas accesibles al usuario. En los últimos años (desde 2013 típicamente), en que este tipo de defectos ha sido superado, se ha hecho mayor énfasis en el cumplimiento de normas vigentes de mayor exigencia: por ejemplo, los transformadores de aislamiento y los sistemas de respaldo de energía, fundamentales en instalaciones para cirugía y otros procedimientos. Esta nueva exigencia explica el aparente aumento del puntaje del grupo de Salas Blancas de la Figura 1 entre 2012-2015 y 2016-2017.

Superados los problemas más graves, se puede mejorar la eficiencia de gestión y buscar aún mejores resultados. Las visitas conjuntas FNR-NIB han sugerido la mejora de la documentación de los eventos de mantenimiento y el seguimiento mediante indicadores de producción. Esto les permitiría a las instituciones seguir progresando en su evolución hacia estándares de excelencia en gestión de equipos biomédicos. Con estos indicadores, los IMAE tendrán la información necesaria para una mejor toma de decisiones, permitiéndoles además del aseguramiento de la calidad de atención médica, la disminución de costos vinculados a fallas imprevistas e incluso posibles demandas por mala gestión. Es también muy importante que la institución posea la documentación de una clara distribución de responsabilidades entre los diferentes actores, como son: personal de mantenimiento, encargados de gestión, operadores, empresas contratadas para mantenimiento y proveedores de equipo nuevo. Por nuestra parte, como entidad auditora, sería interesante someter a evaluación externa la propia herramienta de registro, como hicieron los colegas de las Universidades EIA y CES de Colombia [18]. La evaluación de su herramienta de auditoría HA contempla 7 dimensiones que abarcan desde la pertinencia, claridad, relevancia y capacidad para la toma de decisiones hasta la facilidad de uso. Estas dimensiones podrán servir como guía en ocasión de una futura evaluación de nuestro sistema.

El bajo porcentaje (4%) de IMAE que en 2008-2011 tuvieron deficiencias de documentación es consecuencia de que se trata de IMAE cuyo PEOR rasgo era la falta de documentos. Las demás carencias de estas pocas instituciones estaban resueltas, quedando únicamente el “detalle” de la documentación. Por lo tanto, es dable suponer que en el grupo mayoritario (representado por el 60% de IMAE en la Tabla 5) que tenían equipos peligrosos, también tuvieran a mayor razón una documentación incompleta, hecho que fue considerado en su momento un problema menor en la inducción gradual de una política de seguridad.

El seguimiento de unos 78 IMAE en más de 112 visitas muestra una mejora incuestionable en seguridad eléctrica desde la situación original, con accidentes y lesiones a pacientes, que motivó el convenio FNR-NIB. Sin embargo, el tiempo trascurrido para lograr una situación que está aún lejos de la seguridad certificada que se desea, indica que la inducción de buenas prácticas mediante visitas periódicas es necesaria pero lenta. El objetivo de implementar acompañamiento y docencia fue cumplido, y el camino queda allanado para acciones que instalen los estándares de calidad, y en primer lugar de la documentación formal de procedimientos. El reciente decreto del Ministerio de Salud Pública [10], si bien es de carácter general para todos los centros de salud, sean de complejidad simple o alta, es un primer paso en la dirección que sugiere este estudio. El decreto establece objetivos de buena gestión en el manejo de equipos biomédicos.

Podría pensarse en la difusión de indicadores de buena gestión de equipos biomédicos y buena tenencia de la tecnología en general como inductor indirecto de superación en seguridad eléctrica por parte de las instituciones de salud. En efecto, el usuario que compara las instituciones evaluando varios parámetros, tendría a su disposición también el del buen manejo de instalaciones y equipos. Si bien se trata de un guarismo especializado y parcial, puede contribuir a comparar elementos de calidad de los IMAE para el público.

Completando la acción de formación y acompañamiento de los IMAE, se podría considerar la exigencia de cursar asignaturas específicas en Ingeniería Clínica [19], además de la de contar obligatoriamente con personal con esa capacitación en cada IMAE. Partiendo de un apoyo profesional en ingeniería biomédica muy bajo en 2004, la disponibilidad de técnicos egresados de la Universidad en los últimos tres lustros permite, ahora sí, sugerir que tengan responsabilidades profesionales directas en los IMAE.

AGRADECIMIENTOS

Los autores agradecen al equipo de auditoras y auditores del Fondo Nacional de Recursos y del Ministerio de Salud Pública. Gran parte de las rutinas y modalidades fueron desarrolladas en conjunto con la MSc. Cándida Scarpitta, a quien extendemos el más caluroso agradecimiento. Las visitas fueron realizadas, además de por los propios autores, por los ingenieros Jorge Lobo, Daniel Geido, Gustavo De Martino, Gonzalo Carballo y Adrián Monkas, a quienes se le deben los datos, evaluaciones y sugerencias como docentes del NIB.

Este trabajo está dedicado a la memoria del Dr. Alvaro Haretche, que tuvo la visión de recurrir a la multiplicación del saber por intermedio de la interdisciplina aplicada a la búsqueda de excelencia en atención sanitaria, entendida como servicio al prójimo.

REFERENCIAS

[1] D. Outomuro and L. Mirabile, "Impacto de la tecnología en la práctica de la medicina," Itaes, vol. 15, no. 1, pp. 32-44, 2014.

[2] F. Simini and B. Vienni, "Ingeniería biomédica, interdisciplina y sociedad," Revista Facultad de Ingeniería. Universidad Central de Venezuela, vol. 31, p. 15, 2016.

[3] Presidencia República Oriental del Uruguay, "Sistema Nacional Integrado de Salud." [Online]. Available: *https://www.smu.org.uy/sindicales/documentos/snis/snis.pdf*. [Accessed: 03-May-2018].

[4] Fondo Nacional de Recursos, Uruguay "Qué es el FNR?" [Online]. Available: *http://www.fnr.gub.uy/que_es_fnr*.

[5] Fondo Nacional de Recursos, Uruguay "Técnicas y medicamentos,"2017. [Online]. Available: *http://www.fnr.gub.uy/tecnicas*. [Accessed: 22-Nov-2017].

[6] J. Arango, "Informe de presupuesto 2016," Fondo Nacional de Recursos, Uruguay - Montevideo, Uruguay, 2016.

[7] Á. Haretche, C. Scarpitta, M. Baldizzoni, G. Leiva, R. Gambogi, H. Primus, and L. Gomez, Estándares de evaluación y seguimiento para la mejora de calidad de los IMAE., 2nd ed. Montevideo, Uruguay: Fondo Nacional de Recursos, 2012.

[8] F. Borba and F. Simini, "Electrical Systems in Intensive Care Units of Uruguay," in XXI Congreso Argentino de Bioingeniería X Jornada de Ingeniería Clínica, 2017.

[9] Fondo Nacional de Recursos, Uruguay "Control de las condiciones de seguridad de los equipamientos médicos de los IMAE," El Diario. Med., p. 1, 2004.

[10] Poder Ejecutivo, Decreto N 001-3539/2015 Requisitos de buenas prácticas para el funcionamiento de los servicios de salud. Uruguay: Ministerio de Salud Pública, 2017, p. 9.

[11] O. Gianneo, Á. Haretche, J. Lobo, and F. Simini, "Rutinas de control de calidad de equipos biomédicos de alta complejidad Oscar," XVI Congr. Argentino Bioingeniería, V Jornadas Ing. Clínica Rutinas, pp. 313-316, 2007.

[12] OMS, Introducción a la gestión de inventarios de equipo médico. Organización Mundial de la Salud, 2012.

[13] C. Gonzáles and A. Hernández, Manual de mantenimiento de los servicios de salud: instalaciones y bienes de equipo. Organización Panamericana de la Salud, 1996.

[14] UTE, "Aparatos médicos, aparatos de rayos x," in Reglamento de baja tensión, Montevideo, Uruguay, 2001, p. 4.

[15] UTE, "Instalaciones interiores o receptoras," in Reglamento de baja tensión, Montevideo, Uruguay, 2001, p. 11.

[16] UTE, "Locales de pública concurrencia," in Reglamento de baja tensión, Montevideo, Uruguay, 2001, p. 19.

[17] International Organization for Standardization, "Quality management principles," 2015, p. 20, 2015.

[18] J. E. Camacho-Cogollo, D. M. Torres-Vélez, and T. Chavarría, "Gestión de equipos médicos: implementación y validación de una herramienta de auditoría," Rev. Mex. Ing. Biomed., vol. 38, no. 1, pp. 76-92, 2017.

[19] M. J. Gaitán-González and M. R. Ortiz-Posadas, "Difusión de la ingeniería clínica en eventos académicos.," Rev. Mex. Ing. Biomédica, vol. 28, no. 1, pp. 7-12, 2007.

ARTÍCULO DE INVESTIGACIÓN

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-13

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.4

E-LOCATION ID: **e201821**

ABPSE: Alineador de ADN Basado en Paralelismo a Nivel de Bit y la Estrategia Siembra y Extiende

ABPSE: DNA Aligner Based on Bit-level Parallelism and the Seed and Extend Strategy

D. Pacheco-Bautista, J. Martínez-Oviedo, R. Carreño-Aguilera, I. Algredo-Badillo, S. Sánchez-Sánchez

Universidad del Istmo

RESUMEN

La alineación de ADN es un proceso clave para la reconstrucción de genomas, a partir de los millones de lecturas cortas producidas por las máquinas de secuenciación paralela masiva. Tal proceso suele realizarse mediante algoritmos con elevada complejidad espacial y temporal, requiriendo varias horas para entregar los resultados, así como decenas de GB de RAM. Esto ha motivado la búsqueda de nuevos algoritmos y/o estrategias que permitan disminuir los tiempos de ejecución, mientras se utilizan recursos mínimos de memoria. En este artículo se presenta ABPSE, un nuevo alineador de ADN que combina el algoritmo de Ferragina y Manzini (o índices de FM) y el algoritmo de Myers, mediante la estrategia siembra y extiende. En la siembra, los índices de FM permiten calcular de manera rápida regiones con alta probabilidad de alineación; mientras que en la extensión, el algoritmo de Myers refina la alineación utilizando operaciones basadas en vectores de bits, calculando simultáneamente varias celdas de la matriz de programación dinámica. Los resultados muestran un 96.1% de lecturas alineadas correctamente, un factor de aceleración de 2.45x en relación a BWA-SW y un uso de memoria de apenas 7.6 GB, cuando se alinea el genoma humano completo.

PALABRAS CLAVE: ADN; Bioinformática; Myers; Siembra y extiende; Índices de FM

ABSTRACT

DNA alignment is a key process in the assembly of genomes from the millions of short reads that are produced by massive parallel sequencing machines. Such a process is usually done by means of high spatial and temporal complexity algorithms, which takes hours to deliver the results as well as tens of GB of RAM. This has prompted the search for new algorithms and/or strategies that allow shorter runtimes, while using minimal memory footprint. In this article, we present ABPSE, a new DNA aligner that combines the Ferragina and Manzini algorithm (or FM indexes) and the Myers algorithm, by means of the seed and extend strategy. In the seeding, the FM indices allow a rapid calculation of the regions with high probability of alignment. In the extension, the Myers algorithm refines the alignment using operations based on bit vectors. It simultaneously calculates several cells of the dynamic programming matrix. The results show 96.1% of correctly aligned reads, an acceleration factor of 2.45x in relation to BWA-SW and a memory footprint of only 7.6 GB when aligning the entire human genome.

KEYWORDS: DNA; Bioinformatics; Myers; Seed-and-extend; FM index

Correspondencia

DESTINATARIO: **Daniel Pacheco Bautista**

INSTITUCIÓN: **Universidad del Istmo**

DIRECCIÓN: **Cd. Universitaria S/N, Bo. Santa Cruz Tagojaba, C. P. 70760, Tehuantepec, Oaxaca, México**

CORREO ELECTRÓNICO:

dpachecob@bianni.unistmo.edu.mx

Fecha de recepción:

30 de mayo de 2018

Fecha de aceptación:

29 de noviembre de 2018

INTRODUCCIÓN

La secuenciación de ADN es un proceso que permite obtener el orden de cada uno de los nucleótidos que conforman la molécula de ADN. Tiene una larga lista de aplicaciones y es tecnología clave para la investigación de algunos tipos de cáncer, así como el desarrollo de la medicina genómica, la biología y la agricultura. Las máquinas de secuenciación paralela masiva son tecnologías capaces de secuenciar millones de cadenas de ADN al día [1-3], sin embargo, procesan fragmentos con un número muy pequeño de nucleótidos (entre 35 y 1100), por lo que el resultado de la secuenciación no es un genoma completo, sino pequeñas lecturas cortas que representan fragmentos del mismo. Una manera de reconstruir el genoma a partir de los millones de lecturas cortas es mediante el proceso de alineación [3], el cual consiste en ubicar cada lectura corta tomando como referencia un genoma secuenciado previamente. No obstante, los billones de lecturas cortas, así como la gran longitud del genoma de referencia (3000 millones de nucleótidos para el genoma humano) complican el proceso, además de que deben tomarse en cuenta las diferencias biológicas entre ambas cadenas (inserciones, supresiones o mutaciones de nucleótidos), así como los posibles errores de las máquinas de secuenciación. Para realizar el proceso de alineación se utilizan diferentes algoritmos de elevada complejidad temporal y espacial, algunos basados en programación dinámica [4-5], siendo muy precisos pero requiriendo gran cantidad de recursos computacionales, y otros basados en heurísticas [6-7], los cuales son menos exactos pero más rápidos. Alternativamente los métodos pueden utilizar estimaciones estadísticas [8], basadas principalmente en métodos bayesianos o de máxima verosimilitud o incluso aplicar herramientas de procesamiento digital de señales [9], entre otras técnicas.

La estrategia siembra y extiende

Recientemente se ha optado por combinar las técnicas heurísticas con aquellas basadas en programación dinámica para acelerar la alineación de lecturas, realizando un balance entre velocidad y precisión. La más importante de tales estrategias se denomina siembra y extiende (Figura 1).

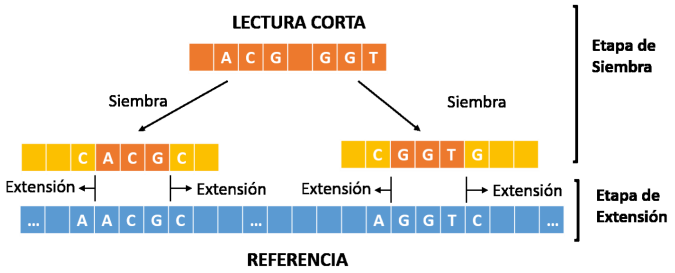


FIGURA 1. La estrategia siembra y extiende.

Durante la etapa de siembra se intenta encontrar una subcadena de la lectura corta (semilla) que se alinee exactamente en uno o más lugares del genoma de referencia. Esta aproximación se basa en la premisa de que si existe un alto grado de similitud de una subcadena de la lectura en una región de la referencia, entonces es más probable que exista un buen alineamiento de toda la lectura corta en esta zona. Para la etapa de extensión se intenta extender la semilla en ambas direcciones, en esta etapa toda la lectura corta es alineada respecto a la cadena de referencia en las zonas encontradas durante la siembra. La extensión permite determinar de forma precisa la existencia de mutaciones, inserciones o supresiones en la lectura respecto a la referencia. Muchos alineadores modernos implementan la técnica siembra y extiende, tal es el caso de BWA-SW [10], BWA-MEM [11], Bowtie2 [12] y Cushaw2 [13]. Para la etapa de siembra, dichos programas realizan un pre-procesamiento del genoma, obteniendo índices de búsqueda eficientes mediante los algoritmos basados en Tablas Hash [14], o en la Transformada de Burrows-Wheeler [15]. Posteriormente, para la extensión implementan algoritmos basados en programación dinámica, tales como el Smith-Waterman [16], o el Needleman-Wunsch [17].

En este artículo se presenta el desarrollo de ABPSE, un alineador eficiente en tiempo y espacio, que realiza la alineación de lecturas cortas mediante la estrategia

siembra y extiende. La siembra utiliza los índices de FM [18] para realizar las búsquedas exactas de semillas, y la extensión el algoritmo de programación dinámica basado en paralelismo a nivel de bit propuesto por Myers [19]. En particular, el algoritmo de Myers está basado en la distancia de Levenshtein para calcular la similitud entre cadenas, tal algoritmo toma las ventajas del paralelismo a nivel de bit del tamaño de palabra del procesador de una computadora, esto es eficiente debido a que los procesadores realizan cálculos con un tamaño entero de palabra en un ciclo de memoria. Básicamente, los cálculos de las puntuaciones y las comparaciones de cadenas son obtenidos simultáneamente mediante una serie de operaciones binarias que comprenden AND, OR, XOR, complementos, desplazamientos y sumas. Después de una profunda revisión de la literatura, se encontró que el algoritmo de Myers no ha sido usado antes en estas aplicaciones.

El resto del artículo se organiza de la siguiente manera, en la siguiente sección se presenta el diseño del software de alineación, iniciando con la descripción del esquema a bloques general propuesto, y culminando con la descripción detallada de los algoritmos y estrategias en cada módulo dentro del mismo. En una sección posterior, se presentan, analizan y comparan los resultados de ejecución obtenidos. Finalmente, se establecen las conclusiones principales.

METODOLOGÍA

El modelo a bloques del programa de alineación desarrollado se muestra en la Figura 2. La implementación del mismo fue realizada utilizando el modelo secuencial de desarrollo de software. Los datos de entrada del alineador son: los archivos con los índices de FM del genoma de referencia, el archivo con las lecturas cortas en formato FASTQ, la cadena de referencia comprimida y un valor de error opcional *n* proporcionado por el usuario, que limita el número de errores permitidos en la alineación. La salida es un archivo con la posición y la trayectoria de alineación en el formato SAM [20].

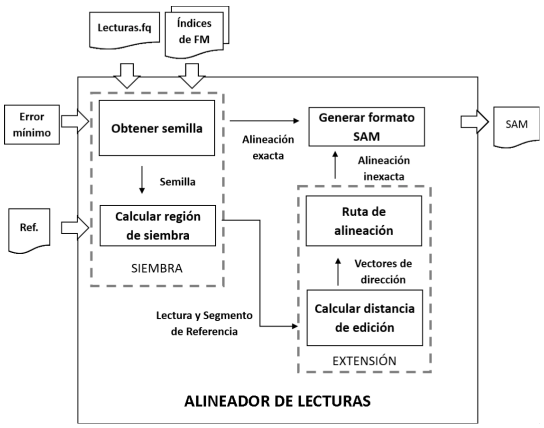


FIGURA 2. Esquema general del alineador genérico.

El proceso inicia en el módulo de siembra, obteniendo las semillas a partir de la lectura corta. Para determinar las semillas, primero se intenta alinear la lectura entera respecto a la cadena de referencia, mediante el algoritmo de búsqueda hacia atrás de Ferragina y Manzini [18]. Si la lectura se alinea de forma exacta entonces se excluye todo el proceso de extensión y se procede de inmediato a generar los resultados de la alineación en formato SAM. En caso contrario, se obtienen las subcadenas de la lectura que se alineen de manera exacta a la referencia, siendo consideradas semillas si tienen una longitud no menor a cierta longitud mínima precalculada. Una vez obtenida una semilla, el segundo paso es calcular la región real donde ésta se ha alineado en la cadena de referencia. El resultado de este paso es una subcadena de la referencia que será usada como entrada en la fase de extensión.

Una vez finalizada la etapa de siembra se inicia la etapa de extensión, la cual tiene como objetivo alinear de forma inexacta la lectura corta y el segmento de la referencia encontrado en la etapa de siembra. El proceso inicia con el cálculo de la distancia de edición entre ambas cadenas utilizando el algoritmo de Myers [19], obteniendo los vectores de bits que representan la matriz de programación dinámica. Posteriormente, se procede a calcular la ruta de alineación utilizando los vectores de bits en un recorrido de la matriz hacia atrás. Finalmente, a partir de los resultados de la ali-

neación se construye el archivo en formato SAM, que representa la salida del alineador y contiene toda la información de la alineación de la lectura corta.

El módulo de siembra

El módulo de siembra obtiene las semillas a partir de cada lectura corta y posteriormente los segmentos de la referencia en las regiones de siembra donde las semillas se han alineado. ABPSE utiliza semillas con máxima coincidencia exacta (SMEM, del inglés *Super Maximal Exact Match*), pues proveen mayor exactitud en la alineación y utilizan menos tiempo de cómputo que otras semillas, como las de longitud fija [21]. En particular, la longitud de una semilla SMEM puede ser tan grande como la lectura corta, de tal manera que se puede determinar inmediatamente cuándo una lectura corta se alinea exactamente a la cadena de referencia sin necesidad de realizar la etapa de extensión.

Para asegurar hallar al menos una semilla que se alinee en forma exacta a la cadena de referencia, la lectura, al inicio, se divide en (n+1) regiones de longitud fija como en la Figura 3, basado en el principio del palomar [22]. También se estableció la longitud mínima de la semilla igual al tamaño de estas regiones, como lo indica la Ecuación 1.

$$\text{L\_mínima\_semilla} = \frac{\text{longitud de lectura corta}}{(n+1)} \qquad (1)$$

La búsqueda de semillas inicia en uno de los extremos de la lectura, intentando encontrar intervalos de alineación exacta mayores a la longitud de semilla mínima. Por ejemplo, si la búsqueda termina dentro de la primera región, la semilla se descarta al no cumplir con el tamaño mínimo (Figura 3a), reanudando una nueva búsqueda en el carácter inmediato a su izquierda. Si la búsqueda, por el contrario, termina en la región dos, se ha encontrado una semilla candidata, y ésta se almacena y se procede a reiniciar una nueva búsqueda a partir del inicio de la región 2 (Figura 3b).

Por otra parte, tomando en cuenta que mientras el valor de n aumenta, el tamaño de la semilla disminuye y en consecuencia se provocan alineaciones falsas en múltiples regiones de la referencia, ABPSE permite valores de n entre 0 y 3, y lecturas cortas con longitud mínima de 60 nucleótidos.

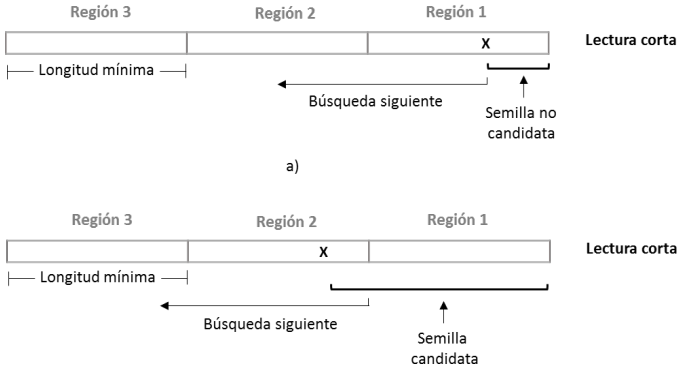


FIGURA 3. Búsqueda de semillas.

Para implementar la etapa de siembra se utiliza el algoritmo de búsqueda hacia atrás basado en los índices de FM. Estos índices son una estructura de datos conformada por la Transformada de Burrows-Wheeler (TBW) de la cadena de referencia, el Arreglo de Sufijos (AS) que contiene todas las posiciones de inicio de los sufijos en la referencia, una Matriz de Ocurrencias (Occ) de los cuatro caracteres del alfabeto Σ = {A, C, G, T} en la TBW y un vector de frecuencia (C). Dichos índices proporcionan una estructura eficiente en espacio para realizar búsquedas exactas de cadenas.

El Algoritmo 1 adaptada de la referencia [18], realiza la búsqueda exacta de cadenas. Las variables k y l representan el intervalo donde aparece el sufijo buscado, P representa la cadena patrón a buscar, i es un apuntador y σ el carácter del patrón que se está procesando actualmente. Las líneas del 5 al 10 representan el núcleo del proceso de búsqueda, básicamente, el ciclo toma cada carácter del patrón una por una y va encontrando los intervalos donde aparece el sufijo que se va formando. Finalmente, el algoritmo retorna el intervalo [k, l] si $k \leq l$ o un intervalo vacío en caso contrario (líneas 11 al 15).

ALGORITMO 1. Búsqueda exacta de cadenas.

```
1   function EXACTMATCH (R, P, C, Occ)
2     i = |P|-1
3     k = 0
4     l = |R|- 1
5     while k ≤ l && i ≥ 0 do
6       σ = P [i]
7       k = C[σ] + Occ[σ,k-1 ] + 1
8       l = C[σ] + Occ[σ,l ]
9       i = i-1
10    end while
11    if k ≤ l then
12      return {k,l}
13    else
14      return {Φ}
15    end if
16  end function
```

Por cada semilla obtenida mediante el algoritmo 1 se almacenan 3 datos importantes: la posición de inicio de la semilla en la lectura corta, el tamaño de la semilla y los valores del intervalo (*k* y *l*). La posición de inicio y el tamaño de la semilla, permiten saber con exactitud a partir de dónde debe extenderse la semilla hacia ambos lados, mientras que el intervalo determina las (*l*-*k*+1) regiones donde la semilla se ha alineado, utilizándose como índices del arreglo de sufijos (AS). En específico, el valor del arreglo de sufijos es utilizado para calcular la posición de inicio y fin del segmento de la referencia que será utilizado en la etapa de extensión, tal como se muestra en la Figura 4.

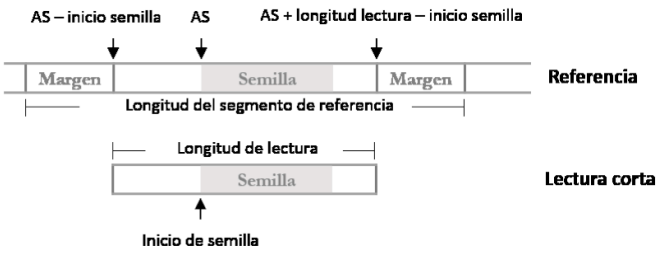


FIGURA 4. Cálculo del segmento de la referencia a utilizar en la etapa de extensión.

Es posible que durante la alineación se realicen operaciones de eliminación de nucleótidos en la lectura corta, lo cual significa que hay inserciones de caracteres en la cadena de referencia y por consiguiente la región requiere un espacio mayor al tamaño de la lectura corta. Es por esta razón que se agregan márgenes en cada extremo de la región previamente encontrada usando un valor de margen igual al número de errores permitidos para la alineación.

El módulo de extensión

La etapa de extensión se implementa utilizando el algoritmo de Myers, el cual se basa en el cálculo de la distancia de edición de Levenshtein, definida como el número mínimo de operaciones de inserción, eliminación o sustitución, necesarias para transformar una cadena a otra. La manera formal de calcular la distancia de edición de Levenshtein es mediante el algoritmo de Programación Dinámica (PD) que utiliza las fórmulas recursivas de la Ecuación (2).

$$C[i,j]=\min\begin{cases} C[i-1,j-1]+\delta_{i,j} \\ C[i-1,j]+1 \\ C[i,j-1]+1 \end{cases} \qquad (2)$$

$$\text{Con:} \quad \delta_{i,j}=\begin{cases} 0, & \text{si } p_i=t_j \\ 1, & \text{en otro caso} \end{cases}$$

Donde C[i,j] contiene la mínima distancia de edición entre el segmento del patrón $P_1,P_2...P_i$ y todos los posibles sufijos del texto T_1, T_2,...Tj. Por ejemplo, dadas dos cadenas similares T= "ATCATGAA" y P= "TCAGT", entonces, la matriz de distancia de edición que representa el cálculo de la similitud entre las dos cadenas es la que se muestra en la Figura 5.

		A	T	C	A	T	G	A	A
	0	0	0	0	0	0	0	0	0
T	1	1	0	1	1	0	1	1	1
C	2	2	1	0	1	1	1	2	2
A	3	2	2	1	0	1	2	1	2
G	4	3	3	2	1	1	1	2	2
T	5	4	3	3	2	1	2	2	3

FIGURA 5. Matriz de distancia de edición entre las cadenas T = "ATCATGAA" y P = "TCAGT".

En tal caso, examinando la última fila de la matriz puede notarse que las subcadenas de T que terminan en la posiciones 4, 6 y 7 están a solo dos transformaciones del patrón P.

Puede notarse que cada celda de la matriz difiere de las celdas vecinas únicamente por 3 valores (1, 0, -1). Mediante esta observación, Myers recodificó la matriz de PD representando solo las diferencias entre las celdas vecinas en cada fila y columna sucesiva, utilizando las fórmulas en las Ecuaciones (3) y (4). Así se obtienen las matrices de deltas mostradas en la Figura 6.

$$\Delta v[i,j]=C[i,j]-C[i-1,j] \quad (3)$$

$$\Delta h[i,j]=C[i,j]-C[i,j-1] \quad (4)$$

		A	T	C	A	T	G	A	A
	0	0	0	0	0	0	0	0	0
T	1	1	0	1	1	0	1	1	1
C	1	1	1	-1	0	1	0	1	1
A	1	0	1	1	-1	0	1	-1	0
G	1	1	1	1	1	0	-1	1	0
T	1	1	0	1	1	0	1	0	1

a)

		A	T	C	A	T	G	A	A
	0	0	0	0	0	0	0	0	0
T	1	0	-1	1	0	-1	1	0	0
C	2	0	-1	-1	1	0	0	1	0
A	3	-1	0	-1	-1	1	1	-1	1
G	4	-1	0	-1	-1	0	0	1	0
T	5	-1	-1	0	-1	-1	1	0	1

b)

FIGURA 6. Representación de la matriz de programación dinámica en deltas. a) Delta vertical Δv. b) Delta horizontal Δh.

Posteriormente, cada columna se almacena mediante dos vectores de bits, VP y VN para la matriz de deltas verticales y, HP y HN para la matriz de deltas horizontales como se muestra en la Figura 7. Cada posición de estos vectores almacenan un 1 cuando se cumplen sus igualdades de acuerdo a las Ecuaciones (5), (6), (7), y (8), en caso contrario almacenan un 0, donde la notación $W_{i,j}$ indica el bit de la i-ésima posición en el entero W de la j-ésima columna. De esta manera, una columna de la matriz original se ha representado en 4 vectores de bits, lo cual reduce el espacio utilizado para calcular la distancia de edición.

$$VP_{i,j}=1 \leftrightarrow \Delta v[i,j]= +1 \quad (5)$$

$$VN_{i,j}=1 \leftrightarrow \Delta v[i,j]= -1 \quad (6)$$

$$HP_{i,j}=1 \leftrightarrow \Delta h[i,j]= +1 \quad (7)$$

$$HN_{i,j}=1 \leftrightarrow \Delta h[i,j]= -1 \quad (8)$$

	A	T	C	A	T	G	A	A
T	1	0	1	1	0	1	1	1
C	1	1	0	0	1	0	1	1
A	0	1	1	0	0	1	0	0
G	1	1	1	1	0	0	1	0
T	1	0	1	1	0	1	0	1

a)

	A	T	C	A	T	G	A	A
T	0	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0
A	0	0	0	1	0	0	1	0
G	0	0	0	0	0	1	0	0
T	0	0	0	0	0	0	0	0

b)

	A	T	C	A	T	G	A	A
T	0	0	1	0	0	1	0	0
C	0	0	0	1	0	0	1	0
A	0	0	0	0	1	1	0	1
G	0	0	0	0	0	0	1	0
T	0	0	0	0	0	1	0	1

c)

	A	T	C	A	T	G	A	A
T	0	1	0	0	1	0	0	0
C	0	1	1	0	0	0	0	0
A	1	0	1	1	0	0	1	0
G	1	0	1	1	0	0	0	0
T	1	1	0	1	1	0	0	0

d)

FIGURA 7. Representación en vectores de bits de la matriz de programación dinámica. a) VP, b) VN, c) HP y d) HN.

Myers demostró que estos vectores pueden obtenerse recursivamente mediante las Ecuaciones (9-14), donde Xv y Xh son vectores auxiliares, y Eq es un vector que codifica la igualdad de caracteres. El análisis se basa en que una columna de la matriz de edición solo puede calcularse a partir de los valores de la columna previa.

$$Xv = Eq \text{ OR } VN_{in} \quad (9)$$

$$VP_{out} = HN_{in} \text{ OR NOT } (Xv \text{ OR } HP_{in}) \quad (10)$$

$$VN_{out} = HP_{in} \text{ AND } Xv \quad (11)$$

$$Xh = Eq \text{ OR } HN_{in} \quad (12)$$

$$HP_{out} = VN_{in} \text{ OR NOT } (Xv \text{ OR } VP_{in}) \quad (13)$$

$$HN_{out} = VP_{in} \text{ AND } Xh \quad (14)$$

El Algoritmo 2 representa el algoritmo de Myers al estilo de programación en C adaptado de [19]. Donde la variable score contiene el último valor de cada columna de la matriz de distancia de edición original, calculada a partir del previo valor de score y el vector HP.

ALGORITMO 2. El algoritmo de Myers.

```
1   Precomputo de Peq[c]
2   VP = 1^m
3   VN = 0^m
4   Score = m
5   for j=1,2,…,n
6       Eq = Peq[tj]
7       Xv = Eq | VN
8       Xh = (((Eq & VP) + VP) ^ VP) | Eq
9       HP = VN | ~(Xh | VP)
10      HN = VP & Xh
11      if HP & 10m-1 then
12         Score += 1
13      else if HN & 10m-1 then
14         Score -=1
15      end if
16      HP <<= 1
17      HN <<= 1
18      VP = HN | ~(Xv | HP)
19      VN =   HP & Xv
20      if Score ≤ k then
21         printf "Coincidencia en " . j
22      end if
23  end for
```

Cálculo de la trayectoria de alineación

El algoritmo previo obtiene únicamente la distancia de edición entre las cadenas, sin embargo, en esta aplicación se requiere de la trayectoria de alineación completa, por lo que fueron necesarias algunas modificaciones. En la Figura 8 se muestran todas las rutas posibles de alineación entre las cadenas P y T, donde las flechas de dirección indican la celda de la cual puede preceder la celda C[i,j] durante el cálculo de la matriz.

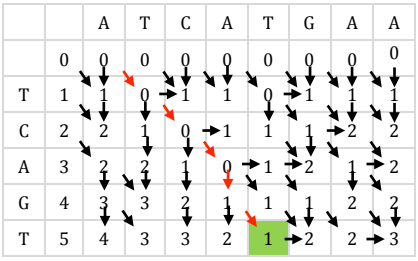


FIGURA 8. Ruta de alineación a partir de la distancia de edición mínima.

En el ejemplo, la mejor distancia de edición es 1, entonces, a partir de la casilla con este valor puede recorrerse la matriz hacia atrás y determinar la ruta de alineación; el proceso se realiza de derecha a izquierda y se apoya en flechas verticales, horizontales y diagonales. Finaliza cuando llega a una casilla de la primera fila de la matriz o se hayan recorrido todas las columnas. Una flecha en diagonal significa una coincidencia o sustitución de caracteres entre las dos cadenas; una vertical, una inserción respecto al patrón; y una horizontal la eliminación de un carácter en el patrón.

Como puede verse, las flechas horizontales y verticales coinciden perfectamente con los unos dentro de los vectores HP y VP de la Figura 7, por lo que pueden utilizarse en el recorrido hacia atrás. Sin embargo, no hay información sobre las diagonales, solo la del vector Eq, lo cual no es suficiente si los caracteres comparados no coinciden entre ellos. Para remediarlo, fue extendida la tabla presentada en el artículo original del algoritmo [19] con las posibles combinaciones de entrada, tal como se muestra en la Tabla 1, donde para clarificar el concepto se han incluido a la derecha ejemplos de combinaciones de entrada que justifican los valores de la salida D. A partir de ahí, pueden derivarse de forma inmediata el vector de la Ecuación 15, el cual representa los movimientos diagonales.

$$D = Eq \text{ or not } (VN_{in} \text{ or } HN_{in}) \quad (15)$$

El vector D, para el caso que se ha tratado como ejemplo en este artículo, se muestra en la Figura 9.

TABLA 1. Tabla de verdad para obtener el vector D.

No.	Δh_{in}	Δv_{in}	D	Δh_{out}	Δv_{out}	Vista de las 4 casillas								
1	-1	-1	1	1	1	1	0		1	1		1	2	1
2	0	-1	1	1	0	0	1		0	1		0	1	0
3	1	-1	1	1	-1									
4	-1	0	1	0	1	1	0		0	0		0	1	1
5	0	0	1	0	0	1	1		0	0		0	0	1
6	1	0	1	0	-1									
7	-1	1	1	-1	1	1	0		0	0		0	1	1
8	0	1	1	-1	0	2	1		1	0		1	0	2
9	1	1	1	-1	-1									
10	-1	-1	0	1	1	1	0		1	1		1	2	1
11	0	-1	0	1	0	0	1		0	1		0	1	0
12	1	-1	0	1	-1									
13	-1	0	0	0	1	1	0		0	0		0	1	1
14	0	0	1	1	1	1	1		0	1		0	1	1
15	1	0	1	1	0									
16	-1	1	0	-1	1	1	0		0	0		0	1	1
17	0	1	1	0	1	2	1		1	1		1	1	2
18	1	1	1	0	0									

	A	T	C	A	T	G	A	A
T	1	1	1	1	1	1	1	1
C	1	0	1	0	0	1	1	1
A	1	0	0	1	0	1	1	1
G	0	1	0	0	1	1	0	1
T	0	1	0	0	1	1	1	1

FIGURA 9. Direcciones en diagonal usando el vector D.

A partir de los vectores VP, HP y D, pueden obtenerse con facilidad todas las rutas de alineación. Por ejemplo, el Algoritmo 3 realiza el recorrido hacia atrás para obtener una posible ruta de alineación. Las entradas del algoritmo son los vectores D y VP que representan las direcciones en diagonal y vertical, *m* es la longitud de la lectura corta, *dmin* es la distancia mínima de edición y *col* la columna donde se encontró dicha distancia. La definición de las variables se realiza en las líneas 2 a 4, donde la variable *CIGAR* almacena todas las operaciones de la alineación, la variable *bit_actual* es una máscara auxiliar que permite identificar en que bit se encuentra el recorrido dentro de cada vector de bits e *i* es un apuntador de índice de la variable *CIGAR*.

El núcleo del programa se encuentra entre las líneas 5 y 17, donde se realiza el recorrido de la matriz hacia atrás, revisando los vectores D y VP hasta que ya no pueda desplazarse más el bit de la máscara auxiliar o cuando el contador de columna *col* sea igual a cero. La ruta de alineación o *CIGAR*, almacena la serie de operaciones que transforman la lectura al texto de referen-

ALGORITMO 3. Recorrido hacia atrás para obtener una ruta de alineación.

```
1   function RUTA_DE_ALINEACION (D,VP,m,
    dmin, col)
2     CIGAR [m + dmin + 1]
3     bit_actual = 1 << (m – 1)
4     i = 0
5     while (bit_actual && col >=0) do
6       if (D[col] & bit_actual) then
7           bit_actual >>= 1
8           col --
9           CIGAR[i++] = 'M'
10      else if (VP[col] & bit_actual) then
11          bit_actual >>= 1;
12          CIGAR[i++] = 'I'
13      Else
14         col--;
15         CIGAR[i++] = 'D'
16      end if
17    end while
18
19  while (bit_actual) do
20      CIGAR[i++] = 'I'
21      bit_actual >>= 1
22    end while
23
24    revertir(CIGAR,i)
25
26    return CIGAR
```

cia, una coincidencia o sustitución de nucleótidos se representa con la letra M, una inserción con la letra I y una eliminación con la letra D. En el proceso, es probable que la lectura corta deba alinearse más a la izquierda del segmento de referencia, lo cual sucede cuando la columna llega a cero antes que la máscara auxiliar, por lo que es necesario agregar operaciones de inserción en el *CIGAR*, lo anterior se realiza en las líneas 19 a 22. Finalmente, en la línea 24 se invierte la cadena resultante antes de ser retornada por la función. El *CIGAR* para el ejemplo que se ha tratado en este artículo se muestra en la Figura 10, donde puede verse claramente la alineación y las operaciones realizadas. Luego, el *CIGAR* se compacta según las especificaciones del formato SAM, obteniendo una versión corta del resultado; en este caso es "3M1I1M". Evidentemente, ambas operaciones pueden ser realizadas por la misma función, lo cual ocurre en el código real del alineador.

Texto	A	T	C	A		T	G	A	A
Cigar		M	M	M	I	M			
Patrón		T	C	A	G	T			

FIGURA 10. El *CIGAR* de la alineación.

Extendiendo el algoritmo para lecturas con longitud mayor a *w*

La descripción anterior es válida cuando el tamaño de la lectura *m* es mayor al tamaño de la palabra *w* del procesador, la cual está limitada a 64 en los procesadores modernos. Para eliminar dicha limitante fue necesario el procesamiento a bloques como se muestra en la Figura 11.

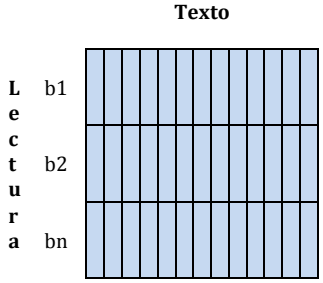


FIGURA 11. Modelo basado en bloques de vectores.

Por cada vector de bits en el algoritmo original, ahora se requieren $b = m/w$ vectores. El cálculo se realiza columna a columna, comenzando con el vector de bits del primer bloque y continuando en forma descendente, tomando en cuenta que, a diferencia de los vectores de bits en los bloques superiores de la matriz, las cuales contienen valores de 0 en el bit menos significativo, el resto de los bloques pueden contener valores de 0 o 1 debido al desplazamiento de bits de los bloques que los anteceden en la columna. Además, los vectores de bits en los bloques del último nivel pueden extenderse más allá del bit que corresponde a la longitud m de la cadena patrón, es decir bitmax = $w - m \pmod w$. Por lo que para calcular el valor de la distancia de edición es necesario un seguimiento preciso de dicho bit, lo cual se realiza mediante una máscara binaria apropiada.

RESULTADOS Y DISCUSIÓN

Para probar la funcionalidad y eficiencia del programa desarrollado se alinearon varios conjuntos de prueba, con un millón de lecturas cortas cada uno, y longitudes de 64, 100 y 128 nucleótidos, a los cromosomas 19, 20, 21 y 22 del genoma humano. Las lecturas cortas fueron generadas de forma artificial utilizando el programa wgsim con una razón de mutaciones de 0.4%, la cual representa el límite de variaciones en el caso del genoma humano [23], donde el 25% de esas mutaciones son indels (inserciones y eliminaciones) y el 70% de los indels son extendidos. El error de secuenciación fue configurado a 0.1%, valor típico en las máquinas de secuenciación actuales [1]. Todas las pruebas fueron realizadas en una computadora con procesador Intel Core i7 de Sexta generación, 16 GB en RAM y 1 TB en disco duro, con sistema operativo Ubuntu 14.04.

En la Tabla 2 se muestran los tiempos de alineación obtenidos, donde puede observarse que incluso los conjuntos de lecturas cortas de 128 nucleótidos, requieren menos de 150 segundos de procesamiento. Lo cual se debe principalmente al paralelismo a nivel

de bit del algoritmo utilizado en la etapa de extensión. También puede observarse una dependencia directa del tiempo de alineación con respecto a la longitud de la lectura, y que los tiempos de alineación permanecen prácticamente constantes aun cuando se incrementa la longitud de las cadenas de referencia.

TABLA 2. Tiempos de ejecución de ABPSE.

Referencia	Longitud de lectura (nts)		
	64	100	128
Chr21	97,61 s.	111,16 s.	140,35 s.
Chr22	106,11 s.	113,87 s.	137,21 s.
Chr19	101,20 s.	131,41 s.	145,19 s.
Chr20	107,33 s.	120,88 s.	143,42 s.

Esto último es muy apropiado, ya que permite la alineación de lecturas a cadenas de referencia muy grandes sin tiempos extras de cómputo, y se debe a la complejidad temporal del algoritmo usado en la etapa de siembra (búsqueda basada en índices de FM), la cual es una función de la longitud de las lecturas cortas y no de la longitud de la cadena de referencia.

TABLA 3. Exactitud de alineación con ABPSE.

Referencia	Tipo de resultado	Longitud de lectura (nts)		
		64	100	128
Chr21	Alineadas	986558	976349	967040
	No alineadas	13442	23651	32960
	Correctas	976101	970002	961044
	Incorrectas	10457	6347	5996
Chr22	Alineadas	985954	976409	967229
	No alineadas	14046	23591	32771
	Correctas	956513	955025	947981
	Incorrectas	29441	21384	19248
Chr19	Alineadas	985117	975625	967397
	No alineadas	14883	24375	32603
	Correctas	962164	963039	956562
	Incorrectas	22953	12586	10835
Chr20	Alineadas	986079	976309	967061
	No alineadas	13921	23691	32939
	Correctas	972866	968010	959421
	Incorrectas	13213	8299	7640

Adicionalmente, se determinó la cantidad de lecturas que se alinearon correcta e incorrectamente a su origen. Esto fue posible debido a que el generador de lecturas artificiales, wgsim, proporciona la ubicación exacta de donde es extraída cada una de sus lecturas. De esta manera, mediante una función de comparación se validan los resultados de ABPSE. En esta prueba, se consideran correctas aquellas alineaciones en un intervalo de ±3 nucleótidos de su posición original, considerando la posibilidad de inserciones o eliminaciones en relación al genoma de referencia. Los resultados se muestran en la Tabla 3, donde puede notarse un porcentaje promedio de 96.2% de alineaciones correctas, 1.39% de alineaciones incorrectas y 2.41% de lecturas no alineadas. Las alineaciones incorrectas son provocadas principalmente por regiones repetidas, lo cual puede mejorarse posteriormente al utilizar lecturas apareadas.

Finalmente, se compararon los tiempos de ejecución del alineador propuesto y el de los populares programas de alineación BWA-SW y BWA-MEM. Ambos programas utilizan la estrategia siembra y extiende, y los índices de FM para la etapa de siembra. En la extensión BWA-SW utiliza el algoritmo de Smith-Waterman mientras que BWA-MEM utiliza una combinación de alineación local y global. La comparación fue realizada utilizando un solo núcleo del procesador en cada uno de los programas. La Figura 12 muestra los resultados al alinear conjuntos de 1 millón de lecturas al genoma humano completo.

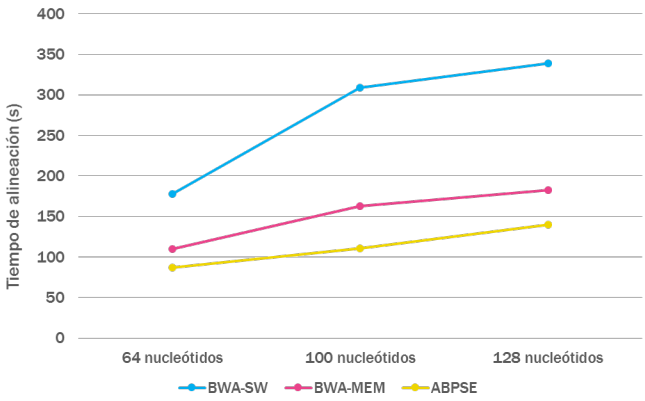


FIGURA 12. Comparación de los tiempos de alineación de ABPSE con los alineadores BWA-SW y BWA-MEM.

Puede observarse como el alineador propuesto proporciona un factor de aceleración superior a 2.45x respecto al programa BWA-SW y 1.36x respecto al alineador BWA-MEM, ambos programas han sido catalogados entre los más rápidos y exactos alineadores de ADN en la actualidad [24-25]. Estas mejoras en la velocidad son muy significantes, puesto que en un proceso normal de alineación, los billones de lecturas que deben alinearse implican días completos de procesamiento, lo cual con ABPSE podría realizarse en solo unas horas. Factores de alineación superiores solo se han reportado mediante aceleradores Hardware basados en unidades de procesamiento gráfico GPUs [26], o en FPGAs [27], aunque estas son excelentes opciones, su elevado costo suele ser una limitante. En esta prueba ABPSE utilizó en promedio solo 7.6 GB de memoria RAM, lo cual permite su ejecución en cualquier computadora convencional.

El número promedio de lecturas mapeadas correcta e incorrectamente, también fue contrastado en esta prueba, obteniendo los resultados de la Figura 13.

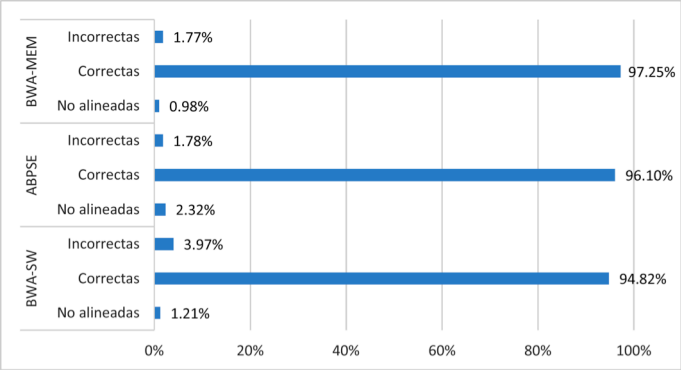


FIGURA 13. Comparación de la eficiencia de ABPSEcon los alineadores BWA-SW y BWA-MEM.

Puede observarse una eficiencia comparable entre los tres programas. En particular, al calcular el error de alineación definida mediante la Ecuación 16, se obtienen valores de 0.0401, 0.0181 y 0.0178 para BWA-SW, ABPSE Y BWA-MEN respectivamente.

$$Error = \frac{No.\ de\ Lecturas\ alineadas\ incorrectamente}{No.\ de\ Lecturas\ alineadas} \qquad (16)$$

Lo anterior refleja una eficiencia muy cercana entre BWA-MEM y ABPSE, siendo mayor a la de BWA-SW. La ligera desventaja en relación a BWA-MEN puede superarse en versiones posteriores al refinar la técnica de sembrado.

CONCLUSIONES

En este artículo se presentó el desarrollo de un programa de alineación de lecturas cortas de ADN que implementa la estrategia siembra y extiende, utilizando la combinación de dos algoritmos muy eficientes. En la etapa de siembra se utilizó el algoritmo de búsqueda hacia atrás basado en los índices de FM que permite realizar búsquedas exactas de cadenas de forma muy rápida independientemente de la longitud de la cadena de referencia. Para la etapa de extensión se implementó el algoritmo de programación dinámica de Myers que calcula la distancia de edición entre dos cadenas. Este último algoritmo fue extendido para obtener la trayectoria de alineación de las cadenas utilizando vectores de bits. Mediante las pruebas realizadas se demuestra que la combinación de dichos algoritmos permite el desarrollo de alineadores de alta velocidad, gracias al paralelismo a nivel de bit en la etapa de extensión. A pesar de que la exactitud del alineador es muy bueno, ésta podría mejorarse si se exploran otros tipos de semillas en la etapa de siembra. Por otra parte, la etapa de extensión podría ser implementada directamente en hardware, donde el tamaño del entero no es un factor limitante, evitando de esta manera la programación a bloques y en consecuencia permitiendo la aceleración al máximo del programa, así como la alineación de lecturas mucho más largas.

REFERENCIAS

[1] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat. rev. genet. 2016; 17(6): 333-351. DOI: 10.1038/nrg.2016.49

[2] Yohe S, Thyagarajan B. Review of Clinical Next-Generation Sequencing. Arch. Pathol. Lab. Med. 2017; 141(11): 1544-1557. DOI: 10.5858/arpa.2016-0501-RA

[3] Pacheco-Bautista D, González-Perez M, Algredo-Badillo I. De la secuenciación a la aceleración hardware de los programas de alineación de ADN, una revisión integral. Rev. mex. ing. bioméd. 2015; 36(3):259-277. DOI: 10.17488/RMIB.36.3.6

[4] Liu Y, Tran TT, Lauenroth F, Schmidt B. Smith-Waterman algorithm on Xeon Phi coprocessors for long DNA sequences. In Cluster Computing (CLUSTER), 2014 IEEE International Conference on; pp. 257-265. IEEE; 2014. DOI: 10.1109/CLUSTER.2014.6968772

[5] Salavert J, Tomás A, Tárraga J, Medina I, Dopazo J, Blanquer I. Fast inexact mapping using advanced tree exploration on backward search methods. BMC Bioinformatics. 2015; 16 (1): 18. DOI: 10.1186/s12859-014-0438-3

[6] Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009; 10(3). DOI: 10.1186/gb-2009-10-3-r25

[7] Drucker TM, Johnson SH, Murphy SJ, Cradic KW, Therneau TM, Vasmatzis G. BIMA V3: an aligner customized for mate pair library sequencing. Bioinformatics. 2014; 30(11): 1627-1629. DOI: 10.1093/bioinformatics/btu078

[8] Agrawal A, Huang X. Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2011; 8(1): 194-205. DOI: 10.1109/TCBB.2009.69

[9] Borrayo E, Mendizabal-Ruiz EG, Vélez-Pérez H, Romo-Vázquez R, Mendizabal AP, Morales JA. Genomic signal processing methods for computation of alignment-free distances from DNA sequences. PloS one, 2014; 9(11): e110954. DOI: 10.1371/journal.pone.0110954

[10] Li H DR. Fast and accurate long-read alignment with Burrows-Wheeler trasform. Bioinformatics. 2010; 26(5): 589-595. DOI: 10.1093/bioinformatics/btp698

[11] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. preprint. 2013;: p. arXiv:1303.3997. [Online]. Available: *hup://arxiv.orglabs/1303.3997*

[12] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9(4): 357-359. DOI: 10.1038/nmeth.1923

[13] Liu Y, Schmidt B, Maskell DL. Cushaw: a cuda compatible short read aligner to large genomes based on the burrows-wheeler transform. Bioinformatics. 2012; 28(14): 1830-1835. DOI: 10.1093/bioinformatics/bts276

[14] Wu TD. Bitpacking techniques for indexing genomes: I. Hash tables. 2016; 11(5): p. 1748-7188. DOI: 10.1186/s13015-016-0069-5

[15] Burrows M, Wheeler DJ. A block sorting lossless data compression algorithm. Reporte Técnico. Palo Alto, California: Digital Equipment Corporation, Systems Research Center; 1994. Report Number: 124.

[16] Smith TF, Waterman MS. Identification of common molecular subsequences. J. Mol. Biol. 1981; 14(1): 195-197. DOI: 10.1016/0022-2836(81)90087-5

[17] Needleman N, Wunsch C. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins. Journal of Molecular. 1970; 48(3): 443-453. DOI: 10.1016/0022-2836(70)90057-4

[18] Ferragina P, Manzini G. Opportunistic data structures with applications. In Foundations of computer science; 2000; Redondo Beach, CA: IEEE. 390-398. DOI: 10.1109/SFCS.2000.892127

[19] Myers G. A fast bit-vector algorithm for approximate string matching based on dynamic programming. Journal of the ACM (JACM). 1999; 46(3): 395-415. DOI: 10.1145/316542.316550

[20] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16): 2078-2079. DOI: 10.1093/bioinformatics/btp352

[21] Ahmed N, Bertels K, Al-Ars Z. A comparison of seed-and-extend techniques in modern DNA read alignment algorithms. In Bioinformatics and Biomedicine (BIBM) 2016 IEEE International Conference on. 2016; p. 1421-1428. DOI: 10.1109/BIBM.2016.7822731

[22] Ahmadi A, Behm A, Honnalli N, Li C, Weng L, Xie XH. Optimized gram-based methods for efficient read alignment. Nucleic Acids Res. 2012; 40(6). DOI: 10.1093/nar/gkr1246

[23] Tishkoff SA, Kidd KK. Implications of biogeography of human populations for 'race' and medicine. Nature Genetics. 2004; 36(11): S21-S22. DOI: 10.1038/ng1438

[24] Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012; 28(24): 3169-3177. DOI: 10.1093/bioinformatics/bts605

[25] Policriti A, Prezza N. Fast randomized approximate string matching with succinct hash data structures. BMC bioinformatics. 2015; 16(9): S4. DOI: 10.1186/1471-2105-16-S9-S4

[26] Hung CL, Lin YS, Lin CY, Chung YC, Chung YF.. CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs. Computational biology and chemistry. 2015; 58: 62-68. DOI: 10.1016/j.compbiolchem.2015.05.004

[27] Pacheco-Bautista D., Carreño-Aguilera, R., Cortés-Pérez E, Gónzalez-Pérez, M, Medel JJ., Acevedo MA, YU W. Nonlinear FM index application for alignment of short DNA sequences using re-parametrization of algorithms. Fractals. 2018; 26(3): 1850023. DOI: 10.1142/S0218348X18500238

ARTÍCULO DE INVESTIGACIÓN

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-10

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.5

E-LOCATION ID: **e201838**

Preparación de un Adhesivo Sensible a la Presión (PSA) con la Incorporación de Nanopartículas de ZnO. Estudio de sus Propiedades Fisicoquímicas y Antimicrobianas

Preparation of a Pressure Sensitive Adhesive (PSA) with the ZnO Nanoparticles Incorporation. Study of its Physicochemical and Antimicrobial Properties

S. N. Ramírez-Barrón[1], *S. Sánchez-Valdés*[1], *B. A. Puente-Urbina*[1], *S. Martínez-Montemayor*[1], *S. C. Esparza-González*[2], *R. Betancourt-Galindo*[1]

[1]Centro de Investigación de Química Aplicada, CIQA

[2]Facultad de Medicina de Saltillo, UAdeC

RESUMEN

Se describe el proceso para obtener un adhesivo sensible a la presión (PSA). Este PSA está formado por un copolímero de acrilato de 2-etilhexil (2-EHA) / metacrilato de metilo (MMA) en una relación 80:20 que se polimerizó mediante una técnica de polimerización en emulsión. Se añadieron nanopartículas de óxido de zinc (NPZnO) a este copolímero, que se sintetizaron previamente y se modificaron superficialmente con 3-aminopropil-3-toxisilano (APTES) y dimetilsulfóxido (DMSO) para mejorar su dispersión en la matriz de copolímero. Los nanocompuestos obtenidos se caracterizaron por espectroscopía infrarroja (FTIR), calorimetría diferencial de barrido (DSC) y pruebas de adhesión al delaminado. Además, se determinó la actividad antimicrobiana contra *S. aureus* y *S. pyogenes*, así como la citotoxicidad en células humanas (HeLa). Los resultados demostraron que la adición de las nanopartículas de NPZnO al copolímero incrementa la temperatura de transición vítrea (Tg) así como las propiedades antimicrobianas del adhesivo mejorando a su vez su adhesión superficial. Con respecto al comportamiento adhesivo, el PSA con NPZnO sin modificar mostró una mayor resistencia al delaminado, esto quiere decir que las nanopartículas incrementan la fuerza cohesiva y proporcionan resistencia a temperaturas elevadas, lo cual sería beneficioso a su aplicación final. Finalmente, los resultados de citotoxicidad mostraron que la incorporación de NPZnO al PSA disminuye la viabilidad celular, sin embargo no se considera tóxico acorde a la norma ISO 10993 test for *in vitro* cytotoxicity.

PALABRAS CLAVE: Nanopartículas de ZnO; modificación de la superficie; polimerización en emulsión; propiedades antimicrobianas; citotoxicidad

ABSTRACT

The process for obtaining a pressure sensitive adhesive (PSA) is described. This PSA is formed by an acrylate copolymer of 2-ethylhexyl (2-EHA) / methyl methacrylate (MMA) in an 80:20 ratio which was polymerized by emulsion polymerization technique. Zinc oxide nanoparticles (NPZnO) were added to this copolymer, which were previously synthesized, and surface modified with 3-aminopropyltretoxysilane (APTES) and dimethyl sulfoxide (DMSO) to improve its dispersion in the copolymer matrix. The obtained nanocomposites were characterized by infrared spectroscopy (FTIR), differential scanning calorimetry (DSC) and T-peel adhesion tests. In addition, the antimicrobial activity against *S. aureus* and *S. pyogenes* as well as the cytotoxicity in human cells (HeLa) were determined. The results demonstrated that the ZnO nanoparticles incorporation enhanced the glass transition temperature (Tg) and the antimicrobial activity of PSA copolymer as well as its surface adhesion. It was confirmed that NPZnO modification with APTES increased its antimicrobial activity. Regarding adhesive behavior, PSA with unmodified NPZnO showed a greater peel resistance. This indicates that these nanoparticles enhances the cohesive force and induces a better high temperature performance, which is beneficial for the final application. Finally, cytotoxicity results showed that the incorporation of NPZnO to PSA decreases the cell viability, however this PSA is not toxic according to the standard ISO 10993 test for *in vitro* cytotoxicity.

KEYWORDS: ZnO nanoparticles; surface modification; emulsion polymerization; antimicrobial properties; cytotoxicity

Correspondencia

DESTINATARIO: **Rebeca Betancourt Galindo**

INSTITUCIÓN: **Centro de Investigación de Química Aplicada, CIOA**

DIRECCIÓN: **Blvd. Enrique Reyna Hermosillo #140, C.P. 25294, Saltillo, Coahuila, México**

CORREO ELECTRÓNICO: **rebeca.betancourt@ciqa.edu.mx**

Fecha de recepción:

18 de septiembre de 2018

Fecha de aceptación:

10 de enero de 2018

INTRODUCCIÓN

Los adhesivos sensibles a la presión (PSA) son materiales viscoelásticos que, en su estado seco y a temperatura ambiente, se vuelven pegajosos cuando se aplica una pequeña presión. Debido a estas propiedades, los PSA tienen una gran cantidad de aplicaciones, desde la industria de la construcción hasta la medicina. Los PSA se han sintetizado mediante polimerización en emulsión y se han incorporado una amplia variedad de rellenos nanométricos para modificar sus propiedades o incluso añadir nuevas características. Las nanopartículas de ZnO (NPZnO) son materiales inorgánicos que se pueden incorporar en una matriz polimérica. Las propiedades intrínsecas del PSA dependen de la fase del polímero y de las propiedades antimicrobianas y citotóxicas del material inorgánico de relleno. ZnO tiene una amplia gama de aplicaciones en ingeniería, medicina y aeronáutica, entre otros. Dicha aplicación incluye dispositivos automotrices, sistemas de comunicación, sistemas biológicos tales como biomateriales para ingeniería de tejidos, polímeros de memoria de forma como interruptores moleculares, biosensores, diagnóstico de laboratorio y liberación de fármacos [1]. Existen diferentes factores que afectan la funcionalidad del PSA, entre los que se encuentran la composición, la morfología de las partículas, la resistencia adhesiva, la flexibilidad, entre otros [2]. Se han publicado varios trabajos relacionados con la síntesis de PSA con diferentes partículas nanoestructuradas, para mejorar las propiedades mecánicas y de liberación del fármaco, o aumentar la conductividad eléctrica. Por lo tanto, Jingyeong O. (2009) [3] prepararon composiciones de nanoarcillas de montmorillonita de poliacrilato (MMT) en diversos contenidos de arcilla por mezcla mecánica y polimerización en emulsión in situ. Descubrieron que la temperatura de transición vítrea (Tg) del polímero y las propiedades viscoelásticas dependen de la concentración final de MMT. Las propiedades mecánicas de PSA usando EA/EHA llena de NaMMT fueron estudiadas por Lofton L. (2004) [4]. Descubrieron que las propiedades adhesivas se ven fuertemente afectadas por el tipo y la cantidad de nanoarcillas incorporadas. Por ejemplo, la fuerza de delaminado y adherencia disminuyó gradualmente con una mayor cantidad de nanoarcilla de MMT. Y se observó un aumento sustancial en el esfuerzo cortante. El contenido máximo determinado de arcilla modificada fue 1% en peso. Mientras que se observó una influencia moderada sobre la adherencia, la resistencia al despegue y el esfuerzo cortante cuando se incorporaron nanoarcillas no modificadas [5, 6].

Otra investigación presenta la incorporación de silicato de sodio a nanoescala a Poli (2-EHA-co-AA) con el fin de mejorar la absorción de sodio-cloxacilina. También Wang *et al.* (2006) [7], desarrollaron un adhesivo de polibutilacrilato y SWNT (nanotubos de carbono de pared simple) para aumentar la conductividad eléctrica. Los SWCNT se funcionalizaron con poli (alcohol vinílico) (PVA), en una proporción de 0,3% en peso. Los resultados demostraron que la conductividad aumenta en diez órdenes de magnitud. También en este umbral, la adhesión cambia y se observó un aumento en la pegajosidad. De la misma manera, la energía de adhesión aumenta aproximadamente un 85% con la adición de PVA-SWNT a la misma concentración. Aunque hay varios informes sobre la preparación de nanocompuestos adhesivos de PSA acrílico, no se han encontrado informes que impliquen la incorporación de NPZnO, modificadas o no modificadas en la superficie, en la matriz de polímero adhesivo PSA.

Por lo tanto, en este trabajo se discute la síntesis de NP esféricas de ZnO y su incorporación a PSA. Las NPZnO se modificaron en superficie con dos agentes diferentes: 3-aminopropiltrietoxisilano (APTES) y dimetilsulfóxido (DMSO) a contenidos de 0.1, 0.2 y 0.3% en peso. Las nanopartículas obtenidas se caracterizaron por difracción de rayos X (XRD) y microscopía electrónica de transmisión (TEM). El compuesto PSA/NP obtenido se estudió mediante microscopía electrónica de transmisión (TEM), resistencia al delaminado y pruebas antimicrobianas contra dos microorganismos diferen-

tes: *Staphylococcus aureus* (*S. aureus*) y *Streptococcus pyogenes* (*S. pyogenes*), y finalmente, se realizaron pruebas de citotoxicidad en los adhesivos obtenidos.

METODOLOGÍA

Materiales

Acrilato de 2 etilhexilo (2-EHA), Metilmetacrilato (MMA), ácido acrílico (AA), persulfato de amonio (APS), Latemul-180, dodecanotiol, hidroquinona (HQ) acetato de zinc ($Zn(Ac)_2 \cdot 2H_2O$), hidróxido de sodio (NaOH), dimetilsulfoxido (DMSO) y aminopropiltrimetoxisilano (APTES) todos con 98% de pureza y de Sigma Aldrich. Etanol fue obtenido de J.T. Baker y agua desionizada de un sistema de columnas de intercambio iónico (Cole-Parmer Instruments).

Preparación de las nanopartículas de ZnO (NPZnO)

NPZnO fueron sintetizadas preparando una solución etanólica de 0.06 M de $Zn(Ac)_2 \cdot 2H_2O$ que fueron colocadas en un reactor. Enseguida la mezcla fue puesta en reflujo a 80 °C por 2 h a 100 rpm. Después de agregar una solución aucosa que contenía 0.22M de NaOH, la mezcla resultante se puso en marcha durante 12 h a temperatura ambiente y el precipitado se hizo pasar por centrifugado tres veces con el 500 mL de etanol y se secó en una estufa a 80 °C durante 24 h.

Modificación superficial

La modificación superficial de las NPZnO fue llevada a cabo colocando 10 g de nanopartículas en un matraz bola de 3 bocas provisto de un condensador al que se adicionaron 100 mL de octano como solvente. Los diferentes agentes modificantes (dimetil sulfóxido y 3 aminopropiltrietoxisilano) se añadieron en una relación en masa de 2:1. La reacción se llevó a cabo mediante reflujo durante 3 h a 80 °C y agitación constante. Después el sistema fue enfriado, el material se pulverizó y secó durante 12 h a 80 °C para su posterior caracterización.

Síntesis de adhesivo sensible a la presión (PSA)

Diferentes concentraciones de nanopartículas (0.1, 0.2 y 0.3%), sin y con modificación superficial con los agentes APTES y DMSO, fueron dispersadas en la mezcla de monómeros 2-EHA/MMA en una relación en masa de 4:1; es decir, se añadieron 16 mL de 2-EHA y 2 mL de MA del agente de transferencia de cadena AA, así como 2 partes por 100 partes de monómero en peso (phm) utilizando un sonificador Branson W700 con un 38% de potencia por 15 minutos, siendo esta la fase oléica en el sistema de polimerización. Asimismo, se preparó una solución micelar en una relación peso surfactante/agua (Latemul 180/H^2O) de 0.8/99.2%, agitando durante 30 minutos; luego se adicionó la fase oléica y se mantuvo por 5 minutos. Después, los componentes de la emulsión se colocaron en un reactor enchaquetado con flujo de nitrógeno para prevenir la oxidación. La reacción se realizó a una velocidad de 400 rpm, a 80 °C, durante 2 h. También se controló el pH con una solución de NaOH al 10% como agente neutralizante para modificar la acidez de la emulsión.

Caracterización

Las NPZnO fueron caracterizadas mediante un análisis de XRD llevado a cabo en un difractómetro SIEMENS D-5000 con radiación CuKα para identificar la fase cristalina de las nanopartículas. Las NPZnO sin y con modificación y el PSA fueron caracterizadas con un JEOL 1200XII TEM para obtener la distribución de tamaños de partícula y la incorporación de las nanopartículas en el PSA. El análisis de FT-IR fue llevado a cabo en un espectrofotómetro Nicolet Magna 5500 utilizando ATR de las muestras de los PSA obtenidos para identificar las bandas de absorción presentes en cada uno de los materiales sintetizados. La determinación de la temperatura de transición vítrea (Tg) se determinó mediante calorimetría diferencial de barrido (DSC) en un DSC modelo 2920 (TA Instruments) a presión constante en una amplitud de ±1 °C cada 60 s a una velocidad de 10 °C/min hasta temperatura ambiente.

Resistencia a la delaminación

Las pruebas mecánicas que se practicaron a los adhesivos sin y con NPZnO, para determinar la fuerza de delaminación y de corte, se realizaron depositando material adhesivo en una película de polietileno, las cuales fueron puestas a secar a temperatura ambiente por 48 h para evaporar el agua y el monómero residual presente. Después se cortaron tiras de 25.4 mm de ancho, preparadas mediante un recubrimiento preciso de 0.03 mm de espesor del adhesivo con y sin nanopartículas aplicado, bajo condiciones fijadas por los estándares de ASTM, a una temperatura de 23 °C y 50% de humedad relativa y bajo los procedimientos establecidos por la ASTM para la evaluación de los materiales sintetizados en este trabajo, Pelaje: D3330/D 3330M-04 y adhesión de corte: D3654/D 3654M-02, la cual marca una velocidad de prueba de 300 mm/min (11.8100 in/min). Se utilizó un rodillo de 2.06 Kg el cual se pasó tres veces por la muestra para ejercer presión sobre ella. Los ensayos se llevaron a cabo en una máquina de ensayos universal marca Instron.

Determinación de la actividad antimicrobiana de las nanopartículas y de los adhesivos

Con la finalidad de evaluar la sensibilidad antimicrobiana de los nanocompuestos y las nanopartículas en estudio, se procedió a la realización de pruebas microbiológicas, las cuales comprenden el método de difusión con disco en agar de acuerdo con los estándares marcados por CLSI (Instituto de Estándares Clínicos y de Laboratorio). En el caso de las nanopartículas se pusieron discos de nanopartículas de 1 cm de diámetro obtenidos por compresión y de los nanocompuestos se colocaron muestras de 2 cm de diámetro provistas de un soporte de polietileno.

La suspensión bacteriana se extendió en 3 planos sobre la superficie de la placa de agar sangre (AS) para *S. pyogenes* y Mueller Hinton (MH) *S. aureus* usando un hisopo de algodón. Los discos de NPs y de los nanocompuestos con diferentes concentraciones de NPs se depositaron sobre el agar inoculado, con al menos 3 cm de separación una de otra y no más de 5 discos por placa de agar. Cabe mencionar que el ensayo se realizó por triplicado. Las placas con agar se incubaron en un ambiente aeróbico a 37 °C, de 18 a 24 h. Al día siguiente se registraron las zonas de inhibición, de cada sistema en estudio, para cada una de las bacterias analizadas, se realizaron los cálculos respectivos para obtener la media de las tres repeticiones realizadas.

Ensayos de citotoxicidad

Se utilizó la línea celular HeLa (ATCC CCL-2), fueron crecidas utilizando Dulbecco's Modified Eagle Medium (DMEM) y en el medio RoswellPark Memorial Institute (RPMI-1640), respectivamente, fueron incubadas a 37 °C en una cámara de CO_2 al 5%. Buffer fosfatos, suero bovino fetal, glutamina y piruvato de sodio (Gibco), y bromuro de 3-(4,5-dimetiltiazol-2-ilo)-2,5-difeniltetrazol (MTT, Sigma Aldrich).

Las células fueron cultivadas en placas de 96 pocillos, donde se sembraron 8,000 células por pozo y se incubaron con los adhesivos en forma de látex sin y con NPZnO, para determinar la citotoxicidad. Después de 24 h, se midió la viabilidad celular mediante el ensayo MTT [8]. Para ello, se añadió a cada pocillo 80 μL del reactivo MTT a una concentración de 5 mg/mL (Sigma-Aldrich St. Louis, Mo. USA), y se incubaron por 4 h.

Lectura de resultado

Pasadas las 4 h se retiró el sobrenadante de cada pocillo y se añadieron 800 μL de DMSO (Dimetilsulfóxido) para disolver los cristales formados. 200 μL se transfirieron a una caja de 96 pozos para realizar la lectura a 595 nm en un lector de placas de ELISA marca BioRad. El control sin NPs se tomó como un 100% de viabilidad celular y a partir de este punto se calculó el porcentaje de viabilidad del resto de los pocillos con la siguiente fórmula. % de viabilidad = (lectura 595 nm tratamiento/lectura 595 nm control sin tratamiento) X100.

RESULTADOS Y DISCUSIÓN

El patrón de difracción de XRD de las NPZnO obtenidas a partir del método de precipitación química se muestra en la Figura 1.

Las nanopartículas obtenidas presentaron una estructura hexagonal tipo wurtzita correspondiente al ZnO. El diámetro promedio de partícula se calculó a través de la ecuación de Debye-Scherrer [9], usando el ancho medio máximo β de las líneas de difracción de rayos X.

$$D = \frac{K\lambda}{\beta \cos\theta} \qquad (1)$$

Donde, D es el diámetro de la partícula, k es la constante de scherrer, λ es la longitud de onda de los rayos X, β la anchura del pico a la mitad del máximo, θ es el ángulo de difracción de Bragg. De acuerdo con esta ecuación el diámetro determinado de las NPZnO mediante esta técnica fue de 17 nm aproximadamente.

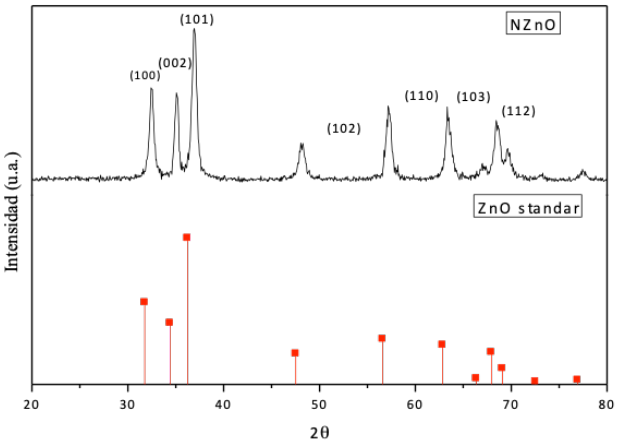
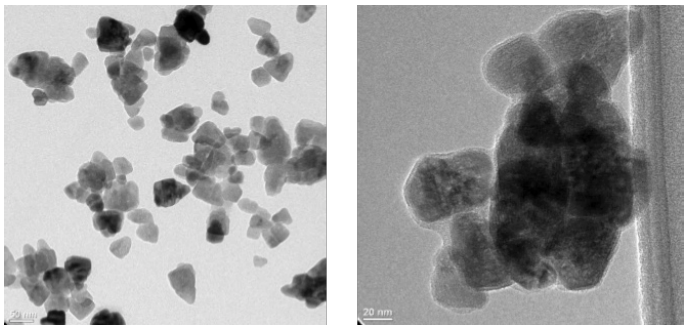


FIGURA 1. Patrón de difracción de Rayos-X de las NPZnO.

La Figura 2a y b muestra una micrografía de TEM de las NPZnO sin modificación y la imagen de las NPs modificadas superficialmente con APTES, respectivamente. La Figura 2a muestra las NPZnO, las cuales presentan un tamaño promedio de 41 nm, estas tienden aglomerarse debido a la alta energía superficial de las nanopartículas. La Figura 2b muestra la imagen de las NPs modificadas en la que se observa un recubrimiento depositado sobre la superficie de las nanopartículas como resultado de la modificación. El espesor del agente de modificación no se aplicó uniformemente sobre la superficie de las NPZnO, como se puede apreciar en la Figura 2b. El espesor del recubrimiento de las nanopartículas se relaciona con el tiempo de funcionalización, temperatura y relación molar utilizada entre otros. El recubrimiento puede estar unido químicamente a la nanopartícula por medio de enlaces Si-O o enlaces covalentes, ya que durante el tratamiento se generan algunos radicales libres del ZnO en la superficie de las nanopartículas y estos pueden interactuar químicamente [10]. Esto produce que las nanopartículas tienden a tener una mejor dispersión cuando son incorporadas a una matriz polimérica, lo cual es atribuido al impedimento estérico entre las nanopartículas que reduce así su tendencia a aglomerarse [11]. Este efecto puede deberse a la interacción que tienen los agentes de modificación con los monómeros, debido a que hay un cambio en el grado de incorporación de las NPs con los agentes de modificación.



a) b)

FIGURA 2. Micrografías obtenidas por TEM de las nanopartículas de ZnO sin modificar a) y b) modificadas con APTES.

La modificación superficial de las NPs puede llevarse a cabo en forma individual o como grupos de partículas. Es muy difícil recubrir sólo partículas individuales, ya que es muy conocido que las nanopartículas se adhieren entre sí debido a su alta energía superficial [10].

En la Figura 3 se muestra el espectro infrarrojo de los PSA con y sin NPZnO al 0.3%. En ambos casos, las bandas características de IR corresponden al estiramiento del enlace C=O de los grupos ésteres presentes en el copolímero en 1734 cm^{-1}. De igual manera las bandas de 1241-1161 cm^{-1} corresponden a la flexión del enlace C-O, mientras las bandas de estiramiento de las cadenas saturadas (C-H) aparecen entre 3000-2880 cm^{-1}, las cuales corresponden a los metilos y metilenos presentes [12]. A diferencia del espectro infrarrojo del PSA, en el FT-IR del PSA/NPs también aparece una señal de alrededor 1578 cm^{-1}, la cual corresponde a los residuos de monómeros (C=C) [13], esto debido a que no se alcanzó una total conversión. Adicionalmente, también en el espectro infrarrojo del PSA/NPs aparece una banda alrededor de 618 cm^{-1}, la cual corresponde a la banda de flexión del enlace Zn-O [14].

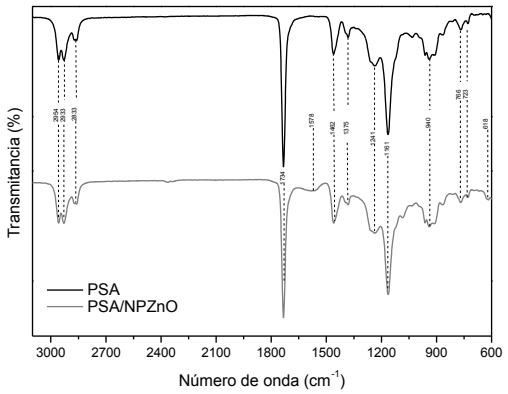


FIGURA 3. Espectro infrarrojo de los adhesivos sensibles a la presión (PSA) antes y después de la adición de nanopartículas.

La Tabla 1 presenta el estudio de las pruebas DSC que fue realizado a los adhesivos PSA/NPZnO, PSA/NPZnO-DMSO, PSA/NPZnO-APTES utilizando diferentes concentraciones de NPZnO (Tabla 1). En el PSA con menor concentración de nanopartículas sin modificar muestra una Tg de -59.35 °C la cual se incrementa conforme se incrementa la concentración de nanopartículas a -47.9 °C, mientras que con las NPZnO-APTES y NPZnO-DMSO se observa también el mismo efecto debido a que la estructura de los adhesivos se vuelve más rígida en presencia de las nanopartículas. Este comportamiento produce un incremento en las propiedades de pegajosidad (Tack) y adhesión de los adhesivos, además que nos indica que las NPZnO producen una alta estabilidad térmica, ya que retrasan la transmisión rápida de calor y lo que puede limitar la descomposición del material [15].

TABLA 1. Transición vítrea (Tg) de nanocompuestos de PSA con diferentes partículas.

Muestra	Contenido de nanopartículas (%)				
	Tg (°C)	0	0.1	0.2	0.3
PSA		-59.35	---	---	---
PSA/NPZnO		---	-50.29	-47.9	-50.71
PSA/NPZnO-APTES		---	-49.21	-49.51	-48.63
PSA/NPZnO-DMSO		---	-48.3	-69.29	-51.9

Actividad antimicrobiana

Una vez modificadas las nanopartículas, se procedió a analizar el efecto antimicrobiano de éstas en comparación con las no modificadas, así como su efecto cuando son incorporadas a los PSA. Estos análisis se realizaron utilizando la técnica de difusión de disco en agar para los microorganismos *S. aureus* y *S. pyogenes* como se muestra en la Figura 4. Las NPs sin modificar y modificadas superficialmente exhibieron un efecto inhibitorio sobre el crecimiento bacteriano, para cada microorganismo. Este efecto puede ser debido a que como las bacterias Gram positivas carecen de una membrana protectora alrededor de las capas de peptidoglicano en la pared celular, esto les permite interactuar y romper la membrana externa más fácilmente y así inhibir, con mayor eficiencia, el crecimiento de bacterias Gram positivas [16]. La diferencia en la actividad antimicrobiana de las NPs modificadas y no modificadas se relaciona con el número de vacancias de oxígeno que son típicas de las propiedades del ZnO. Un aumento en el número de vacancias de oxígeno hace que las NPs se carguen positivamente

y, por lo tanto, realza las interacciones electrostáticas entre las NPs. Para el caso de las NPs modificadas, el efecto inhibitorio se puede atribuir a la presencia de los enlaces Si-O que dan una buena cobertura superficial a las NPs y actúan como una barrera energética. La modificación superficial genera un aumento en la permeabilidad de la membrana y la penetración celular. Esto puede dar lugar a una disminución en la cantidad de proteína en las células expuestas en las NPs modificadas porque el ZnO es altamente reactivo con estas moléculas [17].

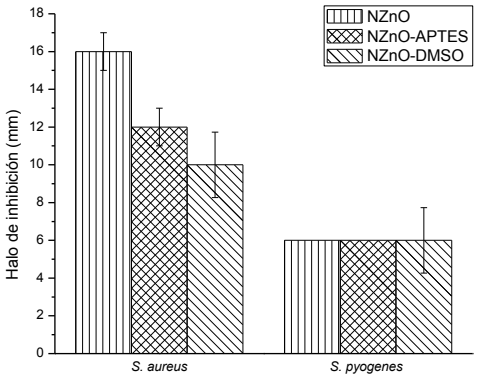


FIGURA 4. Halos de inhibición de las nanopartículas de óxido de zinc sin y con modificación frente a *S. aureus* y *S. pyogenes*.

La actividad antimicrobiana de los PSA con las NPs modificadas y sin modificar presenta un efecto inhibitorio muy similar contra ambos microorganismos cuando se incrementa la concentración de NPs en los PSA, ya que generan un aumento en el número de moléculas de oxígeno activo que causan la muerte celular. Lo que sugiere que las NPs se adhieran a la membrana celular interrumpiendo su respiración e interactuando con un tipo particular de enzimas causando así la muerte celular. Los resultados obtenidos nos indican que el uso de NPZnO sin modificar y modificadas como agentes antimicrobianos en matrices poliméricas produce grandes beneficios para la salud ya que estos microorganismos no generan resistencia a este tipo de NPs, contrario a la resistencia que pueden generar a algunos antibióticos (Figura 5).

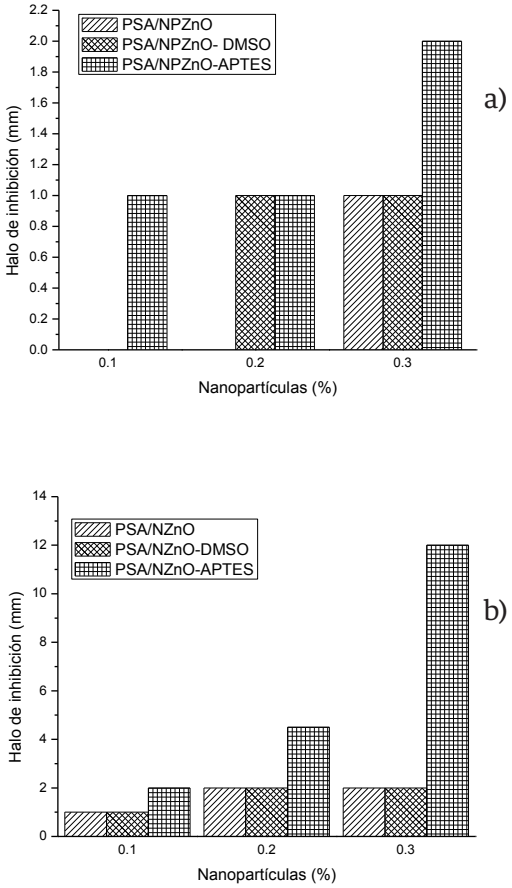


FIGURA 5. Halos de inhibición de los nanocompuestos frente a a) *S. aureus* y b) *S. Pyogenes*.

Viabilidad celular

Los resultados de viabilidad celular evaluada mediante el método MTT, señalan que los PSA sin y con la presencia de NPZnO (Figura 6), poseen una viabilidad del 90 y 80 % con respecto al control sin tratamiento. Esta disminución en la viabilidad celular puede ser atribuida a la liberación del ión Zn^{+2}, el cual es capaz de penetrar en compartimiento ácido de los liposomas e incrementar la generación de especies reactivas de oxígeno (ROS) [18] Se sabe también que el incremento en la disolución de los iones Zn^{+2} en el medio, es capaz de incrementar la citotoxicidad de las partículas, por lo cual es posible que los iones metálicos liberados por NPZnO contribuyeran a la toxicidad [19]. Sin embargo NPZnO, en las concentraciones añadidas, no resultan ser tóxicas, para las células HeLa acorde a la norma.

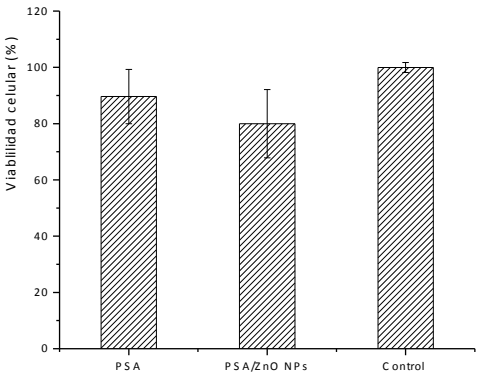


FIGURA 6. **Viabilidad celular de los adhesivos sensibles a la presión (PSA) con y sin de nanopartículas.**

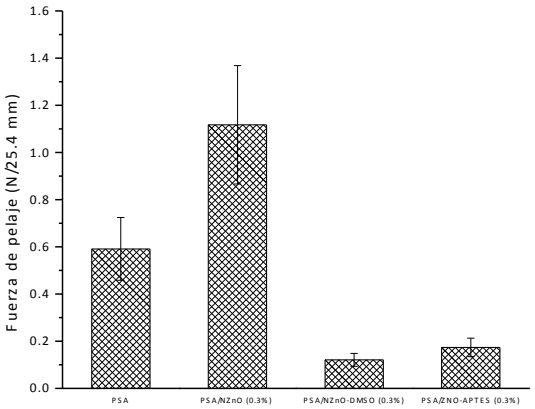


FIGURA 7. **Fuerza de pelado de los nanocompuestos con diferentes nanopartículas.**

Resistencia al pelado (Prueba de adhesión a 180º)

Las propiedades de resistencia a la adhesión de los PSA preparados son las que definen el desempeño y uso final de los PSA (Figura 7). Los PSA sin y con la presencia de NPs muestran un marcado efecto en las pruebas de resistencia al delaminado (peel sterngth). En la muestra que contiene la mayor concentración de NPs sin modificar muestra un valor de 1.1 N/25mm que es el que presenta mejores propiedades. Esto se puede relacionar con el comportamiento en la Tg puesto que el incremento en la Tg produce un incremento en la adhesión [20]. Siendo estos resultados bastante adecuados para pegarse y despegarse de manera segura a la piel y alrededor de la herida, lo que ocasionaría menor trauma al ser removido de la piel.

CONCLUSIONES

Se sintetizó un copolímero de 2-EHA/MMA al cual, después de añadirle NPZnO presenta diferencias en composición y por consecuencia en su Tg. La modificación de las nanopartículas de ZnO incrementa las propiedades antimicrobianas frente a *S. aureus* y *S. pyogenes*, obteniéndose actividad antimicrobiana desde 0.1% del PSA-NPs. La adhesión se ve alterada por la modificación de las nanopartículas presentándose una mayor adhesión en el PSA-NPZnO. Esto es debido a que la modificación de las nanopartículas disminuye la presencia de grupos OH superficiales los cuales son responsables de la adhesión. En cuanto a la citotoxicidad, se muestra que ésta disminuye cuando se le incorporan NPs, sin embargo, esta disminución no es significativa.

AGRADECIMIENTOS

El primer autor agradece el apoyo al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca 263277 otorgada por cursar el doctorado de Tecnología en Polímeros del Centro de Investigación en Química Aplicada. Los autores agraden al CONACYT a través del Laboratorio Nacional de Grafeno (CONACYT-232753) por las facilidades otorgadas. También se agradece el apoyo brindado a: A. Espinoza, F. Zendejo, R. Cedillo, S. Zertuche, A. Herrera.

REFERENCIAS

[1] Rasmussen JW., Martínez E, Louka P, Wingett DG. (2010). Zinc oxide nanoparticles for selective destruction of tumor cells and potential for drug delivery application. Expert Opin Drug Deliv. 7(9), 1063-1077. doi.10.1517/17425247.2010.502560

[2] Mascorro R., Dorantes H. Y Corea M. (2012). Morphology influence over the peeling property in pressure-sensitive adhesives. Rev. Mex. Ing. Quim., 11(2), 323-331.

[3] Jingyeong O. (2009). Synthesis and Adhesion Performance of Polyacrylate-clay Pressure-sensitive Adhesive as nanocomposite by in-situ polymerization. Tesis de Maestría en Ciencias Forestales. Universidad Nacional de Seúl, Corea.

[4] Lofton L. (2004). Clay/polymer nanocomposites for pressure sensitive adhesives. Rohm and Haas Company, Spring House, PA.

[5] Kajtna, J., Šebenik U. (2009). Microsphere pressure sensitive adhesives-acrylic polymer/montmorillonite clay nanocomposite materials. International Journal of Adhesion and Adhesives 29(5), 543-550. doi.10.1016/j.ijadhadh.2009.01.001

[6] Rana, P.K. Sahoo P.K. (2007). Synthesis and pressure sensitive adhesive performance of Poly (EHA-co-AA)/Silicate nanocomposite used in Transdermal drug delivery. Journal of Applied Polymer Science 106(6), 3915-3921. doi.10.1002/app.27034

[7] Wang, T., Lei C.-H, Dalton A. B., Creton C., Lin Y., Fernando K.A.S., Sun Y.P., Manea M., Asua J.M., Keddie J.L. (2006). Waterborne, nanocomposite pressure-sensitive adhesives with high tack energy, optical transparency, and electrical conductivity. Advanced Materials,. 18(20), 2730-2734. doi.10.1002/adma.200601335

[8] Mosmann T. AshaRani PV, Low Kah Mun G, Hande MP, Valiyaveettil S. (1983). Rapid colorimetric assay for cellular growth and survival: Application to proliferation and cytotoxicity assays. J Immunol Methods 65(1-2), 55-63. doi.10.1016/0022-1759(83)90303-4

[9] Chen, L., Zhengb L., Lva Y., Liua H., Wanga G., Rena N., Liua D, Wanga J., Boughtonc R.. Surface and Coatings Technology, 2010. 204(23): p. 3871-3875.

[10] Esteves A.C., Timmons A.B., Trinidae T. Nanocompositos de matriz polimérica: estratégias de síntese de materiais híbridos. (2004) Quim. Nova. 27 (5) 798-806.

[11] Tang E., Cheng G., Ma X., Pang X., Zhao Q. (2006). Surface modification of zinc oxide nanoparticle by PMAA and its dispersion in aqueous system. Applied Surface Science 252(14), 5227-5232. doi.10.1016/j.apsisc.2005.08.004

[12] Kagel N. (1971) Infrared Spectra of Inorganic Compounds Chemical Physics Research Laboratory. Academic Press, Inc.

[13] Taghizadeh SM, Ghasemi D. (2010). Synthesis and Optimization of a Four-component Acrylic-based Copolymer as Pressure Sensitive Adhesive. Iranian Polymer Journal. 19(5), 343-352.

[14] Purcar V., Somoghi R., Georgiana M., Nicolae C., Alexandrescu E., Gîfu I., Gabor A., Stroescu H., Ianchis R., Căprărescu S., Cinteză L. The Effect of Different Coupling Agents on Nano-ZnO Materials Obtained via the Sol-Gel Process. Nanomaterials. 2017. 7, 493.

[15] Jovanović, R., Dubé M.A. (2004). Emulsion-Based Pressure-Sensitive Adhesives: A Review Journal of Macromolecular Science, Part C 44(1), 1-51. doi.10.1081/MC-120027933

[16] Emami-Karvani Z., C. P, African Journal of Microbiology Research, 2011. 5(12): p. 1368-1373.

[17] Yousef J. and Danial. N.E., Journal of Health Sciences, 2012. 2(4): p. 38-42.

[18] Vasile O., Serdarua I., Andronescua E., Truşcă R., Surdua V., Oprea O., Iliea A., Vasile S. (2015). Influence of the size and the morphology of ZnO nanoparticles on cell viability. Comptes Rendus Chimie 18(12), 1335-1343. doi.10.1016/j.crci.2015.08.005

[19] Zhang X.,Wang Z, Mao L., Dong X., Peng O., Chen J., Tan C., Hu R. (2017). Effect of ZnO nanoparticle on cell viability, zinc uptake efficiency, and zinc transporters gene expression: a comparison with ZnO and ZnSO4. Czech J. Anim. Sci. 62(1), 32-41. doi.10.17221/15/2016-CJAS

[20] Wu, L., Wang M., Zhang X., Chen D., Zhong A. (2009). Organic montmorillonite modified polyacrylate nanocomposite by emulsion polymerization. Iranian Polymer Journal 18(9), 703-712.

ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-11

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.6

E-LOCATION ID: **e201801EE1**

Más Allá de La Filogenética: Evolución Darwiniana de La Actina

Beyond Phylogenetics: Darwinian Evolution of Actin

Marcelo A. Moret[1], Gilney Zebende[1], James C. Phillips[3], J. Quetzalcóatl Toledo-Marín[4], Gerardo G. Naumis[4]

[1]Universidade do Estado da Bahia

[2]Department of Physics, State University of Feira de Santana, Bahia, Brazil

[3]Rutgers University, The State University of New Jersey

[4]Universidad Nacional Autónoma de México, Instituto de Física

RESUMEN

La actina es una proteína que se polimeriza para formar citoesqueletos y cuya función es estabilizar y dirigir el movimiento de las paredes celulares. Es una de las proteínas más estables, habiendo evolucionado poco a partir de algas y levaduras, y muy poco desde los peces. Aquí analizamos la evolución de la actina usando las teorías modernas de las interacciones de conformación proteína-agua, y cómo estas han evolucionado para optimizar las funciones de la proteína. Llegamos a la conclusión de que el fracaso del análisis filogenético para identificar positivamente la evolución darwiniana de las proteínas ha sido causado por las limitaciones técnicas propias del siglo XX. Estas limitaciones pueden ser superadas mediante el escalamiento termodinámico y el promedio modular ambos llevados a niveles técnicos del siglo XXI. Los resultados para la actina son especialmente llamativos y reflejan estructuras duales estables, globulares y polimerizadas.

PALABRAS CLAVE: proteínas; evolución Darwiniana; actina

ABSTRACT

Actin polymerizes to form cytoskeletons which stabilize and direct motion of cellular walls. It is one of the most stable proteins, having evolved little from algae and yeast, and very little from fish. Here we analyze actin evolution using modern theories of water-protein shaping interactions, and how these have evolved to optimize protein functions. We conclude that the failure of phylogenetic analysis to identify positive Darwinian evolution has been caused by 20th century technical limitations. These are overcome using 21st century thermodynamic scaling and modular averaging. The results for actin are especially striking, and reflect dual stable structures, globular and polymerized.

KEYWORDS: proteins; Darwinian evolution; actin

Correspondencia

DESTINATARIO: **Gerardo García Naumis**

INSTITUCIÓN: **Universidad Nacional Autónoma de México, Instituto de Física**

DIRECCIÓN: **Circuito Exterior S/N, Ciudad Universitaria, Coyoacán, CDMX, México**

CORREO ELECTRÓNICO: **naumis@fisica.unam.mx**

Fecha de recepción:

20 de agosto de 2018

Fecha de aceptación:

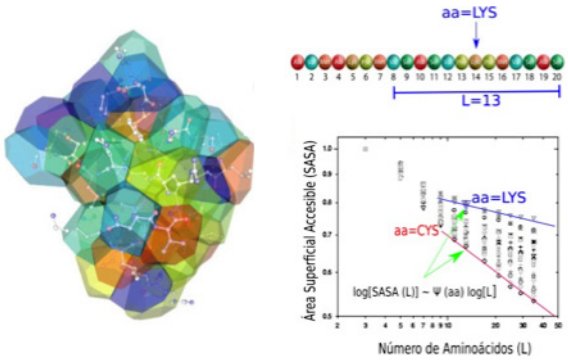
9 de enero de 2019

INTRODUCCIÓN

El advenimiento de bases de datos genómicos muy grandes ha hecho que el análisis filogenético de las secuencias de aminoácidos en proteínas sea un tema atractivo y desafiante, ya que las inferencias con respecto a la selección natural pueden proporcionar información funcional importante [1]. Hay disponibles muchos programas de computadora para el análisis de datos filogenéticos. Actualmente el programa más popular cubre una amplia gama de opciones, mediante un código adaptable y fácil de usar [2], ofreciendo así a los usuarios máximos resultados con un mínimo esfuerzo. El objetivo final de establecer el poder de la selección darwiniana para mejorar la función de la proteína aún no se ha alcanzado [3-5]. Sin embargo, los esfuerzos para cuantificar los relojes moleculares que se iniciaron en la década de 1960 por Pauling y otros ya ha dado resultados positivos [6]. En términos más generales, hay muchas dificultades en filogenómica, y se ha dicho que "más secuencias no son suficientes" [7].

La filogenia cuenta aminoácidos idénticos o similares en sitios específicos utilizando el código BLAST, y se puede refinar de muchas maneras [2], pero todas estas están limitadas por la restricción a sitios únicos. Existe una alternativa a los métodos de sitio único, teniendo la selectividad darwiniana como una característica implícita, y que ha sido corroborada por la identificación de la criticalidad auto-organizada en las áreas superficiales accesibles a solventes (SASA, del inglés “solvent-accessible surface areas”) para más de 5000 segmentos de aminoácidos de proteínas pertenecientes a la moderna Protein Data Base [8, 9].

Esta criticalidad autoorganizada puede entenderse en el contexto más amplio de los sistemas complejos. En estos sistemas, la auto-organización se refiere a la capacidad de un sistema para generar un comportamiento colectivo emergente partiendo de interacciones entre sus constituyentes. Un ejemplo típico en física lo son las fases de la materia, donde las moléculas se auto-organizan al imponérseles condiciones externas como presión o temperatura, dando lugar a comportamientos colectivos, como lo puede ser por ejemplo la resistencia a fluir, lo cual determina si un material se comporta como líquido, sólido o gas. Al variar las condiciones externas aplicadas a un sistema, los cambios entre estos comportamientos no son suaves; de hecho implican discontinuidades en las propiedades termodinámicas, lo cual da lugar a las llamadas transiciones de fase. En 1869 el profesor de química Thomas Andrew encontró que existen condiciones de presión y temperatura donde la diferencia entre fases deja de existir. Así, arriba de cierta presión y temperatura, no es posible distinguir el agua de su vapor. A estos puntos se le llama críticos. Un poco después de su descubrimiento experimental, el físico holandés Van der Waals logró mostrar su existencia de manera teórica. Más aún, Van der Waals logró demostrar que si el fluido se describe escalando las variables termodinámicas en términos de los parámetros en el punto crítico, el fluido llevarse a una descripción universal que ya no depende del sistema en particular. Más allá de la termodinámica, la criticalidad es un comportamiento robusto al cual se llega sin importar los parámetros del modelo y sin necesidad de que el sistema esté en equilibrio. Hay dos características importantes de la criticalidad auto-organizada: una es la capacidad del sistema para mantenerse cerca de ese estado, y la otra es la existencia de la invariancia de escala (fractalidad) [9]. Esta invariancia indica que el sistema se ve igual cuando se observa a una escala diferente. Un ejemplo típico lo constituye la costa de un país, la cual se ve muy parecida al examinarse en mapas de diferente escala. Ello se debe a que existe un mecanismo de erosión y una auto-organización que mantiene a la costa es un estado crítico. Otro ejemplo son las nubes, las cuales también están en un estado crítico y por ello se ven blancas, ya que dispersan la luz de manera igual para todas las longitudes de onda. Como veremos, las proteínas también presentan invariancia de escala [8, 9].



a) b)

FIGURA 1. a) Fragmento de la estructura de una proteína, indicada por esferas unidas por barras. Sobrepuesta a ella se aprecian los poliedros de Voronoi, los cuales delimitan las regiones del espacio en la cuales todos los puntos contenidos dentro del poliedro están más cercanos al átomo situado en el centro del poliedro que a cualquier otro átomo. La superficie exterior neta de estos poliedros permite definir el área superficial accesible a las moléculas del agua (SASA). En la parte de arriba del panel b) se muestra la secuencia de aminoácidos (aa) de la proteína. Para estudiar su relación con la SASA, se escoge un aminoácido, en este caso cistina (aa=LYS), y se considera un intervalo de L aminácidos. Realizando una estadística sobre muchos fragmentos para diferentes L y proteínas, se obtiene la gráfica log-log de la parte b). Para L>9, los datos en la gráfica log-log pueden ajustarse con una recta, indicada en azul, y cuya pendiente da el parámetro Ψ(aa). Considerando otros aa se obtienen rectas similares, por ejemplo para aa=CYS se obtiene la línea roja. Estas pendientes definen la escala de hidropatía MZ del 2007, donde la CYS es el aa más hidrofóbico y LYS el aa más hidrofílico.

En vista de la discusión del párrafo precedente, es claro que el descubrimiento de la criticalidad auto-organizada de la SASA ya implica una selección darwiniana para acercar el proteoma a puntos críticos funcionales. Esto proporciona una plataforma para analizar la evolución de las proteínas individuales, tales como el lisozima c de la clara de huevo de gallina [10], la neuroglobina [11] y muchas otras [12].

El primer paso para desarrollar un nuevo método es probarlo en muchos casos específicos, examinando cada uno de los nuevos datos que proporciona. Debido a que cada familia de proteínas tiene una o más funciones diferentes, uno aprende algo nuevo en cada caso. Sin embargo, ciertos aspectos han ido surgido hasta ahora. Uno de ellos se refiere a la naturaleza del SASA de segmentos auto-organizados estudiados por Moret y Zebende [8]. Para entender este descubrimiento, en la Figura 1 se detallan los pasos seguidos por Moret y Zebende en el año 2007 [8].

Para ello consideraron que la hidropatía está determinada de acuerdo al área accesible al agua, la cual a nivel molecular en la cercanía de la proteína tiene una estructura más parecida al hielo que al agua. El área accesible puede obtenerse usando la construcción de Voronoi, en la cual el espacio se subdivide en poliedros, cada uno centrado en un átomo de la estructura. El poliedro de Voronoi se define de modo que todos los puntos en su interior están más cercanos al átomo central del poliedro que de cualquier otro átomo. Dada la estructura de la proteína, esto permite obtener una imagen como la que se aprecia en la Figura 1 a), y obtener así la SASA, que corresponde al área externa de los poliedros. Posteriormente, Moret y Zebende estudiaron como variaba la SASA para secuencias de aminoácidos (aa) de diferentes tamaños centrados en un aa dado, tal y como se muestra en la Figura 1b. Las longitudes de sus pequeños segmentos L =2N + 1 variaron de 3 a 45, pero el rango interesante resultó ser M <=9 ≤ L ≤ M>=35. A través de éste, encontraron un comportamiento lineal en un diagrama log-log (es decir, una ley de potencias y por lo tanto, auto-similar) para cada uno de los 20 aminoácidos centrados en un segmento dado. Es decir,

$$\log[\text{SASA}(L)] \sim \text{const} - \Psi(aa)\log[L\ (9 \le L \le 35)]$$

Aquí Ψ(aa) es un parámetro del índice de hidropatía para cada aa, y como puede verse en la Figura 1b), se obtiene de la pendiente de la recta que surge al realizar

la gráfica log-log de L contra la SASA. Surge porque los segmentos más largos se repliegan sobre sí mismos, ocluyendo el SASA del aa central. El aspecto más sorprendente de esta oclusión plegada auto-similar es su casi universalidad en promedio a través del proteoma celular, y que es casi independiente del pliegue de la proteína individual. Esta es una demostración dramática del poder de la selectividad darwiniana involucrada en la formación acuosa de proteínas globulares, como se discute en detalle en la referencia [12]. Además, el carácter segmentario de la nueva escala [8] tiene un eco darwiniano: para cada familia de proteínas se puede identificar un ancho de ventana móvil optimizado W*, sobre el cual Ψ(aa) se promedia mejor; este promedio se denota por Ψ(aa,W*). Los perfiles de Ψ(aa,W*) muestran las características funcionales modulares optimizadas por la evolución [12].

El centro del rango de la SASA (una ley de potencias, y que describe estructuras auto-similares), $9 \leq L \leq 35$, es 21. Las proteínas de membrana funcionan en el lado citoplásmico de la membrana celular [13], esta también descrita como un sustrato catalítico que soporta interacciones de tipo proteína-proteínas en el espacio fronterizo interfacial [14]. Las diferencias evolutivas entre la proteína gigante Hub Sr y la tirosina quinasa Syk se describen mejor mediante W* = 21 (+/- 5%) [15]. Es posible considerar a la actina como la proteína de membrana arquetípica, debido a su función esquelética en el soporte de membranas, al mismo tiempo que tiene suficiente flexibilidad para permitir cambios de forma funcionales termodinámicamente clasificados como de segundo orden. Además, la actina empuja las membranas hacia adelante durante el crecimiento celular [16, 17], lo cual es termodinámicamente de primer orden.

La actina ha evolucionado muy poco. Se puede comparar con la ubiquitina, la cual tiene sólo 76 aminoácidos, y se encuentra sin cambios en mamíferos, aves, peces e incluso gusanos. La ubiquitina marca las proteínas enfermas para el reciclaje mediante una cascada de enzimas en tres etapas, iniciada por Uba (E1). Debido a su pequeño tamaño, además de su papel central en biología, la ubiquitina se ha convertido en uno de los más importantes sistemas modelo para estudiar la dinámica de proteínas y ha sido objeto de numerosos estudios, incluidos algunos que combinan enfoques experimentales y computacionales. Un estudio reciente con simulaciones de dinámica molecular a gran escala (> 5000 moléculas de agua) de la Ub globular identificó movimientos conformacionales lentos que implican correlaciones globulares estabilizadoras de cadenas β cerca de los terminales n y c en la escala de tiempo de microsegundos a milisegundos [18]. La función de etiquetado de la Ub es consistente con su perfil hidrofóbico y sugiere un modelo de red elástica agrietada para el blanco común compartido por muchas proteínas enfermas [19]. La Uba (E1) es 14 veces más grande que la Uba y también ha evolucionado muy poco, pero esa pequeña evolución rastrea la nivelación de los extremos hidrofóbicos del parámetro Ψ (aa,W*) [20]. Esta nivelación pivotante refleja la optimización de la dinámica de proteínas por su evolución [12, 21].

RESULTADOS

La actina es una proteína de tamaño mediano con 377 aminoácidos, y es tan estable que para obtener cambios evolutivos sustanciales, se deben comparar humanos con algas (con 84% de identidades según BLAST, 92% positivas). Primero observamos la función Vr(W), es decir, la razón algas / humanos de las varianzas de Ψ(aa,W*), que se muestra en la Figura 2, la cual contiene la primera sorpresa de la actina (de hecho, las funciones de las proteínas son todas diferentes, por lo que generalmente hay sorpresas, especialmente con proteínas extremadamente estables, "casi perfectas" como la actina y la ubiqutina). La mayoría de las veces, W* es el valor de W que maximiza la razón de la varianza Vr, pero aquí W* corresponde a un máximo en la derivada dVr/ dW. ¿Qué ha ocurrido? La actina realiza dos funciones, la estabilización de las formas de las células después de pequeñas distorsiones de

segundo orden durante el funcionamiento, y la polimerización para empujar las membranas hacia adelante durante el crecimiento celular [16, 17], lo cual es es termodinámicamente de primer orden. Aparentemente, ambas funciones se optimizan al equilibrar los valores de W pequeños para W <21 (importante para la polimerización) [16, 17], frente a los valores de W grandes de W (importante para estabilizar formas de células grandes). Este equilibrio elimina el máximo en Vr y genera, en cambio, un máximo en dVr/dW en W* = 21.

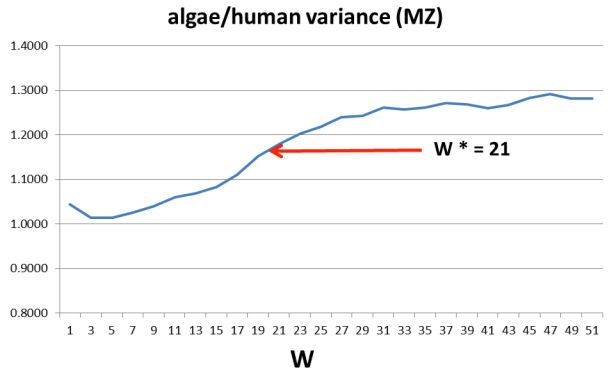


FIGURA 2. Razones de varianza de la actina en función del ancho W de la ventana móvil promediadora. El máximo en dVr/dW ocurre en W*=21. Secuencia humana P68133, secuencia de algas P53500.

Claramente esta imagen resulta atractiva, sin embargo, por sí misma parece poco convincente. Se puede probar al usarlo para perfilar con Ψ(aa,W*=21) los cambios evolutivos de la actina de las algas a los humanos (Figura 3). Por supuesto, con la conservación del sitio ~ 90%, estos cambios son pequeños, pero son sorprendentemente consistentes con estudios previos de nivelación de extremos hidrofóbicos (pivotes elásticos) [12, 21]. Nótese la nivelación de los extremos hidropáticos en la región central 180-280 para humanos en comparación con la actina de algas. El pico correspondiente a un comportamiento hidrofílico cerca del sitio 110 es más profundo en humanos. Finalmente, los puntos extremos de los terminales n y c son ambos más hidrofílicos en la actina humana, lo que facilita la construcción de filamentos y bandas filamentosas más grandes [16, 17].

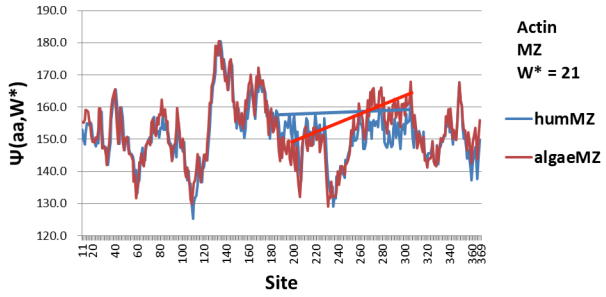


FIGURA 3. Debido a la fuerte conservación evolutiva, los perfiles de Ψ(aa,W*) para la actina en algas y humanos difieren poco, aunque la naturaleza de las diferencias muestra importantes mejoras darwinianas (evolución positiva) (ver texto). Aquí la hidrofobicidad aumenta con valores crecientes de Ψ(aa,W*), y es hidroneutral en 155.

Esta prueba evolutiva se puede llevar un paso más allá al comparar humanos Ψ(aa, 21) con Ψ(aa, 19) en la región central. Por supuesto, las diferencias son pequeñas, por lo que esta comparación es una prueba severa de la precisión de ambas ventanas móviles y la nivelación pivotal impulsada por la evolución Darwiniana.

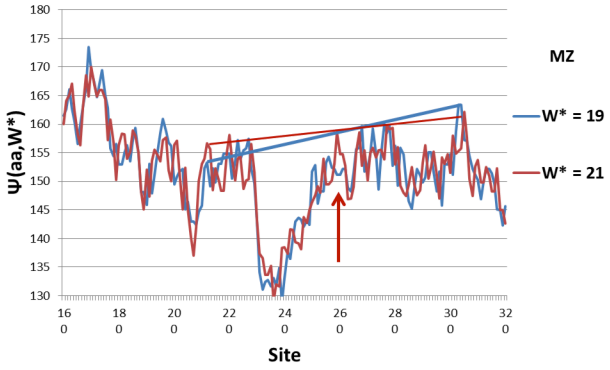


FIGURA 4. Los perfiles de Ψ(aa,W*) para W*=19 y W*=21 son necesariamente muy similares. En la región central que se muestra aquí (con líneas para guiar el ojo), los máximos hidrofóbicos (pivotes elásticos) están más nivelados con W*=21. Cerca del centro de la región de nivel, el perfil W*=21 tiene un pico de nivel extra cerca 260, que está ausente del perfil W*=19. Notar que el mínimo hidrófilo profundo cerca de 240, que está bien conservado de las algas a los humanos, también aquí se cambia poco. En caso de ser fotografías enviarlas en la calidad nativa de la cámara o del medio en donde fueron capturadas.

Como vemos en la Figura 4 y en su leyenda correspondiente, la nivelación pivotante se mejora con W* = 21.

Tal nivelación pivotante precisa se puede utilizar para probar la significancia de las escalas alternativas de hidrofília-hidrofóbia. En el período clásico de la biofísica (antes del 2000), se propusieron no menos de 127 escalas de hidrofília-hidrofóbia. Cada escala tenía sus méritos y se basaba en, como mucho, solo unas pocas docenas de mediciones. Pocos intentos se hicieron para comparar sus exactitudes o aplicabilidad a propiedades distintas de las utilizadas en sus definiciones [12]. Las correlaciones entre escalas fueron típicamente ~ 70%. La escala estándar para las tasas de mutación (BLOSUM 62, utilizada en BLAST) exhibe un mínimo hidroneutral profundo en las tasas de mutación cerca de su centro [22]. Con la escala de MZ del 2007, este mínimo se asocia con alanina (A), glicina (G), el aminoácido más pequeño e histidina (H). La Tabla I de [8] muestra que ninguna de las escalas antiguas coloca los tres aminoácidos en su centro. En términos de la raíz de desviaciones cuadráticas medias respecto al valor promedio de cada escala, las diferencias fuera del centro son 7 veces o más mayores para las otras escalas que para la escala MZ. Por mucho, la escala más popular del período clásico es la escala de 1982 basada en la diferencia de entalpía del agua al aire de los péptidos cortos [23]. Esta escala KD ocupa el segundo lugar después de la escala MZ. En otras palabras, los esfuerzos anteriores al 2000 implicados en la construcción de 127 escalas exploraban en una buena dirección, pero las proteínas son tan complejas que el éxito solo fue posible bio-informáticamente después de que las estructuras PDB se hicieron numerosas y más precisas [8].

A la luz de esta perspectiva histórica, las grandes diferencias en los perfiles hidrofílicos de la actina con la escala de MZ de 2007 y la escala de KD de 1982 que se muestran en la Figura 5 no son sorprendentes. Una de las ventajas fundamentales de la escala MZ es que su carácter fractal está asociado con el enfoque evolutivo de las proteínas hacia una funcionalidad óptima (un punto crítico termodinámico [10]). La criticalidad es característica de la funcionalidad de redes neuronales [24-27] y las proteínas en general [28]. Las matemáticas fractales son bien conocidas por los matemáticos, pero su aplicación a las proteínas se ha desarrollado lentamente [29, 30]. Tal característica de la evolución a veces ha sido objetivo en el diseño de redes informáticas [31].

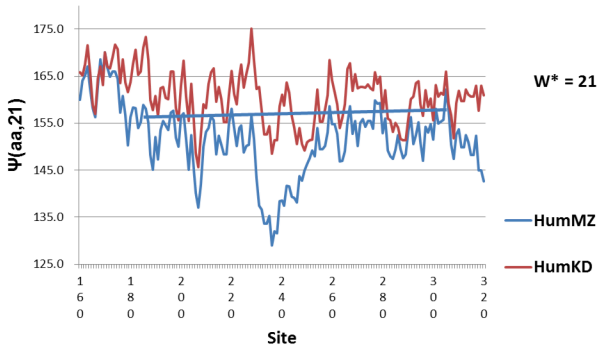


FIGURA 5. Los perfiles humanos se comparan utilizando la escala KD de 1982 (la mejor de las 127 escalas clásicas) y la moderna escala MZ del 2007. Los detalles más fino de la nivelación de pivote están ausentes del perfil KD. Más importante aún, el perfil de KD ha perdido por completo el mínimo hidrofílico profundo cerca de 240, que está bien conservado de algas a humanos (Figura 3). Esto significa que la evolución Darwiniana se puede reconocer en perfiles hidropáticos solo mediante el uso de la escala MZ.

La similitud de sitios ~ 90% se encuentra no solo para algas / humanos, sino también para levaduras / humanos e incluso levaduras / algas. Los perfiles hidropáticos para los dos últimos pares se muestran en las Figuras 6 y 7. Note la nivelación de dos mínimos hidrofílicos en levaduras. La comparación de levaduras y algas en la Figura 7 enfatiza el segmento hidrofóbico estabilizador de algas 265-281. La similitud de sitios 271-281 es <30%, por lo que estas grandes diferencias localizadas ocurren a pesar del 90% de la similitud general de sitios.

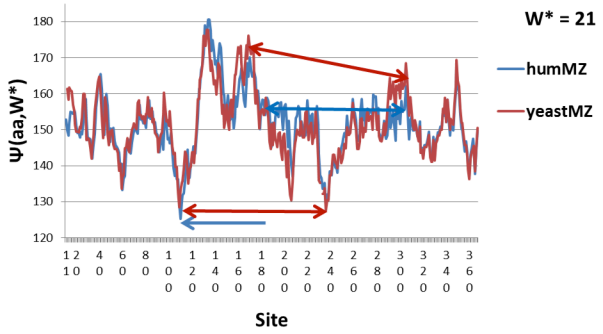


FIGURA 6. Los perfiles de humanos y levaduras se comparan utilizando la moderna escala MZ de 2007. Las diferencias más importantes están marcadas y discutidas en el texto. Secuencia de levadura P60010.

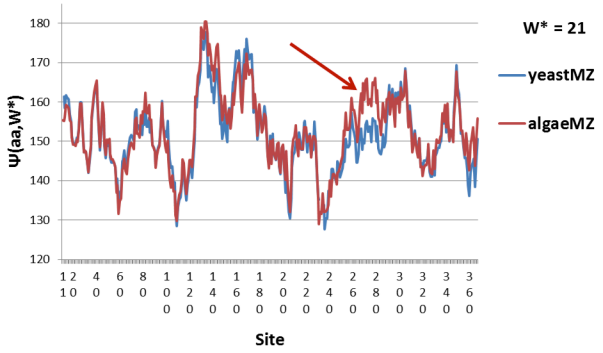


FIGURA 7. Las algas se estabilizan con relación a la levadura mediante un pico hidrofóbico adicional cerca del sitio 275.

¿Qué sucede cuando aumentamos la similitud a más del 98%? Este es el caso al comparar la actina humana con la del pez cebra. Las razones de varianza R de la Figura 8 poseen características interesantes. Se pudo haber anticipado el extremo en W* = 21, pero ¿qué pasa con los extremos de Vr en M <= 9 y de dR / dW en M> = 35? ¿Por qué los valores de corte en la escala MZ de hidropatía aparecen en Vr para la actina de pez cebra y humano? Una razón es que la actina ha evolucionado de pez cebra a humano al optimizar su hidrostructura para dos estados, el globular tal y como se preparó y los polimerizados funcionales [16, 17]. Esto satisface dos condiciones a gran escala y topológicamente opuestas introduciendo correlaciones adicionales más allá de W* = 21, usando exactamente el rango de auto-similitud de MZ para M <≤ 9 ≤ M> = 35. Sin duda, es mucho más que una asombrosa coincidencia.

En los 377 sitios de aminoácidos de la actina de pez cebra/humano, solo hay 5 mutaciones diferentes. La comparación del perfil W* = 21 en la Figura 9 muestra que estos no son accidentales. Las 5 mutaciones diferentes profundizan tres extremos hidrofílicos cerca de 110, 160 y 290. Estos también son las casi expuestas vueltas en la estructura globular en 110, 168 y 283, PDB 1J6Z [32]. En el pez cebra, los mínimos en 110 y 236 están al mismo nivel, mientras que en la actina humana el mínimo en 110 es más profundo.

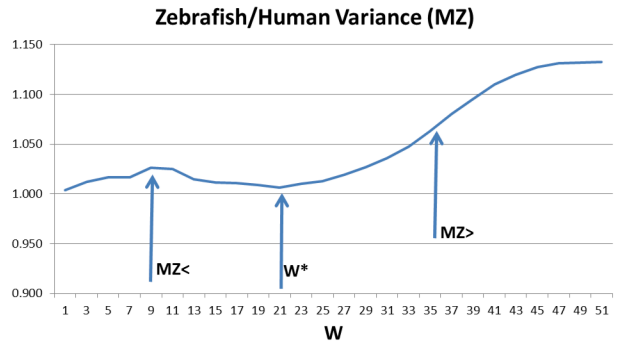


FIGURA 8. La comparación de las proporciones de varianza Vr del pez cebra y humano muestra un extremo en W*=21. Más aún, la función tiene características analíticas en los límites inferior y superior del rango fractal MZ: un extremo en MZ <, y un máximo en dVr/dW en MZ>. La secuencia de pez cebra es la AAH71401.

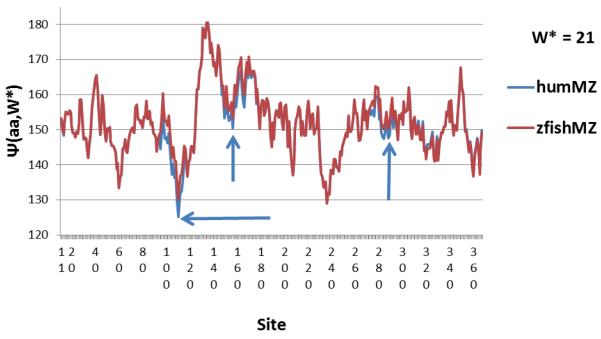


FIGURA 9. Las pequeñas diferencias entre la actina humana y del pez cebra se concentran cerca de tres extremos hidrofílicos.

La forma hidropática de la actina es inusual porque sus estados duales, a decir, globulares y polimerizados, deben ser estables y funcionar de manera reversi-

ble. En la Figura 8 vimos que esta dualidad se refleja en los extremos de Vr en W* = 21 y M <= 9 y un extremo de dVr/dW en W = 35 con la escala MZ. Tal vez tales características analíticas son razonables, dado que la escala MZ en sí misma es fractal y refleja la criticalidad autoorganizada, también rasgo característico del citoesqueleto en la escala celular [33]. ¿Se conserva alguna de esta estructura cuando utilizamos la escala KD [23]? Los resultados que se muestran en la Figura 10 son bastante inesperados. Se esperaría ver una Vr(W) cualitativamente diferente, pero esta función diferente todavía tiene puntos críticos en M <, W* = 21 y M>. Una explicación plausible es que la escala KD se basa en las diferencias de entalpía agua-aire, y estas energías de primer orden están involucradas en la polimerización mediante la unión de los terminales n y c [16, 17].

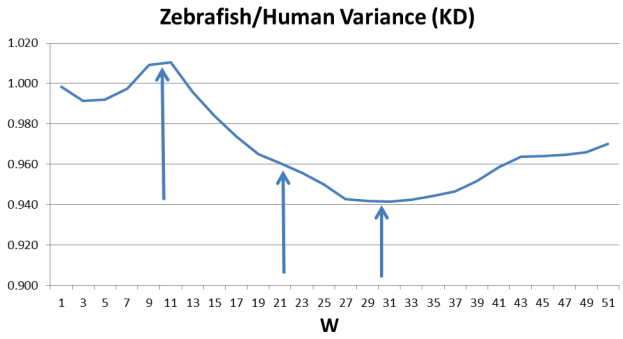


FIGURA 10. La forma de la relación de varianza Vr(W) se muestra aquí usando la escala KD; es cualitativamente diferente de los resultados mostrados en la Figura 8 usando la escala MZ, pero todavía muestra puntos críticos analíticos cerca de M<(máximo), W*=21 mínimo en dVr/dW y M> (mínimo).

DISCUSION

El descubrimiento de fractales termodinámicos en el SASA de> 5000 segmentos de proteínas [8], junto con promedios modulares utilizando ventanas móviles W, ya ha llevado a la observación de muchas conexiones cercanas entre la secuencia y la función de muchas proteínas, especialmente para proteínas de membrana [12]. La selección darwiniana positiva generalmente ha hecho que sea fácil optimizar W y encontrar W*, estableciendo así la selección Darwiniana de pasada. Aquí hemos encontrado tres valores de MZ >> 1 de W*, lo que sugiere que se han producido dos papeles para la evolución de la actina, una para la actina estabilizada en su forma globular y otra para la polimerización de la actina que estabiliza los citoesqueletos celulares. En general, ninguno de estos resultados evolutivos están presentes cuando W = 1 (por ejemplo, las Figuras 2 y 8), por lo que la filogenia ha sido incapaz de identificar la evolución Darwiniana a nivel molecular [2-7].

El uso de términos como pivotes y bisagras sugiere modelos elastométricos. Estos son intuitivamente atractivos [34, 35]. La red creciente de actina se ha modelado como un sistema autoorganizado en criticalidad, en el que los esfuerzos mecánicos de largo alcance que surgen de la interacción con la membrana de plasma proporcionan la presión selectiva que lleva a la organización. La sincronización del citoesqueleto aparece naturalmente como resultado de la criticalidad autoorganizada [35], y se ve facilitada por la nivelación de los pivotes hidropáticos [12]. Este modelo celular se ha reformulado cuantitativamente a nivel molecular evolutivo aquí.

No hemos podido imaginar simulaciones que pudieran derivar estas características críticas de la actina, las cuales involucrarían al menos dos moléculas fusionadas: ~ 750 aminoácidos + agua. Las simulaciones más avanzadas de dinámica molecular en ubiquitina (77 aminoácidos) ahora revelan correlaciones de largo alcance de movimientos consistentes con escalamiento termodinámico [36]. También se puede suponer que la escala fractal universal de las interacciones agua-proteína podría conducir a las características críticas dobles de la actina. Incluso cuando se impone la condición relativamente suave de elasticidad de la columna rígida (por su nombre en inglés, "blackbone elasticity") para la percolación, las fracciones de escala irracionales asociadas con las caminatas aleatorias se reemplazan por fracciones simples [37].

REFERENCIAS

[1] Nielsen, R. Molecular signatures of natural selection. Ann. Rev. Genetics 39, 197-218 (2005). DOI: 10.1146/annurev.genet.39.073003.112420

[2] Tamura, K.; Peterson, D.; Peterson, N.; et al. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Bio. Evolu. 28, 2731-2739 (2011). DOI: 10.1093/molbev/msr121

[3] Suárez-Díaz, E.; Anaya-Muñoz, V. H. History, objectivity, and the construction of molecular phylogenies. Stud. Hist. Phil. Biol. & Biomed. Sci. 39 (4): 451-468 (2008). DOI:10.1016/j.shpsc.2008.09.002

[4] Nozawa, M.; Suzuki, Y.i; Nei, M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc. Nat. Acad, Sci. (USA) 106, 6700-6705 (2009). DOI: 10.1073/pnas.090185510

[5] Nozawa, M.; Suzuki, Y.i; Nei, M. The neutral theory of molecular evolution in the genomic era. Ann. Rev Gen. Human Gen. 11, 265-289 (2010). DOI: 10.1146/annurev-genom-082908-150129

[6] Omland, K. E. Correlated rates of molecular and morphological evolution. Evolution 51, 1381-1393 (1997). DOI: 10.1111/j.1558-5646.1997.tb01461.x

[7] Philippe, H.; Brinkmann, H.; Lavrov, D. V.; et al. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Bio. 9 (3): e1000602 (2011). DOI: 10.1371/journal.pbio.1000602

[8] Moret, M. A.; Zebende, G. F. Amino acid hydrophobicity and accessible surface area. Phys. Rev. E 75, 011920 (2007). DOI: 10.1103/PhysRevE.75.011920

[9] Phillips, J. C (2009) Scaling and self-organized criticality in proteins: Lysozyme c. Phys. Rev. E 80, 051916. DOI: 10.1103/PhysRevE.80.051916

[10] Phillips, J. C. Fractals and self-organized criticality in proteins. Phys. A 415, 440-448 (2014). DOI:10.1016/j.physa.2014.08.034

[11] Sachdeva, V.; Phillips, J. C. Oxygen channels and fractal wave-particle duality in the evolution of myoglobin and neuroglobin. Phys. A 463, 1-11 (2016). DOI: 10.1016/j.physa.2016.07.007

[12] Phillips, J. C. Quantitative molecular scaling theory of protein amino acid sequences, structure, and functionality. *https://arxiv.org/abs/1610.04116*

[13] Ota, M.; Gonja, H.; Koike, R.; et al. Multiple-Localization and Hub Proteins. PLOS ONE 11, e0156455 (2016). DOI: 10.1371/journal.pone.0156455

[14] Wei G, Xi W, Nussinov R, et al. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. Chem. Rev. 116, 6516-6551 (2016). DOI: 10.1021/acs.chemrev.5b00562

[15] Phillips, JC Giant Hub Src and Syk Tyrosine Kinase Thermodynamic Profiles Recapitulate Evolution. Phys. A 483, 330-336 (2017). DOI: 10.1016/j.physa.2017.04.180

[16] Pollard, T. D.; Borisy, G. G. Cellular motility driven by assembly and disassembly of actin filaments. Cell 112, 453-465 (2003). DOI: 10.1016/S0092-8674(03)00120-X

[17] Pollard, T.D.; Cooper, J. A. Actin, a central player in cell shape and movement. Science 326, 1208-1212 (2009). DOI: 10.1126/science.1175862

[18] Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; et al. Picosecond to millisecond structural dynamics in human ubiquitin. J. Phys. Chem. B 120, 8313-8320 (2016). DOI: 10.1021/acs.jpcb.6b02024

[19] Allan, D. C.; Phillips, J. C. Why Ubiquitin Has Not Evolved. Phys. A 491, 377-381 (2018). DOI: 10.3390/ijms18091995

[20] Allan, D. C.; Phillips, J. C. Evolution of the ubiquitin-activating enzyme Uba1 (E1). Phys. A 483, 456-461 (2017). DOI: 10.1016/j.physa.2017.04.144

[21] Osher, S.; Sethian, J. A. Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton--Jacobi Formulations. J. Comp. Phys. 79, 12—49 (1988). DOI: 10.1016/0021-9991(88)90002-2

[22] Sachdeva, V.; Phillips, J. C. Hidden thermodynamic information in protein amino acid mutation tables. Phys A 469, 676-680 (2017).

[23] Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105-132 (1982). DOI: 10.1016/0022-2836(82)90515-0

[24] Eguiluz, V.M.; Chialvo, D.R.; Cecchi, G.A.; et al. Scale-free brain functional networks. Phys. Rev. Lett. 94, 018102 (2005). DOI: 10.1103/PhysRevLett.94.018102

[25] Chialvo, D.R. Emergent complex neural dynamics. Nature Phys. 6, 744-750 (2010). DOI: 10.1038/NPHYS1803

[26] Baliki, M.N.; Geha, PY.; Apkarian, A. V.; et al. Beyond feeling: Chronic pain hurts the brain, disrupting the default-mode network dynamics. J. Neurosci. 28, 1398-1403 (2008). DOI: 10.1523/JNEUROSCI.4123-07.2008

[27] Tagliazucchi, E.; Chialvo, D.R.; Siniatchkin, M.; et al. Large-scale ignatures of unconsciousness are consistent with a departure from critical dynamics. J. Roy. Soc. Interface 13, 20151027 (2016). DOI: 10.1098/rsif.2015.1027

[28] Tang, Q.-Y.; Zhang, Y.-Y.; Wang, J.; et al. Critical fluctuations in the native state of proteins. Phys. Rev. Lett. 118, 088102 (2017). DOI: 10.1103/PhysRevLett.118.088102

[29] Glockle, W. G.; Nonnenmacher, T. F. A fractional calculus approach to self-similar protein dynamics. Biophys. J. 68, 46-53 (1995). DOI: 10.1016/S0006-3495(95)80157-8

[30] Maiorino, E.; L.ivi, L.; Giuliani, A.; et al. Multifractal characterization of protein contact networks. Phys. A 428, 302-313 (2015).

[31] Bentley, P.J. Evolving beyond perfection: an investigation of the effects of long-term evolution on fractal gene regulatory networks. Biosyst. 76, 291-301 (2004). DOI: 10.1016/j.biosystems.2004.05.019

[32] Otterbein, L. R.; Graceffa. P.; Dominguez, R. The crystal structure of uncomplexed actin in the ADP state. Science 293, 708-711 (2001). DOI: 10.1126/science.1059700

[33] Cardamone, L.; Laio, A.; Torre, V.; et al. Cytoskeletal actin networks in motile cells are critically self-organized systems synchronized by mechanical interactions. Proc. Nat. Acad. Sci. (USA) 108, 13978-13983 (2011). DOI: 10.1073/pnas.1100549108

[34] Sinha, N.; Kumar, S.; Nussinov, R. Interdomain interactions in hinge-bending transitions. Structure 9, 1165-1181 (2001). DOI: 10.1016/S0969-2126(01)00687-6

[35] Yang, L.; Song, G.; Jernigan, R. L. How well can we understand large-scale protein motions using normal modes of elastic network models? Biophys. J. 93, 920-929 (2007). DOI: 10.1529/biophysj.106.095927

[36] Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; et al. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin.J. Phys. Chem B 120, 8313-8320 (2016). DOI: 10.1021/acs.jpcb.6b02024

[37] Sampaio, F.; Cesar, I. N.; Andrade, J.S.; Jr.; et al. Elastic Backbone Defines a New Transition in the Percolation Model. Phys. Rev. Lett. 120, 175701 (2018). DOI: 10.1103/PhysRevLett.120.175701

ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-17

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.7

E-LOCATION ID: **e201807EE1**

Signal-Processing tools for core-collection selection from genetic-resource collection

Herramienta de procesamiento de señales para la selección de "Core-Collection" desde colección de recursos genéticos

L. I. López-Flores[1], *Masaru Takeya*[2], *E. Borrayo*[1]

[1]Universidad de Guadalajara

[2]National Agriculture and Food Research Organization

ABSTRACT

Selecting a representative core collection (CC) is a proven and effective strategy for overcoming the expenses and difficulties of managing genetic resources in gene banks around the globe. Because of the diverse applications available for these sub-collections, several algorithms have been successfully implemented to construct them based on genotypic, phenotypic, passport or geographic data (either by individual datasets or by consensus). However, to the best of our knowledge, no single comprehensive datasets has been properly explored to date. Thus, researchers evaluate multiple datasets in order to construct representative CCs; this can be quite difficult, but one feasible solution for such an evaluation is to manage all available data as one discrete signal, which allows signal processing tools (SPTs) to be implemented during data analysis. In this research, we present a proof-of-concept study that shows the possibility of mapping to a discrete signal any type of data available from genetic resource collections in order to take advantage of SPTs for the construction of CCs that adequately represent the diversity of two crops. This method is referred to as 'SPT selection.' All available information for each element of the tested collections was analysed under this perspective and compared when possible, with one of the most used algorithms for CC selection. Genotype-only SPT selection did not prove as effective as standard CC selection did not prove as effective as standard CC selection algorithms; however, the SPT approach can consider genotype alongside other types of information, which results in well-represented Ccs that consider both the genotype and agromorphological diversities present in original collections. Furthermore, SPT-based analysis can evaluate all available data both in a comprehensive manner and under different perspective, and despite its limitations, the analysis renders satisfactory results. Thus, SPT-based algorithms for CC selection can be valuable in the field of genetic resources research, management and exploitation.

KEYWORDS: Core Collection, SPT, Genotype, Genebank

RESUMEN

La selección de una colección núcleo (core-collection) representativa (CC) es una estrategia comprobada y eficaz para superar los gastos y las dificultades de la gestión de los recursos genéticos en los bancos de germoplasma de todo el mundo. Debido a las diversas aplicaciones disponibles para estas subcolecciones, se han implementado con éxito varios algoritmos para construirlos en base a datos genotípicos, fenotípicos, de pasaporte o geográficos (ya sea por conjuntos de datos individuales o por consenso). Sin embargo, hasta donde tenemos conocimiento, no se han explorado adecuadamente conjuntos de datos integrales hasta la fecha. Por lo tanto, los investigadores evalúan conjuntos de datos múltiples para construir CCs representativos; esto puede ser bastante difícil, pero una solución factible para tal evaluación es administrar todos los datos disponibles como una señal discreta, que permite implementar herramientas de procesamiento de señal (SPT) durante el análisis de datos. En esta investigación, presentamos un estudio de prueba de concepto que muestra la posibilidad de asignar a una señal discreta cualquier tipo de datos disponibles de colecciones de recursos genéticos para aprovechar los SPT para la construcción de CC que representen adecuadamente la diversidad de dos cultivos. Este método se conoce como "selección de SPT." Toda la información disponible para cada elemento de las colecciones analizadas se analizó bajo esta perspectiva y se comparó cuando fue posible, con uno de los algoritmos más utilizados para la selección de CC. La selección de SPT de solo genotipo no resultó tan efectiva como los algoritmos de selección de CC estándar; sin embargo, el enfoque SPT puede considerar el genotipo junto con otros tipos de información, lo que da como resultado CCs bien representados que consideran tanto el genotipo como las diversidades agromorfológicas presentes en las colecciones originales. Además, el análisis basado en SPT puede evaluar todos los datos disponibles, tanto de manera integral y bajo diferentes perspectivas, y a pesar de sus limitaciones, el análisis arroja resultados satisfactorios. Por lo tanto, los algoritmos basados en SPT para la selección de CC pueden ser valiosos en el campo de la investigación, gestión y explotación de recursos genéticos.

PALABRAS CLAVE: Core Collection, SPT , Banco de germoplasma, genotipo

Correspondencia

DESTINATARIO: **Ernesto Borrayo Carbajal**

INSTITUCIÓN: **Universidad de Guadalajara**

DIRECCIÓN: **Av. Juárez #976, Col. Centro, C.P. 44100, Guadalajara, Jalisco, México**

CORREO ELECTRÓNICO:

ernesto.borrayo@academicos.udg.mx

Fecha de recepción:

27 de septiembre de 2018

Fecha de aceptación:

9 de enero de 2019

INTRODUCTION

One of the most promising techniques for conserving the diversity of genetic resources is *ex situ* genebank germoplasm collection. A significant effort has been made on a global scale to perserve, characterize, distribute and utilise genetic resource in order to understand their biological phenomena and confront the vulnerable situation regarding the sustainability of future human development [1, 2]. As the size of germoplasm colections increase, it becomes difficult to appropriately manage and extensively evaluate them [3]; thus, the core collection (CC) concept [4] has become a fundamental genetic resource management approach and exploits the potential of a complete collection in terms of viable data management and moetary expenses [5, 6, 7, 8].

Different CCs have different purposes characteristics and evaluation criteria [7, 9, 10, 11]; thus, several different algorithms and informatics tools have been developed and implementend [12, 13, 14, 15] with different approaches for satisfying particular needs of each CC. Because these CCs are constructed mainly on the basis of genotypic, phenotypic, passport or geographic data (either by individual datasets or by consensus) [16], there is a lack of all-inclusive datasets; this limits the possibility of generating a CC that may satisfy most basic and applied genetic resource research programs. To the best of our knowledge, no single comprehensive datasets has been properly explored to date.

One possible method to create a comprehensive dataset is to represent the availabe data as numerical values. Several methods exist that represent genomic information into numerical values [17] and agromorphological traits (ATs) into scores [18]. Through this mapping process, treating each data vector as a discrete signal that can, in turn, be analysed by signal processing tools (SPTs) is possible, thus providing and effective tool for a comprehensive evaluation of datasets. We present a proof-of-concept study that shows the possibility of mapping to a discrete signal any type of data available from genetic resource collections in order to take advantage of SPTs for CC selections; this possibility provides new decision-making criteria for genetic resource management and research.

METHODOLOGY

Mapping data

Each input data must be mapped to a numeric value. This is a fundamental process of the algorithm beacause it enable different datasets to be analysed together, regardless of their nature. In this manner, dissimilar passport data, single nuceotide polymorphisms (SNPs), restriction fragment length polymorphisms (RFLPs), geographic information and phenotypical traits can be included in one comprehensive dataset. To consistently represent each data type, reference tables are implemented according to the nature of each particular data: genetic information (originally represented as character elements) is now represented by a numeral vector, and trait variation, simple sequence repeat (SSR) molecular markers and passport data can be represented as either binary or normalized data depending on the quantitative/qualitative nature of the data. The original data and reference tables for this study are available in supplementary material ??. Data transformation for this study rendered a matrix containing the representation of MC samples (i_1, i_2, i_3, ..., i_n) with (j_1, j_2, j_3, ..., j_n) elements each, where n is the total number of samples, and m is the number of included samples characteristics, represented by a numerical values as $data_{(i,j)}$.

Signal construction

Numerical representations of each jth *data* element can be treated as frequency values in m data time in such a manner that each ith sample is treated as a discrete signal. The i signal correspond to the information behaviour from each sample. This perspective will enable the implementation of SPTs such as the dis-

crete Fourier transform and power spectrum comparision. Although SPTs can be implemented on all data available for each sample, not all data elements contain the same informativeness value to discriminate between samples. To overcome the informative difference in each *j* element of *data*, a principal component analysis (PCA) can be perfomed to rearrange *data* into a new matrix that has the high informative elements of *data* at the beginning and that arranges subsequent elements acording to their informativeness, discarding those whose variance equals 0. This process renders two new matrices: the original *characteristics* mapped vectors matrix (*x*) and rearranged variance value matrix (*X*). Matrix *X*, therefore, contains *n* samples that are formed by a numerical vector with *m=m-* (non informative *characteristics*).

Fast Fourier transform

The main objective of Fourier transform is the decomposition of any signal into a complex histogram of frequencies. Signal function is then represented as a vectorial function whose angle and magnitude determine a sampled point in the signal [19].

The original Fourier model is expressed as follows:

$$\hat{f}(\mathcal{E}) = \int_{-\infty}^{\infty} f(x)\, \mathrm{e}^{-2\pi i x \mathcal{E}} dx \qquad (1)$$

where *x* is the temporal variable, ξ it the frequential variable, *i* is a -1 square root and *e* is the natural exponent.

From equation (1), a derivate can be determined for any point ξ sampled in the signal.

$$f(x)[\cos^{2\pi e \mathcal{E}} + i * \sin^{2\pi e \mathcal{E}}] \qquad (2)$$

Fourier transform can be implemented into any complex numerical series, but in a practical sense, the computational cost increases exponentially.

Thus, fast Fourier transform (FFT) is more often implemented and can be defined according to Cooley-Tukey algorithm [20] as follows:

$$X_k = \sum_{n=0}^{N-1} x_n \mathrm{e}^{-i\pi k \frac{n}{N}} \qquad (3)$$

where *N* is the vector length, *x* is the temporal variable, *i* is a -1 square root and *e* is the natural exponent; in such matter that an euclidean representation - with the angle, magnitude and phase that corresponds to their position in the signal - exists for any signal dot.

Therefore, mapping any signal into a vectorial representation that contains information from every original signal dot is possible. From this complex vector, useful data can be retrieved to establish a comparison between them that indirectly represents the original signal's juxtaposition [21].

Distance matrix computation

Inspired by the genomic signal processing alignment-free distance (GAFD) model [22], each signal corresponding to the PCA-mapped accesions data in a set $\hat{S}_i$ was converted into its frequency representation by applying discrete Fourier transform. Its power spectrum F_i was then computed. Subsquently, the distance d(*i,j*) for a given pair of comprehensive data signal was calculated by obtaining the mean square error (MSE) of their respective power spectra:

$$D(i,j) = \sum_{x} (\hat{F}_i(x) - \hat{F}_j(x))^2 \qquad (4)$$

Finally, a distance matrix (DM) was created by performing a pairwise comparison of all sequences in the set.

In parallel, we construted a point-to-point (RAW) DM on the basis of the MSE given to a pair of signal prior to the PCA analysis.

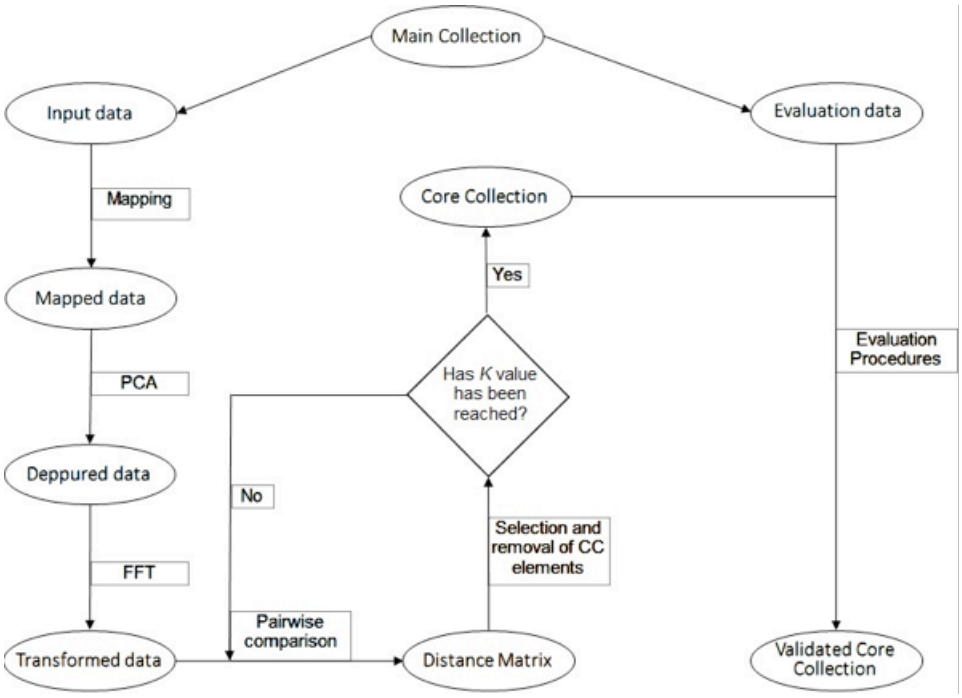


FIGURE 1. General workflow of the FFT-based core collection selection algorithm. PCA: Principal Component Analysis; FFT: Fast Fourier Transfrom; CC: Core Collection.

Core collection selection

Selecting a CC by this method requires the generation of a DM for each sample of the MC; this provides the interrelations among samples and enables adequate selection. A schematic of the complete workflow is present in Fig. 1

In the past, several methodological procedures have been implemented to select K elements from an MC on the basis of information provided by its DM; among such procedures, the most frequently used one is the hierarchical clustering method [11]. However, the current algorithm does not rely on hierarchical clustering for CC selection, instead - similar to the least distance stepwise sampling method [23] - CC elements are selected by an iterative process, where r samples are selected by different criteria (which may be individually implemented) on each iteration.

Selection criteria (based on the MD without hierarchical clustering) for the current algorithm is as follows:

a. The *i*th sample with the most lower distance values among *j*th elements.
b. The *i*th sample with the most higher distance values among *j*th elements.
c. The *i*th sample with a lower distance average.
d. The *i*th sample with a higher distance average.
e. The *i*th sample with a lower overall distance.
f. The *i*th sample with a higher overall distance.

In cases where multiple samples share selection values, an appearance priority will complete the criteria.

An example of selection process is present in Fig. 2 and its final result is present in Fig. 3.

Once the selected samples (r) are included in the future CC, they (along with others that are identical to them (s)) are removed from X for the next iteration; then, a DM_2 with $n_2 = n - r - s$ is calculated. This process will continue Z times until $R >= K$, where $R = (r_1 + r_2 + \ldots + r_Z)$ and K = *predefined* CC *elements desired*.

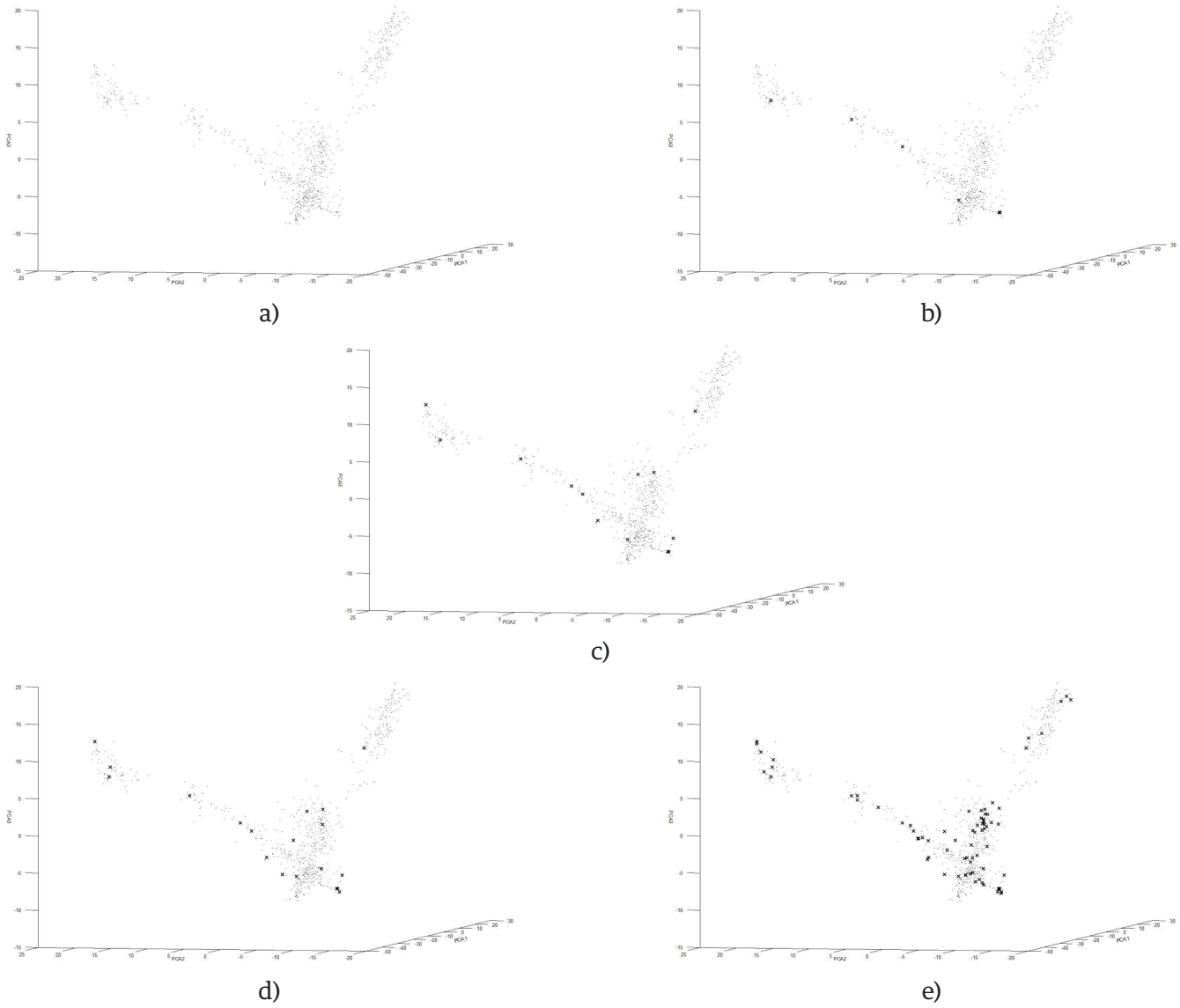


FIGURE 2. **First three principal component's distribution of Rdata (a), methodology's first (b), second (c) and third (d) iteration; final K=72**

Evaluation of the selected core collection

As discussed previously, the best way to evaluate a CC depends on the purpose of that CC, and eve if it can be evaluated from the same dataset from which it was constructed, evaluating it with a diferent dataset [7] is desirable. In this study, we use other datasets for our evaluation whenever possible.The list given below provides the evaluation parameters implemented in this study.

a. The average distance between each MC sample and the nearest CC sample (ANE) can be calculated using the equation as follows:

$$ANE_{tot} = \frac{1}{L}\sum_{k=1}^{K}\sum_{j=1}^{J} D\left(k - cMC_j\right) \qquad (5)$$

where K is all CC elements, k is each CC element and D is the distance between k and each jth cMC

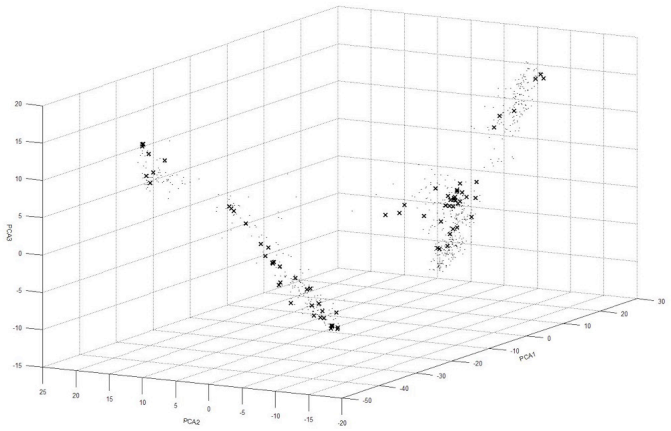


FIGURE 3. **First three principal component's distributions of K=27 CC selection (X) from Rdata MC.**

element whose closest CC element is *k*, including itself, thus rendering *L* total comparisons. The ideal ANE value is 0, where each sample of the CC represents itself and those similar to it. This parameter evaluates the homogeneity of the represented MC diversity.

b. The average distance between each CC sample and the nearest CC sample (ENE) can be calculated using the equation as follows:

$$ENE_{tot} = \frac{1}{L}\sum_{k=1}^{K} D(k - cCC) \qquad (6)$$

where *K* is all CC elements, *k* is each CC element and *D* is the distance between *k* and its closest CC element *cCC*, excluding itself, in *L* total comparisons. With such an evaluation parameter, higher dispersion renders higher scores with the aim of evaluating the dispersion among selected CC elements.

c. The average distance between CC samples (E) can be determined applying the equation as follows:

$$E_{tot} = \frac{1}{L}\sum_{k=1}^{K}\sum_{j=1}^{J} D(k - cCC_j) \qquad (7)$$

where *K* is all CC elements, *k* is each CC element and *D* is the distance between *k* and all other *j*th CC elements *cCC*, excluding itself, in *L* total comparisons. This evaluation parameter indicates higher scores when CC elements have greater distances between themselves.

While previous evaluation parameters are useful for data dispersion analysis, such parameters will not evaluate how well the distribution of the MC is represented on the CC; therefore, the distribution comparisons tests that were included are as follows:

d. The homogeneity test (*F - test* for variances and *t - test* for means; α= 0.05) between the CC and MC for each trait can be represented as a percentage of traits that are statistically different (MD for means and VT for variances) [9].

e. The coincidence rate (CR) can be calculated using the equation as follows:

$$CR = \frac{1}{M}\sum_{m=1}^{M} \frac{R_{CC}}{R_{MC}} \qquad (8)$$

where *R* is the range of each m trait, and *M* represents the number of traits.

f. The variable rate (CV) can be calcuated using the equation as follows:

$$CV = \frac{1}{M}\sum_{m=1}^{M}\frac{CV_{CC}}{CV_{MC}} \quad (9)$$

where *CV* is the coefficient of the variation of each *m* trait in the CC and MC, and *M* is the number of traits. According to Hu *et al*. [10] a valid CC has *CR* > 80 and *MD* <20, which are the limits for the ideal representation of the identity and distribution of the MC.

g. The alleles coverage (CA) can be calculated using the equation as follows:

$$CA = [|1-(|1-ACC|I\,AMC)|] \quad (10)$$

where *ACC* is a set of alleles in the CC, and *AMC* is a set of alleles in the MC; *ACC* meansures the percentage of alleles from the MC that are present in the CC [12].

To compare the obtained CCs with an established methodology, we implemented Core Hunter 2 (CH) [13] as a reference and used it with the program's default parameters on the agrological and genomic datasets.

Experimental datasets

To determine the efficiency of the analysis of data behaviour by point-to-point direct comparison, a synthetic dataset esa constructed using binary data (*Sdata*) with manageable *n* and *m* elements .

To test the algorithm in real biological-context scenarios, the CCs from different Mcs were constructed and evaluated.

To test the algorithm's CCs versus the scores of the MCs, 780 rice (*Oriza sativa* (L.)) accession and 423 foxtail millet (*Setaria italica subspitalica* (L.) *P. Beauv.*) accession data were retrieved from the then National Institute of Agrobiological Sciences (now National Agriculture and Food Research Organization [NARO]) *http://www.gene.affrc.go.jp/databases_en.php* as well as 361 maize (*Zea mays* (L.)) from the International Maize and Wheat Improvement Center public repository.

According to the available data, different datasets were assembled. The 762 SNPs from the 780 rice accesion retrieved from the NARO database (*Rdata*) were divided arbitrarly into two subsets of 331 SNPs each for constructing two smaller datasets (RdataI and *RdataIII*). In addition, ATs were categorized and mapped into the binary data for 273 of the 780 accessions, resulting in 38 variables (*RdataII*). The variables from 423 foxtail millet genotypes with transposon displays [24] were used as a single dataset (*Fdata*). For a subset of 141 accessions (*FdataI*), 9 ATs were categorized and mapped into binary data, resulting in 28 variables (*FdataII*). The maize available information was mapped into 0-1 values (*Mdata*). The substitution tables used during this mapping are presented as supplementary material 1.

Implementation

All procedures were implemented in python 3.6, codes are available as supplementary material 2.

A graphical interface was developed including a SQLite3 database (*https://sqlitebrowser.org/*) in order to store data for future comparison and further analysis. This implementation includes a previously described K-means based CC selection algorithm [25].

RESULTS AND DISCUSSION

Selection and evaluation

The selection criteria were chosen to look for the best possible distribution of selected CC elements within the DM. Although hierarchical clustering has proven to be an effective method for determining collection

structure and sampling CC [26] and although it has been implemented in different crop [27, 28] and included in various selection algortihms [11], hierarchical reconstruction presents the challenge of selecting an appropriate model for biological interpretation that can be applied to everything from unweighted pair-group averages to Markov models in Bayesian estimations [29]. To avoid the challenge of selecting a reconstruction model, we decided to work strictly with the DM. By selecting the items described in this methodology, we aimed to retrieve representative elements from among the distributions of collections; however, because of its iterative nature, this methodology may render high redundancy under certain data distributions. Despite this limitation, the methodology has proven to be capable of selecting representative elements of the MC's diversity.

Evaluation criteria were applied according to Odong *et al*. [7] without excluding the classic criteria used in [9, 10]. The selected CCs render proper results in general terms. As expected, selected CCs did not always reach for optimal values for MD and CR, this is due the fact that it is not the aim of the selection method to render a CC with similar distribution to that of the MC, but to make sure to include as much diversity as possible.

It is our belief that scoring the CC sets obtained with these methodologies will enable genetic resource banks to provide clear descriptors of what their CC strengths and limitations are with respect to the MC from which they come and will provide adequate tools for determining the possible purposes of the selected CCs.

Although several representations of genotypic characteristics (particulary those involving DNA sequences [30, 31, 32]) have been proposed, real-number-based mappings have not been discarded, indeed, this type of mapping has been highly studied for signal analysis even when they share two principal problems: the preferential magnitude of some nucleotides and the non-equidistance of all nucleotides [33, 34]. The arbitrary values selected for SNP's numerical representation of genotypes aim to maintain equidistance relations among purines and among pyrimidines in such a manner that the same distance is also perserved between at least one of them and the undetermined values. ATs are represented as binary data. This representation may prove useful for discrete data but requires a clustering procedure for continuous data. In this study, we arbitrarily generated clusters for the latter and then represented them as the former. Although this implementation may no be the most accurate regarding biological or agronomical significance, it serves as the first approach for testing the feasibility of the use of signal processing techniques when merging several datasets to construct one CC.

RAW versus FFT

The RAW comparison establishes a distance value on the basis of the average distance between each mapped value on each element while the FFT power spectra implementation compares the signals in the frequency domain. Using FFT, establishing a DM on the basis of how data 'shifted' rather than on the basis of average point-to-point comparisons was possible. The FFT approach provides a different DM, where its compared elements are clustered based on the similarity of the shift is in the opposite phase. We expect that the procedure reveal more info about the relations between the individual components within each element.

FFT comparisons of signal without PCA are a good approach for CC selection. Nervertheless, PCA implementation enables us to avoid possible misleads in random data arrangements, as, for example, palindromic data that could result in the same power spectra. Moreover, through PCA, we could organize data according to their levels of impact on the difference between accessions, which --when their magnitudes were obtained-- inherently rendered a representation

TABLE 1. K selected CC scores from MC Sdata Raw and PCA Signal evaluated with Sdata

	Sdata PCA			Sdata RAW		
K	12	18	24	12	18	24
ANE	0.2348	0.2311	0.2164	0.2697	0.2287	0.2164
ENE	0.339	0.3386	0.3401	0.3696	0.3228	0.3214
E	0.5562	0.5622	0.5547	0.5558	0.5333	0.5299
MD	0	0	0	0	0	0
VT	41.6667	50	41.6667	33.3333	58.3333	41.6667
CR	64.8403	71.6918	73.7154	60.6447	75.2465	80.4716
CV	9080.798	61.2074	86.0876	136.6446	139.1418	280.8481
AR	74.3363	81.4159	89.3805	61.9469	77.8761	80.531

TABLE 2. K selected CC scores from MC Fdata Raw and PCA signal evaluate with Fdata

	Edata PCA			Edata RAW		
K	48	72	96	48	72	96
ANE	0.6454	0.6423	0.6407	0.6489	0.6431	0.643
ENE	0.646	0.6472	0.6472	0.65	0.6448	0.6452
E	0.7297	0.7301	0.7301	0.7231	0.7236	0.7239
MD	1.1799	0.59	0.59	1.7699	1.4749	1.4749
VT	50.4425	53.6873	56.6372	50.7375	56.0472	55.1622
CR	83.6883	87.0605	88.9709	83.5334	86.9308	87.7461
CV	0.8494	0.419	0.7357	1.1037	4.74	0.7361
VA	96.3945	97.7652	98.5995	95.3516	97.497	97.4374

TABLE 3. K selected CC scores from MC Rdata Raw and PCA Signal evaluated with Rdata

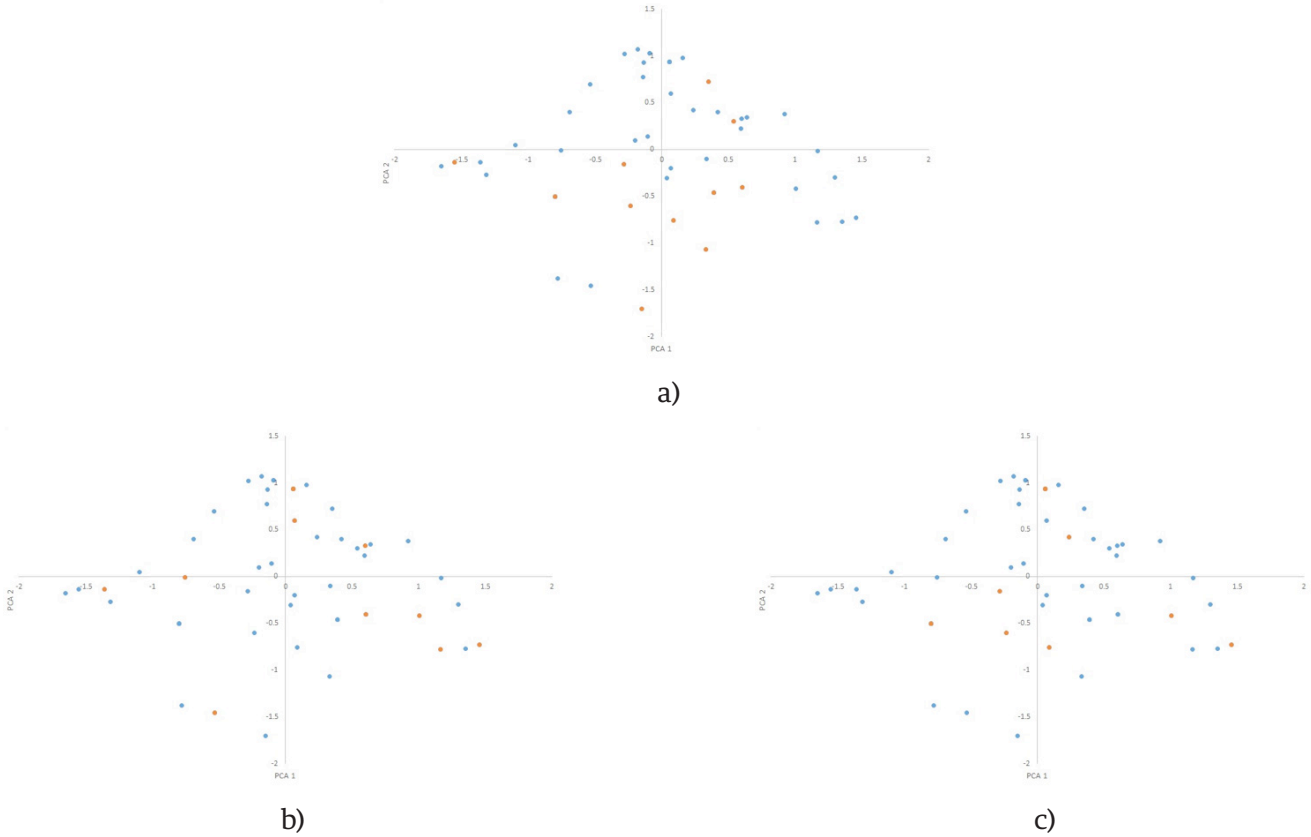
	Rdata PCA			Rdata RAW		
K	48	96	156	48	96	156
ANE	0.6013	0.5966	0.5942	0.6118	0.6052	0.6042
ENE	0.5939	0.5944	0.5981	0.6106	0.6085	0.609
E	0.7105	0.7074	0.7051	0.703	0.7038	0.7054
MD	9.1146	5.9896	3.9062	10.1562	5.4688	4.4271
VT	42.4479	48.6979	58.0729	57.5521	72.9167	70.0521
CR	70.5716	78.477	83.2957	69.9022	78.1045	80.0167
CV	1.0171	0.4343	0.3137	7.9407	0.4375	1.1344
VA	92.6758	96.8992	98.5298	93.9856	98.1823	98.5031

of informativity relations among values. This 'data behaviour' was used as the element for pairwise comparisons, and although this approach clusters differently from RAW comparisons, we belive that it will provide a new perspective for CC selection and open the possibility of further data exploration.

Our first approach was to measure the comparisons under different *K* values. We compared the approach of the RAW signals whit the PCA-FFT- treated signals. Results from *Sdata*, *Fdata*, and *Rdata* are presented in Tables 1-3. As expected most evaluation criteria improved as *K* increased.

The use of FFT signals renders better overall scores than use of RAW signal in *Sdata* and *Fdata*; however, this advantage diminishes in *Rdata*. We speculate that this difference can be explained by the mapping procedures used; further research regarding this matter is ecouraged.

Using the CH's rendered *K* values, we used both CH and FFT to generate the CCs is summarized in Table 4 and in Figs 4-5. Both methodologies rendered similar results, yet PCA rendered better results on parameters representing MC distribution; this could be an effect of the selection method's intrinsic redundancy.



a) b) c)

FIGURE 4. First two principal component's distributions of k=11 CC (orange) selected by CH(a), PCA(b) and RAW (c) in Sdata distribution (blue).

a) b)

FIGURE 5. First two principal component's dristibutions of k=84 CC (orange) selected by CH(a) and PCA(b) in Fdata distribution (blue).

To further this concept, we analized maize data with both K-means and FFT implementation, in order to both contrast with a differt approach and test the interfase. The results are presented in Fig 6.

Thus far, the proposed CC selection method and algorithm appear worthy of further exploration. We are aware that two particular fundamental elements require immediate attention. First, a better mapping

TABLE 4. CCs selected from MC Sdata, Fdata and Rdata using PCA signals and Core Hunter compared with respective same data

	Sdata		Fdata		Rdata	
	PCA CH		PCA CH		PCA CH	
K	12		84		156	
ANE	0.2348	0.2314	0.6407	0.6392	0.5942	0.5952
ENE	0.339	0.3906	0.6474	0.6386	0.5981	0.6047
E	0.5562	0.563	0.7304	0.7176	0.7051	0.7017
MD	0	0	0.59	1.1799	3.9062	5.4688
VT	41.6667	58.3333	56.6372	66.6667	58.0729	86.7188
CR	65.6045	76.1001	88.9709	93.0119	83.2957	89.6723
CV	9080.978	132.6078	0.7357	0.429	0.3137	0.4001
AR	74.3363	76.9912	98.5995	98.4803	98.5298	99.3852

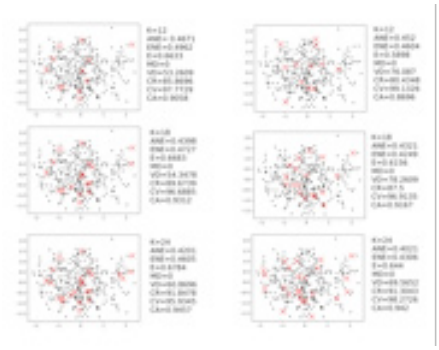


FIGURE 6. First two principal component's distributions of k=12 CC (up), k=18 CC (center) and k=24 CC (bottom); selected by FFT (left) and K-means (right) with their respective evaluation values. Black dots correspond to the complete maize set, while red X represent selected elements for CC.

solution for both genotypic and AT numerical representation needs to be determined. Second, the selection system developed by us is directly based on the DM and is prone to high redundancy in some data distributions. As discussed earlier, this selection system was chosen in order to avoid the problems associated with hierarchical clustering and further allocation selections [13, 35]. Both issues should be addressed in the near future.

Comprehensive data analysis

To demonstrate thet FFT-based CC selection can include and analyse data regardless of its origin, we concatenated corresponding signals from FdataI whit FdataII as well as RdataI and RdaraIII with RdataII to construct Mfdata, MRdataI and MRdataIII. The comprehensive sts were used to construct CCs; the sets were then compared with both their orginal genotype and phenotype MCs. These comparisons are shown in Tables 5-8, and their distributions are represented in Fig. 7-10.

These comprehensive CCs showed overall better scores than genotypic-only CCs when compared with genotypic-only data. On the contrary, there was a better overall score in phenotypic-only CCs when compared against phenotypic-only data.

In the latter case, it should be kept in mind that comprehensive data also consider genotypic data; this could explain why better selections are made when only phenotypic data are considered because genotypic variations may reduce the impact of some phenotypic traits in the PCA analysis.

The generation of a DM based on signal comparisons originating from mixed data construction enables us to explore one of the most interesting applications of this algorithm. By mapping genotypic and AT data, constructing a single signal with all data avaliable for a particular accession is possible. The possibility of including genotypic data with phenotypic traits, geographical locations, climates, habitats, nutritional requirements, symbiotic relationships and so forth provides an opportunity for determining the best information to be included in the selection process in order to cope with the particular objectives for which that CC is beging selected. This concept, in addition to adequate scoring systems, may prove useful in designing tailored CCs that comply with specific research/breeding objetive.

TABLE 5. CCs selected from MC FdataI and MC MFdata PCA signals and evaluated with FdataI and FdataII

	vs EdataI		vs EdataII	
	FdataI	MFdata	FdataII	MFdata
K	24			
ANE	0.6333	0.6356	0.4049	0.4093
ENE	0.6413	0.6423	0.4374	0.4351
E	0.7194	0.7113	0.623	0.5914
MD	1.7668	2.4735	0	0
VT	66.0777	33.9223	46.42	64.2857
CR	89.4908	89.8198	80.677	82.1913
CV	45.7033	35.6847	21.8658	132.1517
AR	91.7647	92.7206	97.5904	94.3775

TABLE 6. CCs selected from MC RdataI , MrdataI, RdataIII and MRdataIII PCA signals and evaluated with RdataI

vs RData				
	RdataI	MRdataI	RdataIII	MRdataIII
K	24			
ANE	0.6148	0.6156	0.6251	0.6169
ENE	0.5989	0.6107	0.621	0.6194
E	0.6962	0.6909	0.6985	0.6934
MD	8.8542	8.5938	7.2917	6.7708
VT	52.0833	63.5417	52.0833	53.3854
CR	80.7367	83.768	81.7278	81.8623
CV	56.3949	59.6279	45.6875	199.9377
AR	86.5097	88.144	86.5651	90.7202

TABLE 7. CCs selected from MC RdataI, MRdataI, RdataIII and MRdataIII PCA signals and evaluated with RdataIII

vs RDataIII				
	RdataI	MRdataI	RdataIII	MRdataIII
K	24			
ANE	0.6285	0.6276	0.6314	0.623
ENE	0.6273	0.6294	0.6368	0.6267
E	0.7036	0.7054	0.7226	0.7056
MD	8.0729	7.5521	7.2917	10.4167
VT	52.8646	60.6771	51.5625	46.875
CR	79.5995	81.0356	79.6809	84.53
CV	28.3673	56.3689	90.0475	60.7279
AR	88.9071	88.7705	87.5956	93.0471

TABLE 8. CCs selected from MC FdataI and MC MFdata PCA signals and evaluated with FdataI and FdataII

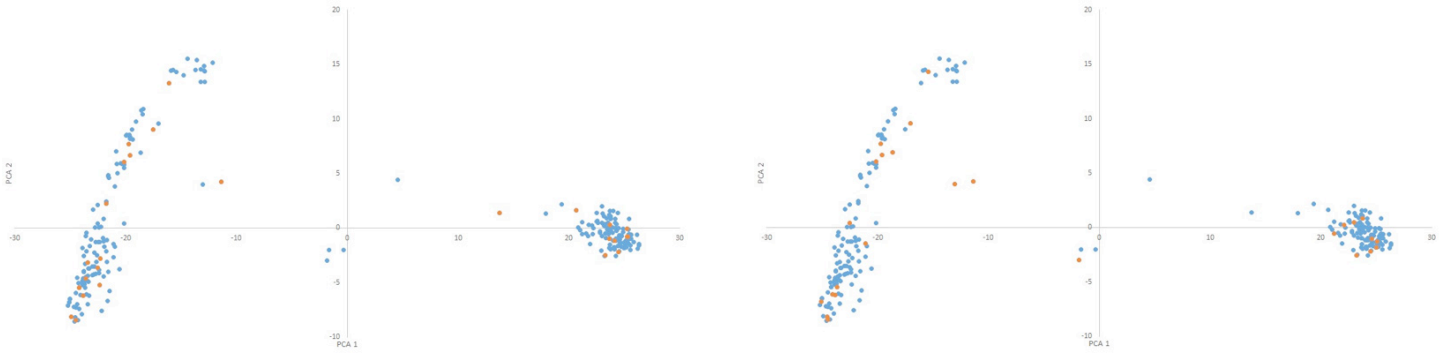
vs RDataII			
	RdataII	MRdataI	MRdataIII
K	24		
ANE	0.4594	0.4652	0.4618
ENE	0.4796	0.4896	0.4742
E	0.6402	0.6205	0.6169
MD	0	5.2632	0
VT	39.4737	42.1053	60.5263
CR	63.8082	61.8988	68.2437
CV	3.8262	2.2285	4.1332
AR	95.4268	98.7805	98.7805

a) b)

FIGURE 7. First two principal component's distributions of k=24 CC (orange) selected by PCA from Fdata(a) and Mdata(b) in FdataI distribution (blue).

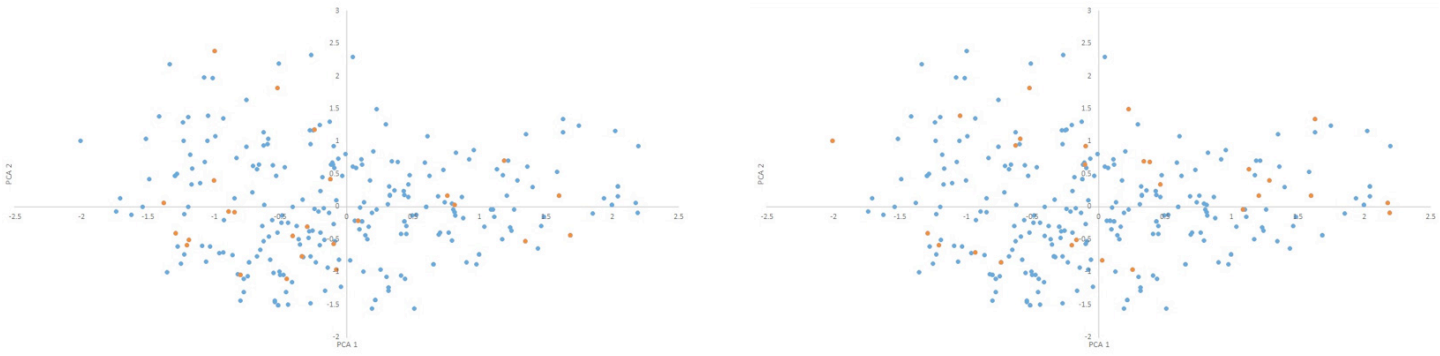
a) b)

FIGURE 8. First two principal component's distributions of k=24 CC (orange) selected by PCA from FdataII(a) and Mdata(b) in FdataII.



a) b)

FIGURE 9. First two principal component's distributions of k=24 CC (orange) selected by PCA from RdataIII (a) and MdataIII (b) in RdataI distribution (blue).



a) b)

FIGURE 10. First two principal component's distributions of k=24 CC (orange) selected by PCA from RdataII (a) and MRDataI (b) in RdataII distribution (blue).

CONCLUSIONS

The use of SPTs in CC selection, as presented in this algorithm, enables us to analyse all avaliable data comprehensively and from different perspectives. Despite its limitations, this signal construction make it possible to analyse all avaliable data regarding each accesion in CC selection with good results.

The efficiency of SPTs in CC selection sugests that the use of these tools in MC analysis may provide useful information not only for CC but also for other purposes.

The implementation of current and other SPTs in all-inclusive MC-mapped signals is worth further exploration, and we belive that it will be an important asset to genentic resource management and exploitation.

AUTHOR CONTRIBUTIONS

ILF performed the implementation, helped with the analysis and manuscript drafting. MT contributed to the design of the algorithm, data analysis and manuscript drafting and correction. EB conceived and designed the algorithm, performed the implementation, analysed the data and wrote the manuscript. All authors have read and approved the final manuscript.

COMPETING INTERESTS

No competing interests were disclosed.

GRANT INFORMATION

This research is supported in part by the SATREPS project by JST and JICA, Diversity Assessment and Development of Sustainable Use of Mexican Genetic Resources and in part by JSPS Grant-in-Aid 25257416.

REFERENCIAS

[1] Biotechnology, P., & Watanabe, K. N. (1999). Plant Genetic Resources and its Global, 7-13.

[2] Dulloo M, Hunter D, Borelli T. Ex situ and in situ conservation of agricultural biodiversity: major advances and research needs. Not Bot Horti ... [Internet]. 2010 [cited 2014 Nov 19];38(2):123-35. Available from: *http://notulaebotanicae.ro/index.php/nbha/article/view/4878*

[3] Upadhyaya H, Gowda C, Sastry D. Plant genetic resources management: collection, characterization, conservation and utilization. J SAT Agric ... [Internet]. 2008 [cited 2014 Nov 19];6(December):1-16. https://doi.org/10.1186/1471-2229-8-106

[4] Brown, A. H. D. (1989). Core collections: a practical approach to genetic resources management. Genome, 31(2), 818-824. https://doi.org/10.1139/g89-144

[5] Guo Y, Li Y, Hong H, Qiu L-J. Establishment of the integrated applied core collection and its comparison with mini core collection in soybean (Glycine max). Crop J [Internet]. 2014 Feb [cited 2014 Jun 6];2(1):38-45. Available from: *http://linkinghub.elsevier.com/retrieve/pii/S2214514113000366*

[6] Studnicki M, MADRY W, Schmidt J. Efficiency of Sampling Strategies to Establish a Representative in the Phenotypic-based Genetic Diversity Core Collection of Orchardgrass (Dactylis glomerata. Czech J Genet Plant Breed [Internet]. 2013 [cited 2014 Jul 10];2013(1):36-47. Retrieved from: *https://goo.gl/vfGhYu*

[7] Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. L. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theoretical and Applied Genetics, 126(2), 289-305. https://doi.org/10.1007/s00122-012-1971-y

[8] Richards C, Volk G. Selection of stratified core sets representing wild apple (Malus sieversii). J Am Soc Hortic Sci [Internet]. 2009 [cited 2014 Jul 31];134(2):228-35. Retrieved from: *http://journal.ashspublications.org/content/134/2/228.short*

[9] Franco J, Crossa J, Warburton ML, Taba S. Sampling Strategies for Conserving Maize Diversity When Forming Core Subsets Using Genetic Markers. Crop Sci [Internet]. 2006 [cited 2014 Jun 19];46(2):854. https://doi.org/10.2135/cropsci2005.07-0201

[10] Hu J, Zhu J, Xu HM. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. TAG Theor Appl Genet [Internet]. 2000 Jul 12;101(1-2):264-8. https://doi.org/10.1007/s001220051478

[11] Wang J, Hu J, Huang X, Xu S. Assessment of different genetic distances in constructing cotton core subset by genotypic values. J Zhejiang Univ Sci B [Internet]. 2008 May [cited 2014 Jul 1];9(5):356-62. Available from: *http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2367373&tool=pmcentrez&rendertype=abstract*

[12] Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, Davenport GF. Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics [Internet]. 2009 Jan [cited 2014 Jun 19];10:243. https://doi.org/10.1186/1471-2105-10-243

[13] De Beukelaer H, Smýkal P, Davenport GF, Fack V. Core Hunter II: fast core subset selection based on multiple genetic diversity measures using Mixed Replica search. BMC Bioinformatics [Internet]. 2012 Jan;13:312. https://doi.org/10.1186/1471-2105-13-31

[14] Jansen J, van Hintum T. Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. Theor Appl Genet [Internet]. 2007 Mar [cited 2014 May 29];114(3):421-8. https://doi.org/10.1007/s00122-012-1971-y

[15] Gouesnard B, Bataillon T. MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J ... [Internet]. 2001 [cited 2014 Jul 1];93-4. Retrieved from *http://jhered.oxfordjournals.org/content/92/1/93.short*

[16] Franco J, Crossa J, Ribaut J. A method for combining molecular markers and phenotypic attributes for classifying plant genotypes. Theor Appl Genet [Internet]. 2001 [cited 2014 Aug 18];103:944-52. Retrieved from *https://goo.gl/H84Nc8*

[17] Kwan HK, Arniker SB. Numerical representation of DNA sequences. Proc 2009 IEEE Int Conf Electro/Information Technol EIT 2009. 2009;307-10. https://doi.org/10.1109/EIT.2009.5189632

[18] Dossou-aminon I, Loko LY, Adjatin A, Ewédjè EBK, Dansi A, Rakshit S, et al. Genetic Divergence in Northern Benin Sorghum (Sorghum bicolor L . Moench) Landraces as Revealed by Agromorphological Traits and Selection of Candidate Genotypes. Sci World J. 2015;2015:e916476. https://doi.org/10.1155/2015/916476

[19] Stein E, Weiss G. The Fourier Transform. In: Introduction to Fourier analysis on Euclidean Spaces. Princeton University Press; 1971.

[20] Cooley J, Tukey J. An algorithm for the machine calculation of complex Fourier series. Math Comput [Internet]. 1965 [cited 2012 Nov 10];297-301. https://doi.org/10.2307/2003354

[21] Nagarajan N, Keich U. FAST: Fourier transform based algorithms for significance testing of ungapped multiple alignments. Bioinformatics [Internet]. 2008 Feb [cited 2011 Sep 17];24(4):577-8. https://doi.org/10.1093/bioinformatics/btm594

[22] Borrayo, E., Mendizabal-Ruiz, E. G., Velez-Perez, H., Romo-Vazquez, R., Mendizabal, A. P., & Alejandro Morales, J. (2014). Genomic signal processing methods for computation of alignment-free distances from dna sequences. PLoS ONE, 9(11), 1-13. https://doi.org/10.1371/journal.pone.0110954

[23] Wang J, Guan Y, Wang Y, Zhu L, Wang Q, Hu Q, et al. A strategy for finding the optimal scale of plant core collection based on Monte Carlo simulation. ScientificWorldJournal [Internet]. 2014 Jan;2014:503473. Available from: *https://goo.gl/gA8LsF*

[24] Hirano R, Naito K, Fukunaga K. Genetic structure of landraces in foxtail millet (Setaria italica (L.) P. Beauv.) revealed with transposon display and interpretation to crop evolution of foxtail millet. ... [Internet]. 2011 [cited 2014 Jul 10];506:498-506. https://doi.org/10.1139/g11-015

[25] Borrayo E, Machida-Hirano R, Takeya M, Kawase M, Watanabe K. Principal components analysis - K-means transposon element based foxtail millet core collection selection method. BMC Genet [Internet]. 2016 Dec 16;17(1):42. https://doi.org/10.1186/s12863-016-0343-z

[26] Odong TL, van Heerwaarden J, Jansen J, van Hintum TJL, van Eeuwijk F a. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet [Internet]. 2011 Jul [cited 2014 Jan 21];123(2):195-205. https://doi.org/10.2135/cropsci2011.02.0095

[27] Cericola F, Portis E, Toppino L, Barchi L, Acciarri N, Ciriaci T, et al. The population structure and diversity of eggplant from Asia and the Mediterranean Basin. PLoS One [Internet]. 2013 Jan [cited 2014 Jun 28];8(9):e73702. https://doi.org/10.1371/journal.pone.0073702 ¿?

[28] Mei Y, Zhou J, Xu H, Zhu S. Development of sea island cotton ('Gossypium barbadense'L.) Core collection using genotypic values. Aust J Crop Sicence [Internet]. 2012 [cited 2014 Jul 3];6(4):673-80. Retrieved from: *http://search.informit.com.au/documentSummary;dn=362661761803357;res=IELHSS*

[29] Redelings BD, Suchard MA. Joint Bayesian estimation of alignment and phylogeny. Syst Biol [Internet]. 2005;54(3):401. https://doi.org/10.1080/10635150590947041

[30] Cristea, P. D. (2002). Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine, 6(2), 279-303. https://doi.org/10.1111/j.1582-4934.2002.tb00196.x

[31] Rosen GL, Sokhansanj B a, Polikar R, Bruns MA, Russell J, Garbarine E, et al. Signal processing for metagenomics: extracting information from the soup. Curr Genomics [Internet]. 2009 Nov [cited 2012 Mar 29];10(7):493-510. https://doi.org/10.2174/138920209789208255

[32] Wang LWL, Schonfeld D. Mapping Equivalence for Symbolic Sequences: Theory and Applications. IEEE Trans Signal Process [Internet]. 2009;57(12):4895-905. https://doi.org/10.1109/TSP.2009.2026544

[33] Almeida JS, Vinga S. Universal sequence map (USM) of arbitrary discrete sequences. BMC Bioinformatics [Internet]. 2002;3(1):6. Retrieved from: *https://www.dropbox.com/s/0w3fllpy9ln250g/Almeida_etal_2002_JZSER_LMO.pdf*

[34] Akhtar M, Epps J, Ambikairajah E. Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction. IEEE J Sel Top Signal Process [Internet]. 2008 Jun;2(3):310-21. https://doi.org/10.1109/JSTSP.2008.923854

[35] Franco J, Crossa J, Taba S, Shands H. A Sampling Strategy for Conserving Genetic Diversity when Forming Core Subsets. Crop Sci [Internet]. 2005 [cited 2014 Dec 4];45(3):1035. https://doi.org/10.2135/cropsci2004.0292

ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-10

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.8

E-LOCATION ID: **e201803EE1**

Mathematical Modeling of the Quorum Sensing in *Vibrio harveyi*

Modelado Matemático de la Detección de Quórum en *Vibrio harveyi*

C. E. Torres-Cerna[1], *E. A. Hernández-Vargas*[2]

[1]Universidad de Guadalajara

[2]Frankfurt Institute for Advanced Studies

ABSTRACT

One of the most used bacteria in the Quorum Sensing (QS) experimental works is the *Vibrio harveyi*, which is used as reporter bacteria to detect the Autoinducers-2 (AI-2) activity of other bacteria. Nevertheless, the description of its QS mechanism by the mathematical modeling is an approach still unexploited. For biological systems, it is necessary to consider the high variability of the experimental data, thus identifiability and parametric reliability analyses must be performed before a model could be used. The following work describes a methodology for parameter fitting and parametric identifiability analysis in a model that describes the dynamics of AI-2 in *V. harveyi* bacteria. Identifiability analyses showed that all parameters are identifiable, but parametric dependency analyses showed two linearly dependent parameters. According to our results, the model is adequate to describe the AI-2 dynamics in *V. harveyi*.

KEYWORDS: Mathematical modeling; Vibrio harveyi; AI-2; Parameter estimation; Parameter dependency

RESUMEN

Una de las bacterias más utilizadas en los trabajos experimentales de detección de quorum (QS) es la *Vibrio harveyi*, que se utiliza como bacteria reportera para detectar la actividad de Autoinductores-2 (AI-2) de otras bacterias. Sin embargo, la descripción de su mecanismo de QS por medio del modelado matemático es un enfoque aún no explotado. En el caso de los sistemas biológicos, es necesario considerar la alta variabilidad de los datos experimentales, por lo que deben realizarse análisis de identificabilidad y fiabilidad paramétrica antes de que un modelo pueda ser usado. El siguiente trabajo describe una metodología para el ajuste de parámetros y el análisis de la identificabilidad paramétrica en un modelo que describe la dinámica de la AI-2 en las bacterias *V. harveyi*. Los análisis de identificabilidad mostraron que todos los parámetros son identificables, pero los análisis de dependencia paramétrica mostraron dos parámetros linealmente dependientes. De acuerdo con los resultados, el modelo es adecuado para describir la dinámica AI-2 en *V. harveyi*.

PALABRAS CLAVE: Modelado matemático; Vibrio harveyi; AI-2; Estimación de parámetros; dependencia paramétrica

Correspondencia

DESTINATARIO: **Esteban Abelardo Hernández Vargas**
INSTITUCIÓN: **Frankfurt Institute for Advanced Studies**
DIRECCIÓN: **Ruth-Moufang-Straße 1, 60438 Frankfurt am Main, Germany**
CORREO ELECTRÓNICO: **vargas@fias.uni-frankfurt.de**

Fecha de recepción:

31 de agosto de 2018

Fecha de aceptación:

10 de enero de 2019

INTRODUCTION

Quorum Sensing (QS) is a mechanism of bacterial gene regulation used to coordinate collective behaviors in a population [1]. Among the known bacteria that use the QS mechanism, the *Vibrio harveyi* is one of the most versatile, mainly because uses the Autoinducer-2 (AI-2) as a signaling molecule which is known as an interspecies signaling molecule [2]. Despite many models have been proposed to describe the QS mechanism [3], just a few of these models are focused on QS mechanism that uses AI-2 as a signaling molecule [4, 5].

Mathematical modeling has become an important tool in many sciences, mainly because of its capability to describe different aspect and relations between the elements of a system. Normally, mathematical models contain a set of parameters which either can be inferred from experimental data, or need to be estimated, and before a mathematical model can be considered reliable, the unknown parameters need to be estimated [6].

When experimental data from the real system is available, the model parameters can be estimated by minimizing a cost function which measures the error between the experimental data and model outcome. Nevertheless, because the quality or quantity of experimental data, the model parameters can present estimation problems, like non-identifiability or parameter uncertainty. These problems are very common in mathematical models that describe biological systems, due to their stochastic nature and the noise added by the experimental measurements [3].

Some methodologies have been proposed and successfully used to tackle these problems. Raue et al. present a methodology to identify the non-identifiability based on the likelihood profile, this approach can determinate the practical and structural non-identifiability [7]. Additionally, Xue *et al*. present a methodology to determinate the parameter uncertainty if the experimental data distribution is unknown [8]. These approaches were satisfactory used in [9–11], where were applied in parameter estimation of mathematical models which describe biological systems. In both models, parameters practically non-identifiable were identified and fixed for further estimations, which enhanced the parametric estimation.

Here, we propose a mathematical model that describes the production and uptake of Autoinducer-2 (AI-2) in bacteria *V. harveyi*. In our model, the parameters are identifiable but exist a parametric dependency. We found that fixing a dependent parameter reduces the confidence interval of the remaining parameters, enhancing the parameter identifiability. Based on the estimation results, the model can be useful to describe the AI-2 dynamics of *V. harveyi*. Unlike other QS mathematical models, ours describes the AI-2 dynamics as a function dependent on the bacterial growth, which could offer a new approach to develop control mechanism based on the bacterial growth media.

METHODOLOGY

Quorum Sensing in *Vibrio harveyi*

The QS mechanism of *V. harveyi* has been well characterized and a brief description is presented in Figure 1. Briefly, *V. harveyi* uses three different signaling molecules, CAI-1, HAI-1, and AI-2, produced by the CqsA, LuxM, and LuxS proteins, respectively. These molecules freely cross the cell membrane and accumulate in the extracellular space till reach a threshold and are sensed by membrane proteins. Each signaling molecule has a cognate membrane protein, CqsS for CAI-1, LuxN for HAI-1, and LuxP-LuxQ for AI-2. Once sensed, the membrane proteins reduce the phosphorylation activity over the LuxU, and this, in turn, reduces the LuxO phosphorylation. This reduction activates the LuxR protein expression which represses and activates many genes, like genes related to the bioluminescence and biofilm formation [12–14].

Mathematical model

In the *V. harveyi* QS mechanism three different AIs are produced and detected, nevertheless, to simplify our model, we only considered the AI-2 dynamic. Our model was made based on the next assumptions *i)* the AI-2 production by the LuxS synthase is dependent on cell growth [15]; *ii)* it is considered that all produced AI-2 freely cross the membrane to the extracellular space.

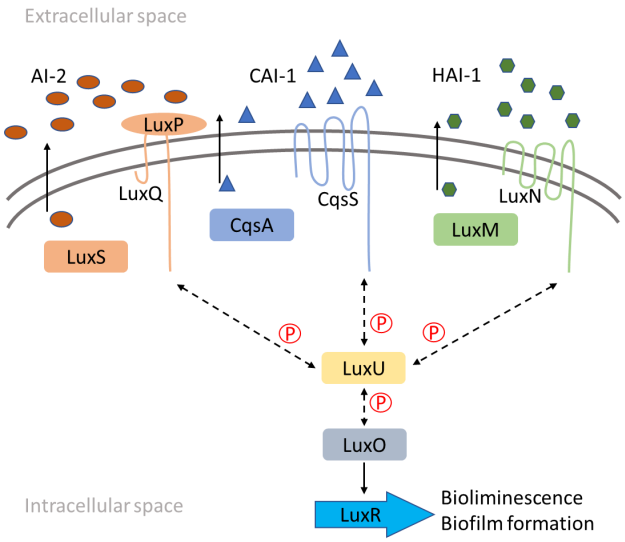


FIGURE 1. The Quorum sensing mechanism in *V. harveyi*.

The *V. harveyi* QS model is composed of the Gompertz function to adjust the *V. harveyi* growth curve (X), and the extracellular AI-2 concentration (A). The model is described as follows:

$$X(t) = X_0 + C\, e^{-e^{-B(t-M)}} \tag{1}$$

$$\frac{dA}{dt} = \mu_A - \mu_{XA} \tag{2}$$

The bacterial growth dynamic is described using a Gompertz function in Equation (1), where X_0 is the initial bacterial concentration, C is the asymptote of the function, and represent the maximum bacteria concentration, B is the slope of the function which represents the growth rate, and M is the saturation time.

Equation (2) describes the A dynamics, μ_A is the AI-2 synthesis. μ_{XA} is the uptake of A by the bacteria. μ_A is presented below.

$$\mu_A = k_A \left(\frac{X(t)^{n_1}}{X(t)^{n_1} + k_{m1}^{n_1}} \right) \tag{3}$$

where k_A is the A velocity production, and k_{m1} is an affinity constant. The uptake of A is described by a function μ_{XA} as follows:

$$\mu_{XA} = k_{XA} \left(\frac{X(t)^{n_2}}{X(t)^{n_2} + k_{m2}^{n_2}} \right) A(t) \tag{4}$$

where k_{XA} is the uptake rate by the bacteria, and k_{m2} is an affinity constant.

Parameter estimation

Parametric estimation of the mathematical model can be understood as the search of values for parameters set (θ) that minimize the difference between the model outcome y_i and experimental data y_i as close to zero as possible. This search is restricted by the system dynamics, algebraic restrictions and systems constraints. Focused on this aim, the Sum of Square of Weighted Residues (*SSWR*) has been used successfully in others works as cost function [10, 16, 17], and is defined as follows.

$$SSWR(\theta) = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\frac{y_i^j - \bar{y}_i^j}{\max(y^j)} \right)^2 \tag{5}$$

where j and i represents the number of variables and experimental data, respectively, $\bar{y}$ is the set of experimental data points, and y is the model outcome. Since the integration routine of Equation (2) requires dense data sets at different times depending on adaptive step size, inputs in each estimation are approximated by a linear interpolation. The minimization of Equation (5) implies a non-linear optimization problem with several variables that can be solved using a global optimi-

zation algorithm. In this work, we used the Differential Evolution (DE) algorithm [18] to estimate the optimal parameter values.

Parameter identifiability

A parameter is identifiable if can be determined by a value within a confidence interval with a desired probability. The parameter identifiability plays an important role for analysis of parametric models because the parameters define the model performance and its adaptability under different conditions [7, 19].

In order to analyze the parameter identifiability in Equations (1) and (2), we used the approach based on the profile likelihood presented by Raue *et al.* [7]. This approach offers insights into the parametric identifiability. Additionally, this approach explores the practical and structural non-identifiability, two phenomena related to parameters.

Briefly, the approach consists on defining a set of values for each parameter, centered at its optimized value, and re-optimize the remaining parameters minimizing the *SSWR* [7]. The objective of this approach is to explore the parameter search space around the optimal value of each parameter, while the model outcome is re-optimized estimating the remaining parameters [9].

Parameter uncertainty and dependency

Due to the stochastic nature of biological systems, when a mathematical model that describes a biological phenomenon is developed, is necessary to consider the data variability. Additionally, the measuring methods normally add noise to the measurements, incrementing the data variability. The bootstrap method is a statistical tool to determinate the parameters accuracy and parameters dependency.

For parametric bootstrap is necessary to know the data distribution, which is normally unknown. To tackle this issue, the weighted bootstrap method assigns to the cost function a vector of random weights from an exponential distribution with mean and variance one [8]. This method has been used successfully in others similar works [9, 10]. In each weighted bootstrap repetition, the model parameters are re-optimized, and after enough repetition, the confidence interval is calculated. The 95% confidence interval for each parameter is calculated between the 2.5 and 97.2% quantiles. From the confidence interval, the distributions of parameters and dependency among parameters are plotted.

NUMERICAL RESULTS

Initially, the parameters were estimated to find a set of parameters values that adjust the model outcome to experimental data. The experimental data were taken from [14], using the Plot Digitizer program to obtain the numerical data from the graphics [20]. Because the growth function is independent, firstly we estimate the growth function parameters and used the best fit values as constants for estimations of remaining parameters. The best fit values of growth function parameters are presented in Table 1.

TABLE 1. Parameter values of the growth function. These values are fixed in further estimation.

Parameter (units)	Best fit
X_0 (OD_{600})	0.00091
C (OD_{600})	3.2357
B (t)	0.3514
M (t)	10.7743

The remaining model parameters were estimated using the values in Table 1, and the best fit value set was used to calculate the profile likelihood [7], this method has been successfully used in previous similar works [9, 10]. For each parameter, a vector is defined with values centered at its best-estimated value and use to explore its parameter space. The profile likelihood results can be seen in Figure 2. The graphics

show a concave shape that means there is a parameter value that minimizes the model error, and the model parameters are identifiable.

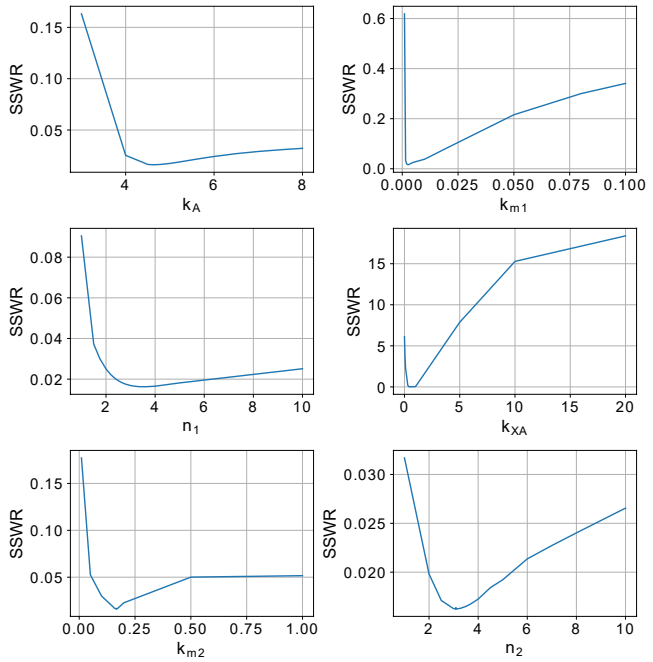


FIGURE 2. **Profile likelihood graphics.**

After identifiability analysis, we perform the weighted bootstrap method to analyze the parameters uncertainty and parameter dependency and compute the confidence interval. Firstly, 500 weighted bootstraps repetitions were made, re-optimizing the parameters on each repetition. Then, the 95% confidence interval was calculated, using the 2.5 and 97.5% quantiles, and the intervals are presented in Table 2.

TABLE 2. **Parameters confidence interval and best fit.**

Parameter (units)	Best Fit	Confidence interval	
		2.5% quantile	97.5% quantile
$k_A\ (\mu M \cdot t^{-1})$	4.6361	4.3507	15.0
$k_{m1}(OD_{600})$	0.0025	0.0021	0.0312
n_1	3.5412	0.8991	4.3129
$k_{XA}\ (t^{-1})$	0.4769	0.4477	2.7818
$k_{m2}\ (OD_{600})$	0.1674	0.0704	1.9998
n_2	3.0957	0.2375	5.8004

Parameter k_A is the parameter with the larger interval confidence. On the other hand, k_{m1} has the smaller interval confidence of all parameters. That means, k_A can variate along a long interval and the model is still suitable, but the smaller interval confidence of k_{m1} means that the model is more sensible to it. Based on the confidence interval, the distribution of parameters is depicted in Figure 3.

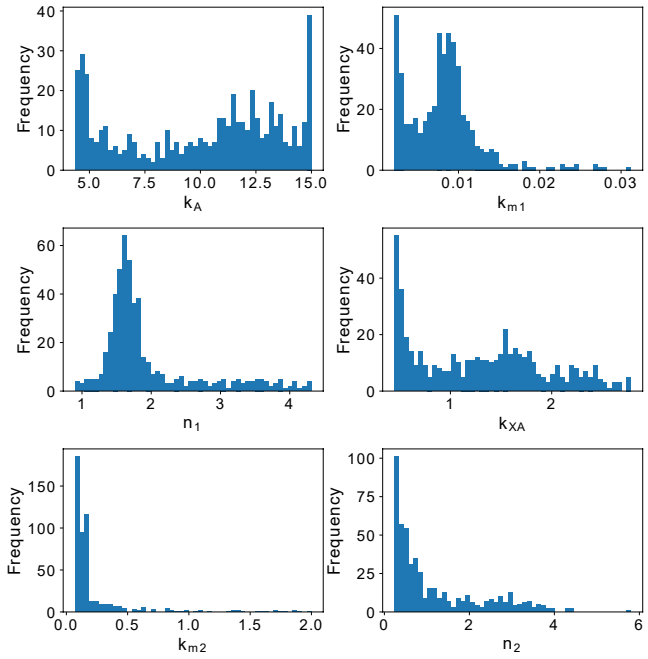


FIGURE 3. **Graphs of the parameter distribution.**

In Figure 4, parameters k_A and k_{XA} show a linear dependency, increasing the value of k_A, the estimation of k_{XA} also increases. These parameters can not be estimated independently. This behavior is consistent with the real system, if the velocity production of AI-2 (k_A) increases, is necessary that the AI-2 uptake rate (k_{XA}) also increases to balance the extracellular AI-2 concentration.

To improve the parameter estimation, one of these parameters must be fixed for further estimations. This approach has been successfully used to reduce the parameter estimation process, which helped to reduce the computational cost and improves the model fit [10].

In this work, they fix the parameters based only on the profile likelihood because they found that some parameters were structurally non-identifiable. In our work, profile likelihood of all parameters showed that all parameters are identifiable, but parametric dependency analysis showed up that some parameters are dependent on each other. Fixing a dependent parameter helps to improve the model fit and to reduce the parameters confidence interval [9].

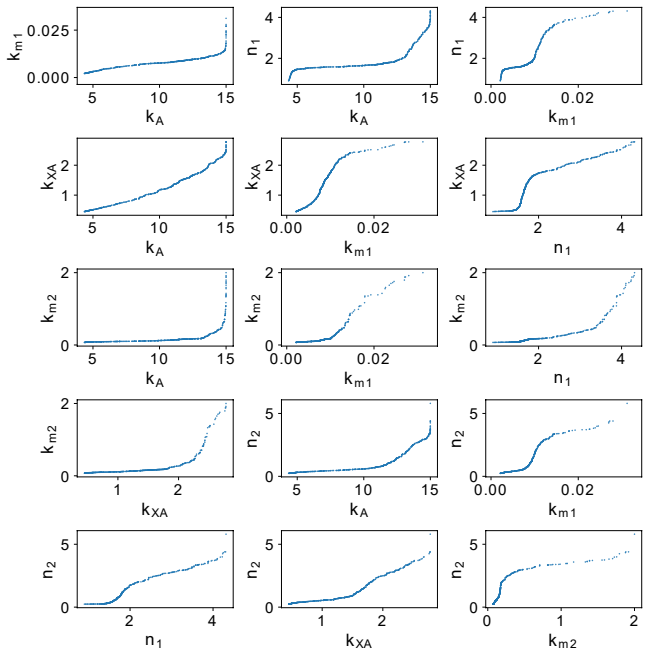


FIGURE 4. Parameter dependency.

Using this approach, we fixed k_{XA}=0.4769 for further estimations and to analyze the remaining model parameters. The selection of k_{XA} as the fixed parameter was to analyze the impact of this approach in a parameter with a large confidence interval. Once k_{XA} fixed the likelihood of remaining is computed, and the graphics are depicted in Figure 5. There is a remarkable improvement in the identifiability of most of parameters. Confidence interval, parameter distributions and dependency among parameters were re-calculated after a new set of 500 weighted bootstraps repetitions. The confidence interval and best fit parameters value are depicted in Table 3.

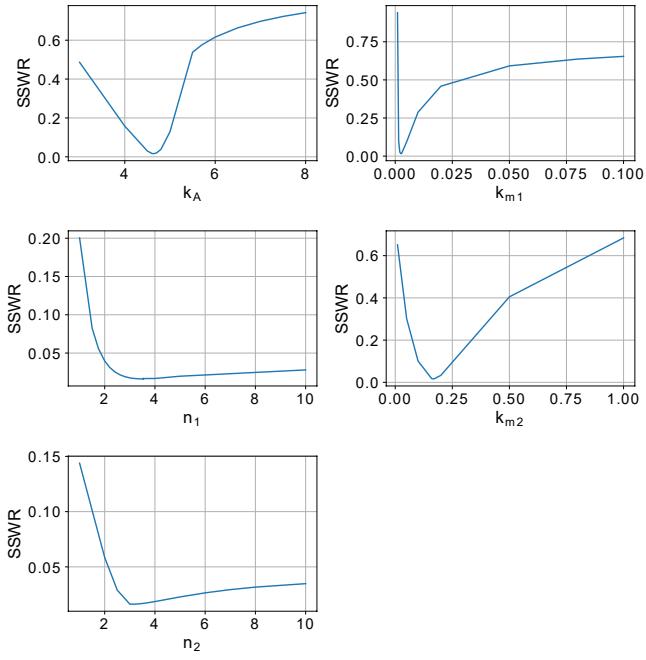


FIGURE 5. Profile likelihood graphics after fixing the parameter k_{XA}.

Remarkably, the best-fit values are the same as that of the previous estimations. Fixing the parameter k_{XA} does not affect the parameter estimation but reduce the computational cost and improves the confidence interval.

Also, the estimated parameters get a distribution more defined, as can be seen in Figure 6, where the parameter k_A varies less when parameter k_{XA} is fixed.

TABLE 3. Parameters confidence interval and best fit value after estimations with k_{XA} fixed. *means that parameter was fixed in estimations.

Parameter (units)	Best Fit	Confidence interval	
		2.5% quantile	97.5% quantile
k_A $(\mu M \cdot t^{-1})$	4.6361	1.8900	4.6649
$k_{m1}(OD_{600})$	0.0025	0.00001	0.0072
n_1	3.5412	0.2023	4.4556
k_{XA} (t^{-1})	0.4769*		
k_{m2} (OD_{600})	0.1674	0.1582	2.5
n_2	3.0957	0.8611	4.1206

The reduction of confidence interval also improved the parameter distribution, which is visually evident in parameters k_A, k_{m1}, and k_{m2} in Figure 6.

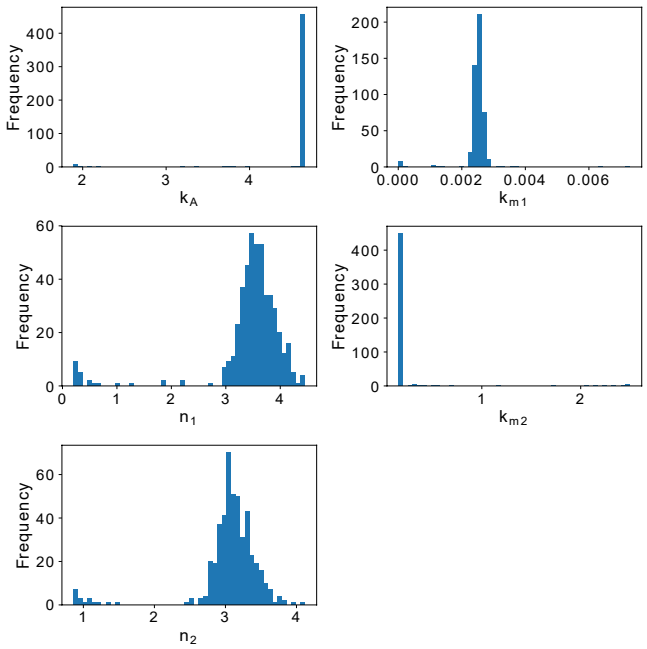


FIGURE 6. Graphs of the parameter distribution after fix parameter k_{XA} in weighted bootstrap repetitions.

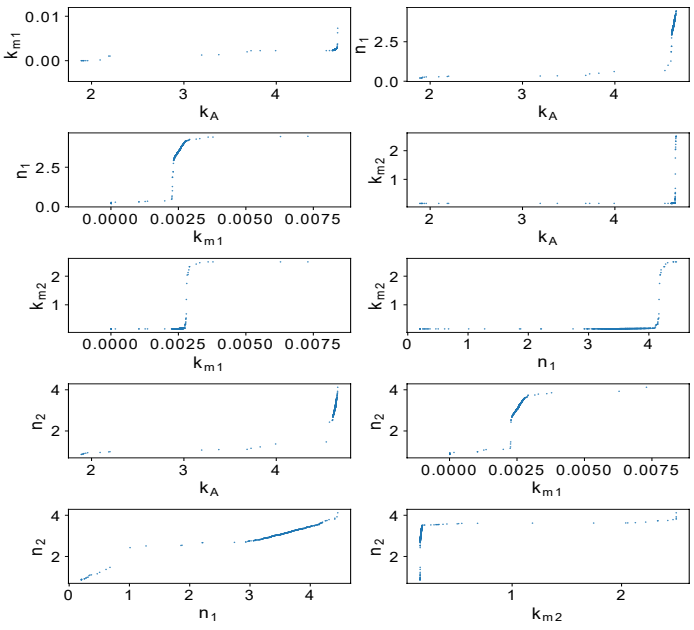


FIGURE 7. Parameter dependency after estimations fixing k_{XA}.

Nevertheless, as in Figure 3 there is one tail distribution in parameters, which could be attributed to the estimation method used is stochastic and its random nature.

The parameter dependency is presented in Figure 7. Parameters n_1 and n_2 present a behavior like parameters k_A and k_{XA}, which seems to be dependent. Nevertheless, based on their distribution (Figure 6) and confidence interval, we considered that the parameter dependency is not significative to affect the model performance or parameters identifiability. After parameter analysis, we consider that by fixing k_{XA} the remaining parameters are identifiable.

The model performance is presented in Figure 8, the model outcome in blue lines, and experimental data in closed circles [14]. Parameters values are presented in Table 1 for Equation (1), and Table 3 for Equation (2). As can be seen, the model presents a good performance to adjust the experimental data, despite there

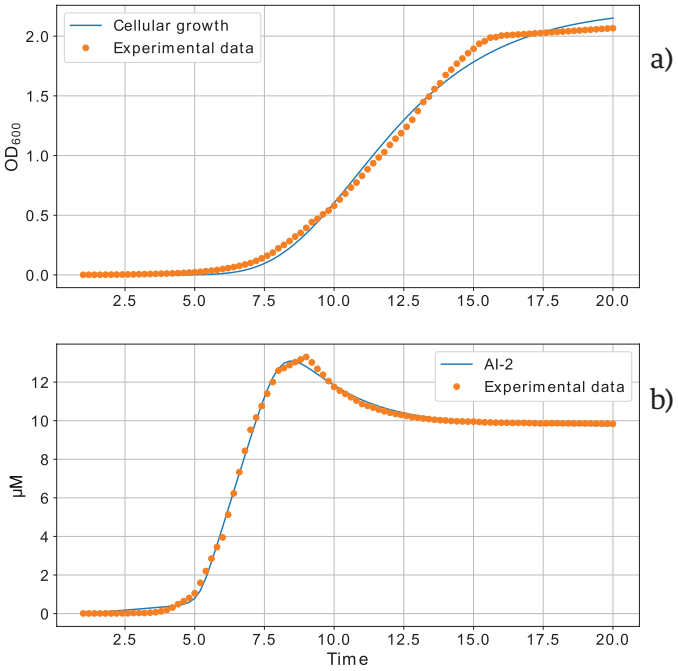


FIGURE 8. Model Outcome. (A) is the cellular growth, and (B) is the AI-2 extracellular concentration.

are misalignments when the experimental data change drastically, about the hour eight for AI-2 dynamics, and hour 15 for the cellular growth.

CONCLUSIONS

In this paper, we presented a mathematical model that describes the AI-2 dynamics as a function of the bacterial growth in the *V. harveyi* bacteria, and the model viability was probed by the parameter identifiability analysis as can be seen in Figure 8, our model can represent adequately the experimental data from [14].

Despite the identifiability analysis showed that all parameters were identifiable, the parameter dependency analysis showed that k_A and k_{XA} were dependents, additionally, the dependent parameters do not present a clear distribution.

Despite both parameters are identifiable, they can increase or decrease arbitrary without enhance the model adjustment.

Fixing k_{XA}, the identifiability and confidence interval of remaining parameters was improved and the distribution of k_A showed a clear tendency to a centered value. This is a new way to tackle the identifiability problem and enhance the model parameters viability.

As future work, we propose a deeper analysis about the influence of bacterial growth on the AI-2 dynamics. Additionally, further analysis can be realized for a better understanding about the effect of a dependent parameter over the other, this could be useful to controls the parameters tendency, which can mean a saving of resources during the experiments.

REFERENCIAS

[1] Waters CM, Bassler BL. Quorum Sensing : Communication in Bacteria. Annu Rev Cell Dev Biol. 2005;21(1):319-46. DOI 10.1137/090757009

[2] Pereira CS, Thompson JA, Xavier KB. AI-2-mediated signalling in bacteria. FEMS Microbiol Rev. 2013 Mar;37(2):156-81. DOI 10.1111/j.1574-6976.2012.00345.x

[3] Pérez-Velázquez J, Gölgeli M, García-Contreras R. Mathematical Modelling of Bacterial Quorum Sensing: A Review. Bull Math Biol. 2016 Aug 25;78(8):1585-639. DOI 10.1007/s11538-016-0160-6

[4] Drees B, Reiger M, Jung K, Bischofs IB. A Modular View of the Diversity of Cell-Density-Encoding Schemes in Bacterial Quorum-Sensing Systems. Biophys J. 2014 Jul 1;107(1):266-77. DOI 10.1016/J.BPJ.2014.05.031

[5] Bressloff PC. Ultrasensitivity and noise amplification in a model of V. harveyi quorum sensing. Phys Rev E. 2016 Jun 28 ;93(6):062418. DOI 10.1103/PhysRevE.93.062418

[6] Fu L, Li P. The Research Survey of System Identification Method. 2013 5th Int Conf Intell Human-Machine Syst Cybern. 2013;2:397-401. DOI 10.1109/IHMSC.2013.242

[7] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmuller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009;25(15):1923-9. DOI 10.1093/bioinformatics/btp358

[8] Xue H, Miao H, Wu H. Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. Ann Stat. 2010 Aug ;38(4):2351-87. DOI 10.1214/09-AOS784

[9] Nguyen VK, Binder SC, Boianelli A, Meyer-Hermann M, Hernandez-Vargas E a. Ebola virus infection modeling and identifiability problems. Front Microbiol. 2015 Apr 9;6(April):1-11. DOI /10.3389/fmicb.2015.00257

[10] Torres-Cerna CE, Alanis AY, Poblete-Castro I, Hernandez-Vargas EA. Batch Cultivation Model for Biopolymer Production. Chem Biochem Eng Q. 2017 Apr 15 [cited 2017 Apr 24];31(1):89-99. DOI 10.15255/CABEQ.2016.952

[11] Nguyen VK, Klawonn F, Mikolajczyk R, Hernandez-Vargas EA. Analysis of Practical Identifiability of a Viral Infection Model. PLoS One. 2016;e0167568. DOI 10.1371/journal.pone.0167568

[12] Waters CM, Bassler BL. The Vibrio harveyi quorum-sensing system uses shared regulatory components to discriminate between multiple autoinducers. Genes Dev. 2006 Oct 1 ;20(19):2754-67. DOI 10.1101/gad.1466506

[13] Rutherford ST, van Kessel JC, Shao Y, Bassler BL. AphA and LuxR/HapR reciprocally control quorum sensing in vibrios. Genes Dev. 2011 Feb 15 [cited 2018 May 31];25(4):397-408. DOI 10.1101/gad.2015011

[14] Anetzberger C, Reiger M, Fekete A, Schell U, Stambrau N, Plener L, et al. Autoinducers Act as Biological Timers in Vibrio harveyi. Misra R, editor. PLoS One. 2012 Oct 26 ;7(10): e48310. DOI 10.1371/journal.pone.0048310

[15] Xavier KB, Bassler BL. LuxS quorum sensing: more than just a numbers game. Curr Opin Microbiol. 2003 Apr;6(2):191-7. DOI 10.1016/S1369-5274(03)00028-6

[16] Patwardhan PR, Srivastava a. K. Model-based fed-batch cultivation of R. eutropha for enhanced biopolymer production. Biochem Eng J. 2004;20(1):21-8. DOI 10.1016/j.bej.2004.04.001

[17] Khanna S, Srivastava AK. Optimization of nutrient feed concentration and addition time for production of poly(β-hydroxybutyrate). Enzyme Microb Technol. 2006 Sep;39(5):1145-51. DOI 10.1016/j.enzmictec.2006.02.023

[18] Storn R, Price K. Differential Evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces. J Glob Optim. 1997;11(4):341-59. DOI 10.1023/A:1008202821328

[19] Miao H, Xia X, Perelson AS, Wu H. On Identifiability of Nonlinear ODE Models and Applications in Viral Dynamics. SIAM Rev. 2011 Jan;53(1):3-39. DOI 10.1137/090757009

[20] Huwaldt JA, Steinhorst S. Plot Digitizer 2.6.2. *http://plotdigitizer.sourceforge.net* (1 March 2015)

ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-7

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.9

E-LOCATION ID: **e201808EE1**

Breve Descripción de la Biología Sintética y la Importancia de su Relación con otras Disciplinas

Brief description of Synthetic Biology and the importance of its relationship with other disciplines

L. A. Muñoz-Miranda[1], I. Higuera-Ciapara[1], A. C. Gschaedler-Mathis[1], L. C. Rodríguez-Zapata[2], A. Pereira-Santana[1], L. J. Figueroa-Yáñez[1]

[1]Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco (CIATEJ)

[2]Centro de Investigación Científica de Yucatán, A. C. (CiCY)

RESUMEN

La biología sintética (SynBio) es una disciplina de reciente aparición que sirve para diseñar o re-diseñar sistemas biológicos y otorgarles cualidades mejoradas o nuevas cualidades. En la SynBio el diseño de nuevos sistemas biológicos requiere de herramientas moleculares muy precisas, tales como: a) la bioinformática, b) la secuenciación NGS (Next Generation Sequencing), el ensamble y/o síntesis de ADN c) y la edición de genomas a través de CRISPR-Cas9. En la SynBio encontramos además otras disciplinas con un perfil más hacia el ámbito social, las cuales tocan aspectos éticos, legales, filosóficos y económicos, considerándose así una multidisciplina. La SynBio está propiciando el desarrollo de nuevas tecnologías (emergentes) partiendo de una óptica ingenieril. En la SynBio, al ADN se le entiende de forma práctica y abstracta como una serie de partes que se pueden ensamblar en cierto orden para obtener los productos deseados una vez que se conoce la funcionalidad de cada parte. La SynBio ha dado pie a una nueva concepción de la economía a nivel mundial y por consecuencia se ha tomado muy seriamente el termino Bioeconomía como una nueva disciplina que transformará a las sociedades.

PALABRAS CLAVE: Biología Sintética; Edición de genomas; Bioeconomía

ABSTRACT

Synthetic biology (SynBio) it is considered a very recent discipline. View as a tool serves to design or re-design biological systems, giving them improved qualities or new qualities. In the SynBio, the design of new biological systems requires very precise molecular tools, such as: a) bioinformatics, b) sequencing NGS (Next Generation Sequencing), assembly and synthesis of DNA c) and CRISPR-Cas9 genome editing. Within the SynBio there are other social profile disciplines which concerned to ethical, legal, philosophical, and economic, and for that reason it is considered a multidiscipline. The SynBio is promoting the development of new (emerging) technologies based on an engineering perspective. In SynBio, DNA is understood in a practical and abstract way as a series of parts that can be assembled in a certain order to obtain the desired products once the functionality of each part is known. The SynBio has given rise to a new conception of the economy worldwide and consequently the term Bioeconomy is already taken very seriously as a new discipline that will transform societies.

KEYWORDS: Synthetic Biology; Genome editing; Bioeconomy.

Correspondencia

DESTINATARIO: **Luis Joel Figueroa Yáñez**

INSTITUCIÓN: **Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco (CIATEJ)**

DIRECCIÓN: **Av. Normalistas #800, Colinas de La Normal, C, P. 44270, Guadalajara, Jalisco, México**

CORREO ELECTRÓNICO: **lfigueroa@ciatej.mx**

Fecha de recepción:

29 de septiembre de 2018

Fecha de aceptación:

11 de enero de 2019

INTRODUCCIÓN

La conformación de toda sociedad humana es moldeada por un adecuado suministro de productos que suplan las necesidades básicas para su desarrollo presente y futuro. La demanda de dichas necesidades básicas, que van desde medicamentos, alimentos, materias primas, entre otros, han ido en constante aumento debido a la creciente población humana. Ante todo esto, la Biología Sintética (SynBio) ofrece nuevas alternativas sustentables para la adquisición de productos que suplan dichas necesidades. La SynBio es una disciplina que se ha desarrollado de manera gradual a partir del siglo pasado y para entenderla como concepto es útil situarle dentro de varios periodos históricos que involucran hitos importantes en el impulso del conocimiento científico [1]. Dentro de los acontecimientos científicos más relevantes que le han dado a la SynBio una presencia central en esta revolución bio-industrial se encuentran los siguientes: el descubrimiento y los avances logrados en la regulación génica (e.g. el operón LacZ); el perfeccionamiento de las técnicas de ADN recombinante y la ingeniería genética; el lanzamiento del proyecto genoma; la secuenciación Sanger; el perfeccionamiento de la técnica de la Reacción en Cadena de la Polimerasa (PCR por sus sigla en inglés); la generación de bases de datos mundiales tales como NCBI, GeneBank y EMBL-net; la optimización de nuevas tecnologías de Secuenciación de Nueva Generación (NGS´s) para la decodificación del ADN; el ensamble Gibson; el uso de la herramienta de edición de genomas CRISPR-Cas9, los modelos computacionales de célula completa; así como la reciente creación de la bacteria sintética *Micoplasma mycoides* JCVI-syn 3.0 [1–3].

Definición del concepto Biología Sintética

Para entender el concepto de SynBio es necesario que primero definamos un sistema biológico como el conjunto de partes similares que trabajan en armonía para cumplir alguna función fisiológica determinada. La interacción de sistemas biológicos da lugar a una red de sistemas que, a su vez, pueden organizarse para dar lugar a procesos de mayor complejidad. El estudio de los sistemas biológicos ha dado lugar al desarrollo de nuevas técnicas y metodologías enfocadas al conocimiento de la adaptación, la evolución y la interacción entre organismos. Dichas herramientas han abierto el camino para el estudio de los organismos desde una nueva perspectiva: la generación de nuevos sistemas biológicos no existentes en la naturaleza para la obtención de productos de interés para el hombre.

La SynBio se puede definir entonces como una disciplina útil para diseñar y construir nuevas partes, mecanismos y sistemas biológicos, o re-diseñar sistemas biológicos existentes y otorgarles nuevas y mejores cualidades con un propósito definido. A partir de ella, es posible desarrollar nuevas metodologías para estudiar la funcionalidad de los propios sistemas biológicos. Los sistemas biológicos sintéticos deben reunir características muy específicas para ser considerados como tales, por ejemplo: a) ser computacionalmente predecibles, b) deben ser medibles, c) controlables y d) transformables (adicionar funciones y/o regular funciones existentes) [4, 5].

Dependiendo del nivel en el que se observa un sistema (ADN/ARN, proteínas, metabolitos, interacciones intracelulares y redes regulatorias) es posible conocer el perfil dinámico de un organismo. En conclusión, la SynBio ha evolucionado en una ciencia esencialmente interdisciplinaria que estudia las interacciones de los múltiples componentes y el comportamiento colectivo de una célula u organismo [6].

La SynBio y su relación con otras disciplinas

El diseño de nuevos sistemas biológicos a través de la SynBio involucra la interacción de una gran variedad de disciplinas científicas tales como la química, la biología, la física, la ingeniería, las matemáticas, la esta-

dística, las ciencias computacionales, entre otras más. Por consecuencia, la SynBio se considera un campo necesariamente multidisciplinario el cual presenta un gran potencial tecnológico para el desarrollo de nuevos productos, entre los cuales podríamos destacar: medicamentos a bajos costos, biocombustibles, plásticos biodegradables, la implementación de la terapia génica y/o molecular, por mencionar algunos. Cabe señalar que especialidades aparentemente distantes como la bioseguridad y la bioética forman parte fundamental en la base del conocimiento de esta disciplina, aunque no pertenezcan al campo tecno-científico [4, 7].

En un esfuerzo por hacer asequible la concepción de los principales elementos que integran a la SynBio, Lee y colaboradores [4] proponen tres etapas para la construcción de un sistema biológico sintético: a) la decodificación y análisis de los genomas de los diferentes sistemas biológicos, b) la síntesis de las partes que integran un genoma y el ensamble de cada una de ellas, y por último c) el uso de las herramientas de edición de genomas. Entre los factores esenciales para el impulso de esta disciplina encontramos: las nuevas herramientas bioinformáticas que relacionan, almacenan y procesan grandes bases de datos; el desarrollo del súper-computo; así como el diseño de nuevos plásmidos mejorados que incluyen gARN's y Cas9. Sin embargo la utilización de la herramienta de edición de genomas CRISPR-Cas9 ha sido el detonador en el avance de ella debido al enorme potencial que tiene en la generación del conocimiento básico y aplicado de genes funcionales y su regulación [4, 8].

Disciplinas sociales que impactan directamente a la SynBio

Por la relevancia de los múltiples elementos antes mencionados que conforman a la SynBio, es importante hacer énfasis en otras disciplinas que involucran aspectos éticos, legales, filosóficos y económicos. De hecho, tal es la importancia de estas disciplinas que al día de hoy continua el debate sobre la adjudicación del descubrimiento y las variantes ligadas a la herramienta CRISPR-Cas9 [8]. Un ejemplo de los muchos debates legales alrededor de CRISPR-Cas9 más recientes, es la revocación de una solicitud de patente otorgada inicialmente al Broad Institute del Massachusetts Institute of Technology de la Universidad de Harvard la cual ha sido rechazada por la oficina de patentes europea [9].

Es importante mencionar que los investigadores involucrados en el desarrollo de CRISPR-Cas9 es amplio; sin embargo, de manera ética y otorgando el debido reconocimiento, cabe mencionar algunos de los más influyentes: Yoshizumi Ishino, Francisco Mojica, Guadalupe Juez, Ruud Jansen, Eugene Koonin, Rodolphe Barrangou, Philippe Horvath, Jennifer Doudna, Blake Wiedenheft, Martin Jinek, Emmanuelle Charpentier, Krzysztof Chylinski, Feng Zhang, Karl Deisseroth, Edward Boyden, George Church, entre otros [10].

La SynBio es tema de un gran debate ético y filosófico a nivel mundial debido al impacto en su uso como herramienta. Al día de hoy existen diversas publicaciones referentes al tema ético de su aplicación tales como: Synthetic Biology and Morality de Kaebnick y Murray, Synthetic Biology, Social and Ethical Challenges de Balmer y Martin, Synthetic Biology at the Limits of Science de Nordmann, Life by design: Philosophical perspectives on synthetic biology de Bensaude.

El crecimiento exponencial de la SynBio en la solicitud de patentes

Van Doren y col., [11] realizaron un estudio sobre la relación de la tendencia que han seguido las solicitudes de patente ligadas al tema de biología sintética desde 1990 al 2010 y observaron que: a) existe un aumento en la actividad anual de solicitudes, de acuerdo al desarrollo de la SynBio y con base en sus aplicaciones, b) cuando se observa por área de aplicación, las principales patentes parecen ser más relevantes para los sectores médico, energético e industrial, c)

entre los principales solicitantes se encuentren los EE.UU., Japón y detrás un considerable número de países europeos y d) las universidades así como las empresas son los principales sectores en la solicitud de distintos tipos de patentes. Las solicitudes de patente en el tema de la SynBio han incrementado principalmente en áreas como la biotecnología médica, industrial y energética promovidas especialmente por universidades e industrias.

Percepción de la SynBio en la sociedad como una herramienta biotecnológica

La SynBio como concepto se percibe tanto positiva como negativamente en la sociedad a nivel mundial; sin embargo, es innegable el gran potencial que posee en áreas como la medicina. Posiblemente, es en este campo donde menor resistencia podría sufrir respecto a su uso ya que seguramente obtendrá su total legitimidad debido a lo que esta representa en el área de la terapia génica. En resumen, su uso es de enorme interés en el campo de la salud y la industria en general [12].

En encuestas realizadas en los EE.UU sobre la percepción existente respecto a la SynBio, esta es catalogada como una disciplina peligrosa que tendría que ser sujeto de regulación; sin embargo muchos países alrededor del mundo aún están trabajando sobre los temas regulatorios [13]. Por otro lado, hoy en día se enfrenta la disyuntiva respecto a que, la SynBio debe ser un conocimiento científico abierto al público; en el aspecto ético y legal esto plantea un gran dilema debido al impacto que existe al modificar el genoma de los sistemas biológicos sin el conocimiento científico básico de la función de los genes. Diferentes agrupaciones a nivel mundial han surgido para "democratizar la SynBio", tales como la ya extinta Glowing Plant, así como las agrupaciones aún vigentes como iGEM y Biohackers DIYBIO [14, 15]. En México la legislación acerca del tema aún está en discusión, sin embargo, existe guías internacionales en donde podemos basarnos para el adecuado trabajo experimental relacionado a la SynBio [16].

Alcances y perspectivas de la SynBio

En el 2009 Gibson y col., [17] perteneciente al grupo de Craig Venter, reportaron el diseño, síntesis y ensamble de un genoma completo. Sintetizaron 1.08 Mb del genoma de la bacteria *Mycoplasma mycoides* llamado JCVI-syn1.0 el cual iniciaron a partir de la información de la secuencia genómica digitalizada. Después trasplantaron el genoma químicamente sintetizado dentro de *Mycoplasma capriculum* (célula receptora) la cual fue controlada solo por el cromosoma sintético diseñado de *M. mycoides*. A raíz de estos grandes descubrimientos, la SynBio se ha convertido en un enorme campo de exploración y de oportunidades surgiendo así empresas como Human Longevity Incorporation y Caribou Biosciences, Inc. [18, 19].

En 2015, William C. Campbell, Satoshi Ōmura y Youyou Tu obtuvieron el premio Nobel de fisiología o medicina por haber sintetizado una droga antimalárica al modificar la regulación de la vía del mevalonato mediante la introducción de doce genes de *Artemisia annua* en la levadura *S. cerevisiae*. La SynBio al ser una herramienta molecular tan poderosa y por estar basada en tres grandes pilares de conocimiento: la bioinformática, el ensamble y síntesis de genomas y la edición de genomas, es objeto de atención mundial; a raíz del entendimiento de la funcionalidad de los genes, especialmente en humanos, el tema de la terapia génica, por ejemplo, se convierte en un tema muy relevante [20].

Existen tecnologías y/o herramientas emergentes que se basan en el uso del conocimiento generado por la SynBio, tales como: bioinformática de súper cómputo cuántico; biobricks, para construir circuitos que generan nuevos diseños de plásmidos; ensamble de genes ssODN's; optogenética; silenciamiento, reemplazo y regulación de genes mediante herramientas de edición (ZNF, TALEN, CRISPR-Cas9); nanoingeniería; selex o aptámeros para hacer dispositivos para diagnóstico; dispositivos o simuladores metabólicos (órgano chips); terapia génica; *phantoms*, que simulan propiedades

electromagnéticas de tejidos y órganos; teletransportador biológico para la colonización de planetas y almacenamiento de datos utilizando memoria de ácidos nucleicos (NAM) [21–28].

El surgimiento del término Bioeconomía en la SynBio

Flores-Bueso y cols., [2] publicaron que a raíz del desarrollo de todas estas nuevas herramientas relacionadas a la SynBio, se está revolucionando la industria biotecnológica y abriendo nuevos mercados, conduciendo a una nueva área emergente, denominada Bioeconomía. En su trabajo mencionan que las empresas relacionadas a la SynBio se han expandido a casi 40 países y casi 700 organizaciones, además de haberse fundado casi 530 nuevas agencias relacionadas al tema; por lo tanto, es evidente que está impactando totalmente la demanda creciente relacionada con la alimentación, la salud, la energía, etc., además de ser vista como una herramienta biotecnológica que puede ser útil para aminorar los efectos del cambio climático.

La Biología Sintética tiene un mercado actual valuado en alrededor de 3.9 billones de dólares, y se estima que crezca cerca del 24.4% anual. Para el 2021, se predice que alcance los 11.4 billones de dólares. Si la SynBio toma las riendas de la Bioeconomía, su contribución podría ser un fenómeno poco esperado, pues se cree que puede dar lugar al desarrollo social y de comunidades con base en las iniciativas de integración y aceptación de la industria, gobiernos y mercados.

CONCLUSIONES

La biología sintética continúa diariamente su desarrollo vertiginoso con un futuro muy prometedor. El conocimiento científico logrado en el siglo XX y parte del siglo XXI en materia de biología molecular y genética, ha permitido la conceptualización y aplicabilidad de nuevas tecnologías y disciplinas encaminadas a la salud, industria y energía.

La SynBio es una herramienta con altas expectativas y potencial socioeconómico. Es probable que su campo de acción crezca de manera exponencial en la próxima década con base a los nuevos desarrollos, patentes e industrias que se están creando y desarrollando. El traslado de los beneficios de esta tecnología del laboratorio al campo va de la mano con las regulaciones que aplican a la salud, al ambiente, a la seguridad alimentaria y la gobernabilidad, por mencionar algunas. Sin embargo, la falta de información y la manera de comunicar las ventajas y desventajas de esta tecnología es importante abordarse debido a que mucho del cuestionamiento social proviene de la falta de conocimiento o simplemente se debe a un entendimiento distinto en términos científicos básicos. La bioética como herramienta de certidumbre o de razón será finalmente una disciplina muy poderosa a la hora de responder sobre la pertinencia de la aplicabilidad de organismos diseñados o de la edición de genomas en humanos como parte de una terapia génica. Si la regulación y la aceptación social en la aplicación de estas tecnologías se articula será de mucho beneficio para la sociedad.

REFERENCIAS

[1] Cameron DE, Bashor CJ, Collins JJ. A brief history of synthetic biology. Nat Rev Microbiol 2014; 12: 381-390. DOI: 10.1038/nrmicro3239

[2] Flores Bueso Y, Tangney M. Synthetic Biology in the Driving Seat of the Bioeconomy. Trends Biotechnol 2017; 35: 373-378. DOI: 10.1016/j.tibtech.2017.02.002

[3] Hughes RA, Ellington AD. Synthetic DNA synthesis and assembly: Putting the synthetic in synthetic biology. Cold Spring Harb Perspect Biol; 9. Epub ahead of print 2017. DOI: 10.1101/cshperspect.a023812

[4] Lee BR, Cho S, Song Y, et al. Emerging tools for synthetic genome design. Mol Cells 2013; 35: 359-370. DOI: 10.1007/s10059-013-0127-5

[5] Li, Yinqing, Yun Jiang, He Chen, Weixi Liao, Zhihua Li, Ron Weiss and ZX. Modular construction of mammalian gene circuits using TALE transcriptional repressors. Nat Chem Biol 2015; 11: 207-213. DOI: 10.1038/nchembio.1736

[6] Liu D, Hoynes-O'Connor A, Zhang F. Bridging the gap between systems biology and synthetic biology. Front Microbiol 2013; 4: 1-8. DOI: 10.3389/fmicb.2013.00211

[7] Pretorius IS. Synthetic genome engineering forging new frontiers for wine yeast Synthetic genome engineering forging new frontiers for wine yeast. Crit Rev Biotechnol; 8551. Epub ahead of print 2016. DOI: 10.1080/07388551.2016.1214945

[8] Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science (80-) 2014; 346: 1258096-1258096. DOI: 10.1038/nj0503

[9] Mitchell S. CRISPR-Cas9 - from gene editing to EPO patent law editing?, *https://www.udl.co.uk/insights/crispr-cas9-from-gene-editing-to-epo-patent-law-editing (2018, accessed 28 September 2018)*

[10] Lander ES. The Heroes of CRISPR. Cell 2016; 164: 18-28. DOI: 10.1016/j.cell.2015.12.04

[11] van Doren D, Koenigstein S, Reiss T. The development of synthetic biology: A patent analysis. Syst Synth Biol 2013; 7: 209-220. DOI: 10.1007/s11693-013-9121-7

[12] Sha H, Wang D, Yan D, et al. Chimaeric antigen receptor T-cell therapy for tumour immunotherapy. Biosci Rep 2017; 37: BSR20160332. DOI: 10.1042/BSR20160332

[13] Research Associates H. Awaren ess & Impressions Of Synthetic Biology, *https://www.cbd.int/doc/emerging-issues/emergingissues-2013-07-WilsonCenter-SynbioSurvey-en.pdf* (2013).

[14] Goodman C. Engineering ingenuity at iGEM. Nat Chem Biol 2008; 4: 13. DOI: 10.1038/nchembio0108-13

[15] Loera M. Biohacking en México: talento y visión - red synbioMX, *https://synbiomx.org/2017/07/08/biohacking-en-mexico-talento-y-vision/* (2017, accessed 19 September 2018).

[16] Howard J, Murashov V, Schulte P. Synthetic biology and occupational risk. J Occup Environ Hyg 2017; 14: 224-236. DOI: 10.1080/15459624.2016.1237031

[17] Gibson DG, Young L, Chuang RY, et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 2009; 6: 343-345. DOI: 10.1038/nmeth.1318

[18] Fleming N. Edit Your Future with a Career in CRISPR. NatureJobs. NatureJobs. DOI: 10.1038/nj0503

[19] Kowalski H. Human Longevity, Inc. Completes $220 Million Series B Financing - Human Longevity, Inc., *https://www.humanlongevity.com/human-longevity-inc-completes-220-million-series-b-financing/* (2016, accessed 19 September 2018).

[20] Ro DK, Paradise EM, Quellet M, et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 2006; 440: 940-943. DOI: 10.1038/nature04640

[21] Deisseroth K. Optogenetics. Nat Methods 2011; 8: 26-29. DOI: 10.1038/nmeth.f.324

[22] Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science (80-) 2012; 337: 1628. DOI: 10.1126/science.1226355

[23] Gaj T, Gersbach CA, Barbas CF. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 2013; 31: 397-405. DOI: 10.1016/j.tibtech.2013.04.004

[24] Chao R, Yuan Y, Zhao H. Recent Advances in DNA Assembly Technologies. FEMS Yeast Res 2015; 15: 1-9. DOI: 10.1111/1567-1364.12171

[25] Darmostuk M, Rimpelova S, Gbelcova H, et al. Current approaches in SELEX: An update to aptamer selection technology. Biotechnol Adv 2014; 33: 1141-1161. DOI:10.1016/j.biotechadv.2015.02.008

[26] MacDonald IC, Deans TL. Tools and applications in synthetic biology. Adv Drug Deliv Rev 2016; 105: 20-34. DOI: 10.1016/j.addr.2016.08.008

[27] Kitada T, DiAndreth B, Teague B, et al. Programming gene and engineered-cell therapies with synthetic biology. Science (80-); 359. Epub ahead of print 2018. DOI: 10.1126/science.aad1067

[28] Dietzel A. Microsystems for pharmatechnology: Manipulation of fluids, particles, droplets, and cells. 2016. Epub ahead of print 2016. DOI: 10.1007/978-3-319-26920-7

ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-10

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.10

E-LOCATION ID: **e201802EE1**

Microscopio como Lector de Absorbancia con Utilidad en Análisis Clínicos

Microscope as Absorbance Reader with Utility in Clinical Analysis

I. J. Orlando-Guerrero[1], *C. H. Bravo-Delgado*[1], *Z. J. Hernández-Paxtián*[1], *A. A Aguilar-Felipe*[2]

[1]Universidad de la Cañada

[2]Laboratorio Clínico Huatulco

RESUMEN

La descentralización de laboratorios clínicos, se encuentra en desarrollo, lo anterior ha llevado a diseñar instrumentos que ofrecen resultados rápidos, confiables y al lado del paciente, esta tendencia se le conoce como *prueba en el punto de atención* (point of core testing POCT) y brinda la posibilidad de dar un seguimiento inmediato al padecimiento. El objetivo de esta investigación fue la implementación de un medidor de absorbancia, empleando la estructura de un microscopio óptico, al cual se le ha adaptado un filtro de luz, y una cámara digital. Lo anterior permite obtener valores de intensidad promedio de imágenes sólidas microscópicas de reacciones enzimáticas, y a partir de ellas estimar absorbancia o concentración. Como resultados se presentan rectas de calibración de absorbancia y mediciones de concentraciones para los casos de violeta de genciana, un kit de glucosa oxidasa y muestras problemas de pacientes voluntarios. Concluimos que existe un error de medición menor de ±1mg/dL comparados con las mediciones de un lector de placas de Elisa y un analizador de química sanguínea semiautomatizado. Teniendo en cuenta lo anterior el sistema resulta ser una alternativa viable para la estimación de absorbancia y aumenta la funcionalidad de microscopios ópticos en clínicas de salud.

PALABRAS CLAVE: Prueba en el punto de atención; lector de absorbancia; micro-espectrofotómetria

ABSTRACT

The decentralization of clinical laboratories is under development, which has led to the design of instruments that offer fast, reliable and patient-side results, this trend is known as point of core testing (POCT) and It offers the possibility of giving an immediate follow-up to the disease. The objective of this investigation was the implementation of an absorbance meter, using the structure of an optical microscope, to which a light filter and a digital camera have been adapted. This allows to obtain values of average intensity of solid images of enzymatic reactions, and from them to estimate absorbance or concentration. As results, we present absorbance calibration lines and concentration measurements for cases of gentian violet, a glucose oxidase kit and samples of volunteer patients. We conclude that there is a measurement error of less than ± 1mg / dl compared with the measurements of an Elisa plate reader and a semi-automated blood chemistry analyzer. Taking into account the above, the system turns out to be a viable alternative for estimating absorbance and increasing the functionality of optical microscopes in health clinics.

KEYWORDS: Point of Core Testing: Absorbance Reader; Microspectrophotometer

Correspondencia

DESTINATARIO: **Israel Jesús Orlando Guerrero**
INSTITUCIÓN: **Universidad de la Cañada**
DIRECCIÓN: **Carretera Teotitlán-San Antonio Nanahuatipán Km 1.7 S/N, Paraje Titlacuatitla, C.P. 68540, Teotitlán de Flores Magón, Oaxaca, México**
CORREO ELECTRÓNICO: **iorlando@unca.edu.mx**

Fecha de recepción:

28 de agosto de 2018

Fecha de aceptación:

11 de enero de 2019

INTRODUCCIÓN

El empleo de microscopios invertidos se ha difundido ampliamente en los últimos años, esto debido a que se pueden observar cultivos celulares sin una previa preparación, dicha actividad puede ser resaltada gracias a técnicas de contraste de fase, con las cuales se evita el uso de biomarcadores. Sin embargo, los microscopios ópticos no solo pueden ser empleados para observación, también permiten cuantificar concentraciones de solutos presentes en una célula, tal es el caso de la reciente técnica llamada micro-espectrofotómetria UV-Vis [1], esta técnica añade un valor sustancial a microscopios ópticos, ya que no lo limita a la observación, sino que también expande sus capacidades como un medidor de absorbancias o concentraciones. Lo anterior permite a clínicas de salud tener una doble funcionalidad en un solo instrumento y por ende realizar análisis rápidos, confiables y al lado del paciente, como lo dicta la tendencia POCT [2].

En la presente investigación se emplea un microscopio invertido como un lector de absorbancia, esto se logra añadiendo al sistema un filtro de luz, una cámara digital CMOS para la captura de imágenes digitales sólidas y un análisis numérico de las mismas.

Sensor CMOS como medidor de absorbancia

La absorbancia (A) es función del registro de intensidades que son obtenidas por el detector, es decir, es una comparación entre la intensidad de referencia (Ir) (intensidad medida en el detector sin muestra) y la intensidad de la muestra (I). La absorbancia puede ser calculada en función de dichos términos, aplicando la conocida ley de Lambert Beer como sigue [3]:

$$A = log_{10} I_r/I \qquad (1)$$

Por ser una comparación, la sensibilidad del detector es importante en los límites inferiores y superiores de medición admisibles del detector, por ejemplo, si el detector es capaz de medir absorbancias del orden de 5, significa que este puede medir intensidades muy bajas del orden de 0.001% de la intensidad de referencia, lo anterior permite medir concentraciones muy altas de un soluto. Por el contrario, absorbancias del orden de 0.001 significa que el detector debe ser capaz de medir intensidades altas del orden de 99.77 % de la intensidad de referencia, por lo tanto, puede medir concentraciones muy bajas de un soluto [3].

Habitualmente en espectrofotometría se emplean fotomultiplicadores como detectores, los cuales generan una corriente que es proporcional a la intensidad incidente, por lo tanto, el cálculo de la absorbancia se realiza empleando la Ecuación (1), pero en términos de intensidad de corriente.

Sin embargo, en la actualidad las cámaras digitales CMOS han empezado a remplazar este tipo de detectores ya que pueden ser empleadas como discriminadores de niveles de luz, es decir como sensores de luminancia. Para este caso, la luminancia es calculada a partir de la siguiente ecuación:

$$L = f^2 N/kts \qquad (2)$$

Dónde: L es la luminancia en candela/metro [2], N es el valor del pixel en la imagen digital en formato 8 o 16 bits, t es la velocidad de obturación (cuadros por segundos), f es el número de apertura, S es la sensibilidad del detector ISO, y k es una constante de calibración obtenida a partir de luminancias estándares [4].

Debido a que la absorbancia es obtenida por una división, esta resulta ser la misma en términos de luminancia, valor de pixel en 8 bits o 16 bits. Lo anterior se confirma en la Tabla 1, donde se muestra un valor de I e Ir y su cálculo de absorbancia en términos de las tres magnitudes. Debido a que es posible estimar la absorbancia en niveles de gris, en esta investigación usamos el formato de valor de pixel de 8 bits para su cálculo.

TABLA 1. Calculo de absorbancia, empleado la luminancia y dos formatos del valor del pixel.

	Pixel 8	Pixel 16	Luminancia
I	235	60395	148208.589
I_r	255	65635	160822.0859
A	0.036	0.036	0.036

Por otro lado, el rango de medición inferior de absorbancia de una CMOS, puede estar limitado por la corriente de oscuridad, la cual es debida a electrones generados térmicamente en total oscuridad, por ejemplo, la CMOS empleada en esta investigación, tiene un ruido promedio de oscuridad de 2±1pixel en niveles de gris, lo cual equivale a una medición de absorbancia de 1.4065, considerando una intensidad de referencia de 255. El límite superior de medición de absorbancia, está limitado por la diferenciación entre niveles de gris a intensidades altas, para este experimento se observaron diferencia significativa entre niveles de gris hasta 251±1 pixel, lo cual equivale a una absorbancia de 0.0068. Por lo anterior el rango de medición de absorbancia de la CMOS es de 2.0986 a 0.0068.

METODOLOGÍA

Sistema óptico

Para la generación de las imágenes sólidas de solutos, se empleó un microscopio invertido marca Olympus modelo CKX41[5], (ver Figura 1a), trabajando en campo claro, la descripción del funcionamiento del MLA (microscopio medidor de absorbancia) se describe a continuación: La luz incoherente proveniente de la fuente de tungsteno es filtrada a luz cuasi cromática por medio de un filtro pasa banda modelo FB510-10-Ø1 (marca thorlabs [6]), cuya longitud de onda central es de CWL=510 ± 2 nm, y un ancho medio de FWHM = 10 ± 2 nm, este filtro es insertado en el control deslizante modelo IX2-SLP en la poción reservada para iluminación en campo claro, las otras dos posiciones empleadas para iluminación en contraste de fase no son empleadas en el experimento, lo anterior se muestra en la Figura 1b.

FIGURA 1. a) Sistema optico empleado para la captura de imágenes solidas de reacciones enzimaticas y b) filtro y control dezlizante para iluminacion cuasi cromática.

Posteriormente la luz cuasi cromática es dirigida a una placa de ELISA de 96 posos por medio de un condensador que genera iluminación tipo Köhler, cada poso puede ser seleccionado por el usuario por la platina x,y del microscopio. La intensidad e imagen de cada poso es colectada por medio del objetivo de 60x (N.A=0.85), de la marca edmundoptics y dirigida a una cámara digital CMOS Infinity 1 de 2.0 megapixeles de la marca lumera, con tamaño de pixel de 4.2µmx4.2µm y una resolución de imagen de 1600x1200 pixeles [7]. La intensidad promedio y la absorbancia para cada poso de la imagen digital es calculada por medio de un programa el cual es descrito en la siguiente sección. Cabe mencionar que el experimento es llevado a cabo en un cuarto oscuro, con el fin de evitar la luz parasita.

Imagen sólida

La imagen sólida de una reacción enzimática, en esta investigación, se refiere a la imagen digital obtenida por un microscopio conectado una cámara digital, en

este caso no es el interés la estructura presente en la imagen digital, sino el valor promedio de los pixeles en la imagen de luminancia [8], el cual es empleado en el cálculo de la absorbancia. El procedimiento para obtener este valor es el siguiente: Se captura una imagen sólida con una cámara digital CMOS, esta imagen es descompuesta en sus componentes RGB, las cuales son empleadas para obtener una imagen de luminancia relativa de acuerdo a la norma CIE XYZ [9] posteriormente se calcula la intensidad promedio Im de todos sus pixeles, aproximando al valor más próximo. Este valor es empleado para calcular la absorbancia correspondiente a la imagen sólida empleando la Ecuación (1) y considerando Ir=251. Este procedimiento se muestra en la Figura 2.

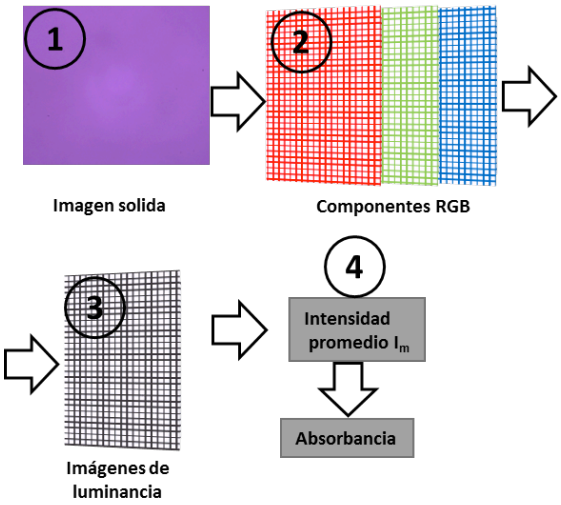


FIGURA 2. Procedimiento de conversión de imagen sólida a valor de absorbancia.

Un ejemplo de una secuencia de imágenes solidas de violeta de genciana para concentraciones en aumento, son presentadas en la Figura 3, en cada caso se obtienen el promedio de intensidad de los pixeles en la imagen de luminancia, y como se puede observar esta disminuye conforme aumenta la concentración del colorante, mientras que la absorbancia aumenta progresivamente. En estas imágenes solidas no existe alguna estructura, ya que la intensidad promedio y el valor de absorbancia es la de interés.

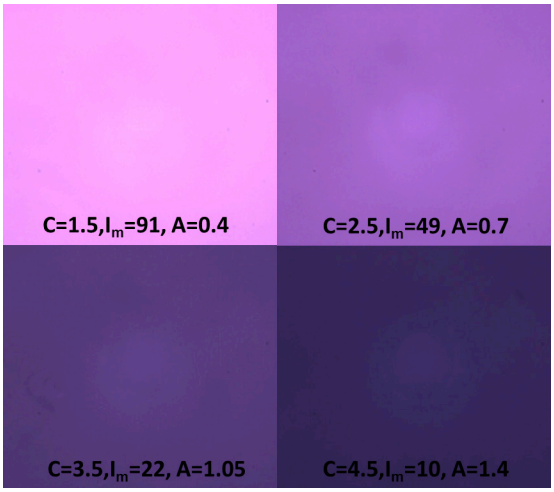


FIGURA 3. Secuencia de imágenes solidas de colorante violeta de genciana para concentraciones de 1.5 a 4.5 mg/dL. Para cada imagen C=concentración mg/dL, Im=Intensidad promedio de pixeles en la imagen, A=Absorbancia.

Programa para el cálculo de Absorbancia y concentración

Se implementó un programa que realiza mediciones de absorbancia y concentración a partir de imágenes solidas de solutos, este programa se realizó en el software Labview 2015 y utilizando el toolkit visión [10], este último permite la activación y captura de la cámara CMOS, además de contar con funciones especializadas en el procesamiento de imágenes.

El programa cuenta dos opciones, medición de absorbancia (A) y medición de concentración (C), que el usuario puede elegir de acuerdo a sus necesidades (ver Figura 4). Para el caso de *medición de absorbancia*, el programa solicita al usuario introducir el ***número de lecturas (1)*** a realizar, posteriormente el usuario ubica el primer poso a leer en la placa de Elisa utilizando la **platina x, y (2)** del microscopio; a continuación el programa activa la cámara CMOS y **captura una imagen (3)** de tamaño 1600x1200 pixeles. Las imágenes RGB de la imagen captura son extraídas y empleadas para obtener una imagen de **luminancia (4)** de la cual se obtiene la **intensidad promedio I_m (5)** de todos sus

pixeles, redondeado al valor más próximo. Este valor es empleado para calcular la absorbancia correspondiente a la imagen sólida en turno (Ecuación (1) y considerando Ir=251). El proceso de conversión de imagen solida a medición de absorbancia, es mostrado en la Figura 2. Posteriormente el programa retorna al paso 2 y el usuario ubica un nuevo poso a leer. Finalmente, las n mediciones de absorbancia son almacenadas en un **vector (6)** que contiene las relaciones de concentración conocidas y sus mediciones de absorbancia.

Para la opción de la ***medición de concentración*** el vector que contiene las relaciones de concentración y absorbancia es utilizado para obtener una **recta de calibración (7)** por medio de una regresión lineal, esta recta es del tipo:

$$A=mc+b \rightarrow c=A-b/m \qquad (2)$$

Dónde: A es la absorbancia, m es la pendiente de la recta, c es la concentración, y b es la intersección con el eje de la absorbancia. Por último, el programa estima **concentración (8)** a partir del despeje de la recta de calibración en términos de la concentración (ver Ecuación (3)). El diagrama de flujo del programa es mostrado en la Figura 4.

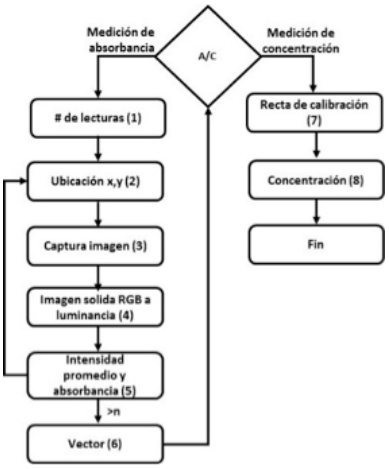


FIGURA 4. Pasos que realiza el programa para obtener mediciones de absorbancia y concentración de reacciones enzimáticas.

RESULTADOS

Colorante de utilidad clínica

Con el único fin de analizar el desempeño del MLA, se realizó comparación cuantitativa de mediciones de absorbancia entre un lector de elisa modelo H Reader 1 y el sistema propuesto. Para lo anterior, se emplearon diferentes concentraciones de un colorante de utilidad clínica como lo es el violeta de genciana ($C_{24}H_{28}N_3Cl$ al 1 % en agua destilada), el cual usualmente se emplea para la realización de curvas de calibración en determinación de metabolitos en sangre. Se preparó una solución stock de violeta de genciana a una concentración de 100mg/dL, y se obtuvieron disoluciones en agua destilada de 1.5 a 6.5 con variación de 1mg/dL (C en la columna uno de la Tabla 2). Las mediciones de absorbancia para los dos lectores son mostradas en la siguiente tabla.

TABLA 2. Tabla comparativa de mediciones absorbancia y concentración en ELISA y MLA, para diferentes concentraciones de violeta de genciana.

C	Abs-E	Abs-MLA	C-MLA	E-C
1.5	0.439	0.441-(91)	1.506	0.006
2.5	0.708	0.709-(49)	2.383	0.117
3.5	1.06	1.057-(22)	3.518	0.018
4.5	1.371	1.400-(10)	4.636	0.136
5.5	1.682	1.701-(5)	5.618	0.118
6.5	1.972	1.923-(3)	6.342	0.158

Se puede observar en la Tabla 2 que las mediciones de abs-E y abs-MLA (mediciones de absorbancia del lector de ELISA y el sistema propuesto respectivamente) son muy cercanas unas a otras, por lo tanto, podemos intuir que el sistema MLA es comparable con lectores de absorbancia comerciales.

Por otro lado, con el fin de cuantificar la exactitud en la medición de concentraciones por parte del MLA, se obtuvo la recta de calibración del sistema aplicando

una regresión lineal a los datos de concentración y Abs-MLA obteniendo un R2 =0.9959. Dicha ecuación lineal es despejada con respecto a la concentración para su correspondiente cálculo, como se muestra en la siguiente ecuación:

$$C\text{-}MLA = A + 0.0207/0.3064 \quad (4)$$

Dónde: C-MLA es la concentración estimada en mg/dL, A es la absorbancia, 0.0207 es la intersección en el eje de la absorbancia y 0.3064 pendiente de la recta de calibración. Así mismo el error en la medición en la concentración (E-C) es mostrado en la cuarta columna de la tabla dos, el cual es calculado a partir del valor absoluto de la diferencia de la concentración experimental con la concentración calculada con la ecuación cuatro, en esta columna se observa que el error aumenta conforme aumenta la concentración, esto puede deberse a que la diferencia entre intensidad de niveles de gris de imágenes digitales de soluciones, a concentraciones altas es muy pequeña. Por ejemplo, para concentraciones de 5.5 y 6.5mg/dL la variación de intensidad es tan solo de dos en una escala de grises. Lo anterior también se puede verificar observando la Figura 5, ya que ambas graficas se separan más a concentraciones altas.

Recta de calibración de glucosa

Fue realizada una recta de calibración con el reactivo líquido para la determinación fotométrica de glucosa en suero o plasma del Grupo Mexlab [11], con el objetivo de cuantificar el desempeño del MLA comparado con un lector de ELISA empleado en la investigación. La recta de calibración del sistema aplicando una regresión lineal a los datos de concentración y Abs-MLA fue obtenida con un R2 =0.9991. Dicha recta de calibración es despejada con respecto a la concentración para su cálculo, como se muestra en la siguiente ecuación:

$$C = A - 0.0166/0.0085 \quad (5)$$

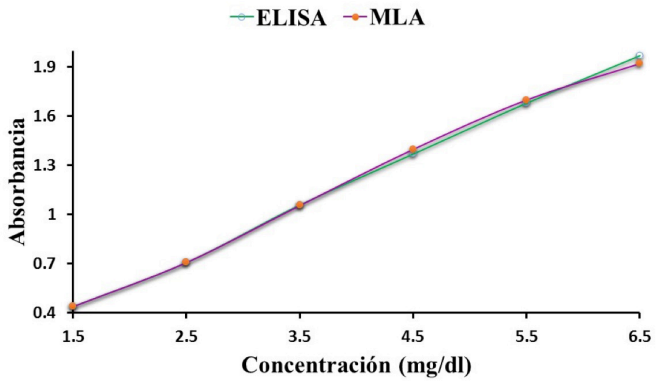


FIGURA 5. Grafica comparativa de mediciones de absorbancia de abs-E y abs-MLA.

Dónde: C-MLA es la concentración estimada por el sistema propuesto en mg/dL, A es la absorbancia, 0.0207 es la intersección en el eje de la absorbancia y 0.3064 pendiente de la recta de calibración. En la tabla tres se muestran las mediciones de absorbancia de ambos lectores (Abs-E y Abs-MLA), columna dos y tres, mientras que en la cuarta y quinta columna son mostrados las concentraciones estimadas con la Ecuación cinco (C-MLA) y su error respectivamente (E-C).

TABLA 3. Mediciones absorbancia y concentración en elisa y MLA, para diferentes concentraciones del reactivo líquido para la determinación fotométrica de glucosa en suero.

C	Abs-E	Abs-MLA	C-MLA	E-C
75	0.669	0.651-(56)	74.635	0.365
100	0.894	0.868-(34)	100.164	0.164
125	1.119	1.077-(21)	124.752	0.248
150	1.344	1.286-(13)	149.341	0.659
175	1.569	1.497-(8)	174.164	0.836
200	1.794	1.701-(5)	198.164	1.836

La comparación de las rectas de calibración para los dos sistemas se muestra en la Figura 6, donde se puede notar que a concentraciones altas ambas se empiezan a separar, posiblemente porque la cámara digital CMOS tiene menos sensibilidad a concentraciones altas.

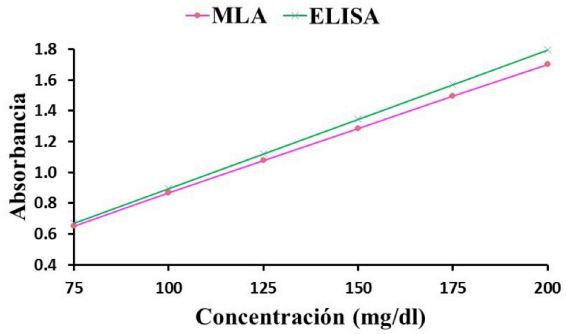


FIGURA 6. Gráfica comparativa de mediciones de absorbancia de Abs-E y Abs-MLA, para el caso del reactivo líquido para la determinación fotométrica de glucosa en suero.

Utilidad clínica del MLA

Uno de los objetivos principales de esta investigación, es el empleo del MLA como una alternativa a lectores de absorbancia convencionales, para análisis clínicos, en este caso el sistema tiene una doble funcionalidad, ya que puede servir como microscopio o cuantificador de concentraciones.

Por lo anterior se realizó la determinación de glucosa en muestras problemas de pacientes voluntarios de la Universidad de la Cañada, otorgándose al individuo una copia del consentimiento informado para el aseguramiento de su participación en esta investigación.

Se formaron dos grupos de tres pacientes, el primer grupo son individuos clínicamente sanos y el otro grupo fueron diagnosticados previamente como hipoglucémicos. A estos grupos se les extrajo 5 mL de plasma sanguíneo sin coagulante, después de dejar coagular la muestra, se continuo a centrifugar las muestras a 3500 rpm durante 5 minutos, para la obtención de suero sanguíneo. Una vez obtenidas la muestras, se realizó la reacción enzimática de Trinder God-Pod [11] y finalmente las mediciones de concentraciones fueron estimadas empleando el analizador de química sanguínea semiautomatizado (AQSA), modelo Easykem plus y el MLA, para este último caso, la concentración fue estimada a partir de la ecuación cinco correspondiente a la recta de calibración glucosa. Los resultados de ambos equipos son comparados en la Tabla 4.

TABLA 4. Tabla comparativa de mediciones absorbancia y concentración en AQSA y MLA, para seis pacientes.

PC	C-AQSA	Abs-MLA	C-MLA	E-C
1	398.56	----	----	----
2	282.199	----	----	----
3	145.812	1.254(14)	145.576	0.236
4	92.408	0.798(40)	91.929	0.479
5	73.953	0.651(56)	74.635	0.682
6	62.435	0.548(71)	62.517	0.082

PC: Número de paciente. C-AQSA: Medición de concentración con AQSA en mg/dL. Abs-MLA y C-MLA: Lecturas de absorbancia y concentración del sistema propuesto respectivamente. E-C: Error en concentración, calculado a partir de C-AQSA menos C-MLA.

Como puede observarse en la tabla cuatro, los pacientes uno y dos presentaron rangos de medición de absorbancia fuera del rango, permitido por el MLA, mientras que los otros cuatro pacientes están dentro del rango de medición, también puede observarse que los errores de medición no fueron mayores a 1mg/dL.

DISCUSIÓN

De acuerdo a las mediciones obtenidas con lectores de absorbancia comerciales, estas son semejantes a las mediciones obtenidas con el MLA, arrojando un error no mayor a 1mg/dL para los casos de los resultados del colorante, glucosa y muestras problemas de pacientes voluntarios (ver Tabla 2, Tabla 3, y Tabla 4). Para estos casos las rectas de calibración presentaron un R2 alto y una pendiente baja, lo cual indica que el sistema propuesto presenta una buena exactitud y precisión en mediciones arbitrarias (ver Figura 5 y Figura 6). Sin embargo, el límite de detección está acotado, debido a la sensibilidad de la CMOS que se empleó, ya que, a concentraciones altas, la intensidad de luz que llega al detector es muy poca y esta puede ser confundida con el ruido térmico generado por la CMOS, (ver Tabla 2).

Por otro lado, para intensidades altas (concentraciones bajas) la sensibilidad del sistema se reduce ya que los cambios de nivel de gris en la imagen solida son insignificantes (ver Figura 6). Sin embargo, empleando un rango de medición de absorbancia de 0.4 a 0.9, el sistema propuesto presenta una alta linealidad. Si se desea aumentar el rango de medición, el sistema debe de sufrir dos modificaciones, la primera consistiría en tener una cámara digital con un ruido térmico más bajo y la segunda consistiría en aumentar la intensidad de la fuente de iluminación, en ambos casos, una recta de calibración debe ser obtenida para una correcta medición en el MLA.

Por otro lado, la conjunción de la recta de calibración con el algoritmo, es de suma importancia, ya que este último calcula las concentraciones a partir de la recta de calibración es decir a partir de la pendiente y la intersección en el eje de absorbancia, por lo que un cálculo no adecuado de estos, arrojaría resultados erróneos (ver Ecuaciones 3, 4 y 5). Teniendo en cuenta lo anterior el sistema resulta ser una alternativa viable para la estimación de absorbancia, y proporciona a los microscopios ópticos una doble funcionalidad.

CONCLUSIONES

La adaptación de un filtro de luz, una CCD y un algoritmo computacional, a un microscopio óptico, aumentan sus cualidades como instrumento de medición, en este caso como medidor de absorbancia, para estimar concentraciones desconocidas. Lo anterior brinda la posibilidad de realizar mediciones cerca del punto de atención del paciente, ya que los microscopios ópticos son de uso común en clínicas de salud. A su vez el sistema brinda la posibilidad de la reducción de reactivos al poder calcular absorbancias de áreas microscópicas, por otro lado, en trabajos futuros, esta propuesta se pretende seguir validando, de acuerdo a normas establecidas, patrones de referencia y múltiples ensayos, con el objetivo de determinar los criterios de validez del método, así como la exactitud del método, precisión, linealidad, entre otros.

REFERENCIAS

[1] Grahn HF, Geladi, P. Technique and Applications of Hyperspectral Image Analysis. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, England: John Wiley & Sons, 2007.

[2] Christopher P Price. Regular review Point of care testing. bmj volume 322 26 may 2001.

[3] Greenfield Sluder and David E. Wolf. Digital Microscopy 3RD edition, Volume 81, Pages 64-67 (2007).

[4] Peter D. Hiscocks, P.Eng. Measuring Luminance with a Digital Camera, Syscomp Electronic Design Limited, February 16, 2014.

[5] Olympus, inverted microscope CKX41, 2016.

[6] Thorlabs. (2003). FB510-10. agosto 7,2018, de Thorlabs Sitio web: *https://www.thorlabs.com/thorproduct.cfm?partnumber=FB510-10*

[7] Lumenera Corporation. (2017). INFINITY1-2. julio,2018, de Lumenera Corporation Sitio web: *https://www.lumenera.com/infinity1-2.html*

[8] Anastas Dakashev,Stanch Pavlov,Krasimira Stancheva, Application of Digital Camera and Digital Image Processing Techmique for Molecular Absorption Analysis in the Visible Spectrum, Universal Journal of Chemistry Vol. 1(4), pp. 129 - 134 DOI: 10.13189/ujc.2013.010401

[9] Carranza-Gallardo, J. "Manejo de las fórmulas de diferencias de color vs límites de aceptabilidad". Memorias del Simposio de metrología 2002, Centro Nacional de Metrología, México., 2002.

[10] National Instruments. (2005). NI Vision. julio,2018, de National Instruments Sitio web: *http://www.ni.com/pdf/manuals/371007b.pdf*

[11] Grupo Mexlab. Bio-Glucosa, Reactivo para la determinación fotométrica de glucosa en suero o palsma y otros fluidos biológicos. Sitio web: *http://www.grupomexlab.com*

ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-18

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.11

E-LOCATION ID: **e201805EE1**

A code biology analysis of the regulatory regions in cell lines

Análisis de biología de códigos de las regiones reguladoras en líneas celulares

Omar Paredes[1], *Isaías May-Canche*[1], *Elena Fimmel*[2]

[1]Universidad de Guadalajara

[2]Mannheim University of Applied Sciences

ABSTRACT

Coding sequences are widely studied for their relevance in protein synthesis. However, higher organism genomes, such as human genomes, has a small amount of them, and a larger proportion of non-coding sequences. ENCODE and Epigenomic Roadmap projects discovered that regulatory functions are carried out in the non-coding regions of the human genome. These regulatory functions are part of the regulatory machinery that yields different gene expression profiles, thus, different cell lines. Whereas different environmental elements, i. e. histone modifications, DNA methylation, and other epigenomic phenomena, determine the regulatory function of genome part, the sequences' composition where these functions take place could also influence regulatory machinery. In this work, we explore the non-coding regulatory sequences and lexica build with subsequences between 3 and 16 nucleotides to evaluate the difference between the sequence composition of the regulatory regions in the cell lines. Our results show that the lexica corresponding to the regulatory regions are different based on their complexity/degeneracy, moreover, the lexica of regulatory regions in different cell lines are also different. These features suggest that non-coding sequences are an active element of the regulatory machinery and the histone code that are involved in cell differentiation.

KEYWORDS: Lexicon Complexity; Regulatory Regions; Code Biology

RESUMEN

Las secuencias codificantes han sido ampliamente estudiadas por su relevancia en la síntesis de proteínas. Sin embargo, los genomas de organismos complejos, como el humano, tiene una porción menor de estas secuencias y una mayor proporción de secuencias no codificantes. Los proyectos del ENCODE y Epigenomic Roadmap describieron que las funciones reguladoras se llevan acabo en las regiones no codificantes del genoma humano. Estas funciones reguladoras son parte de la maquinaria reguladora que produce diferentes perfiles de expresión genética, por tanto, diferentes líneas celulares. Mientras diferentes elementos del entorno, como las modificaciones en las histonas, metilación del ADN y otros fenómenos epigenéticos, determinan la función reguladora que tienen una porción del genoma, la composición de la secuencia donde estas funciones son llevadas a cabo también podrían influir en la maquinaria reguladora. En este trabajo, se exploraron las secuencias de las regiones no codificantes y los léxicos generados con las subsecuencias entre 3 y 16 nucleótidos, para evaluar las diferencias entre la composición de las secuencias de las regiones reguladoras en las líneas celulares. Los resultados muestran que los léxicos correspondientes a las regiones reguladoras son diferentes con base en su complejidad/degeneración, así mismo, los léxicos de las regiones reguladoras en distintas líneas celulares son también distintos. Estos detalles sugieren que las secuencias no codificantes son elemento activo de la maquinaria reguladora y del código histónico que participan en la diferenciación celular.

PALABRAS CLAVE: Complejidad de Léxico; Regiones Reguladoras; Biología de Códigos

Correspondencia

DESTINATARIO: **Omar Paredes**
INSTITUCIÓN: **Univerisdad de Guadalajara**
DIRECCIÓN: **Blvd. Gral. Marcelino García Barragán #1421, C. P. 44430, Guadalajara, Jalisco, México**
CORREO ELECTRÓNICO: **omar.paredes@alumnos.udg.mx**

Fecha de recepción:

21 de septiembre de 2018

Fecha de aceptación:

11 de enero de 2019

INTRODUCTION

DNA sequences are carriers of hereditary material in all living organisms [1]. The hereditary information in DNA is stored as a code made up of four chemical bases, adenine (A), guanine (G), cytosine (C), and thymine (T), written in triletter words (codons) without delimiters that are decoded after copying into a complementary RNA (transcription) into a matching protein sequence in a process called translation. In the 2000s, the Human Genome Project estimated that only approximately 2% of human genome consists of coding sequences and the remaining large part of the DNA (non-coding regions) does not serve as a template for protein sequences [2].

However, ENCODE and Epigenomic Roadmap consortiums evidenced that there are regulatory functions in the apparently non-functional sequence of the human genome [3, 4]. Both consortiums located the regulatory regions in 127 cell lines based on epigenomic profiles [4, 5], and thus, they implemented an experimental whole-genome validation of the histone code.

The histone code is a set of rules that maps the histone modifications to chromatin packaging events and leads to regulatory functions in gene expression [6-8]. Altogether, these events build a regulatory machinery that depends on the environmental context, shows diverse gene expression profiles and, hence, a diversity of cell lines [9, 10].

Elements of the previously mentioned context, that possibly determine a cell line, are locations where the chromatin packaging events happen [11-15].

A way for studying such phenomena is suggested within an emergent discipline, Code Biology, which considers life events, for instance, as maps between organic signs and organic meanings [16-18]; in this work represented by genomic sequence and regulatory function, respectively. The Code Biology approach includes a methodology for identifying organic codes consisting of three steps: (i) demonstrating the existence of two sets linked by an organic code; (ii) identifying the decoder of the organic code, called adaptor; and (iii) validating an arbitrary nature of the organic code (compare, for instance, Hofmeyr [19]).

In particular, in [20], the histone code was examined from this viewpoint: "we try to show how simple combinations of essential elements such as histone modifications can participate in sophisticated cellular features such as the structure of the genome. Here code is identified, where an input system (histone modifications) is translated into an output system (chromatin states) via adaptors (epigenetic regulators or transcription factors). Such a code has distinct importance in gene regulation and consequently for the cellular phenotype".

In this exploratory work, we implement genomic signal processing and natural language techniques to explore the sequences of regulatory regions and evidence that indeed these sequences play an important role in the regulatory machinery.

METHODS

In this work, the regulatory regions of three types of human cell lines are being explored to identify differences between the regulatory machineries in these cell lines at the sequence level. A workflow of the methodology in this work shown in Figure 1.

In order to perform this preliminary analysis, we choose the cell lines: H1 cells, Primary T CD8+ naive cells and Brain hippocampus middle, that represent pluripotent cells, first culture, and differentiated cells.

We download the files of the three cells corresponding to 14 regulatory regions (Table 1) proposed by the Epigenomic Roadmap Map project from the database of the mentioned project (*http://www.roadmapepigenomics.org/data/*, August 2018).

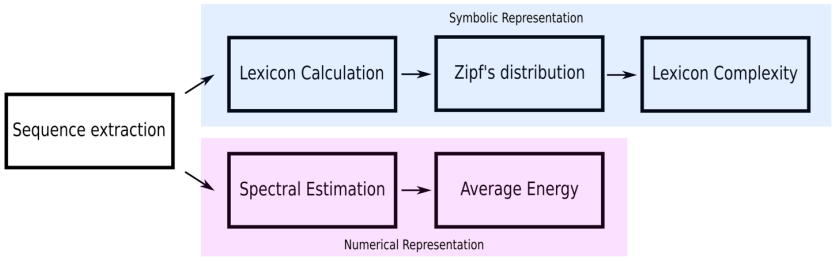


FIGURE 1. **Methodology workflow.**

TABLE 1. **Regulatory regions proposed by the Epigenomic Roadmap Consortium [4].**

Regulatory regions	
Abbreviation	**Name**
TssA	Active TSS
TssAFlnk	Flanking active TSS
TxFlnk	Transc. at gene 5'and 3'
Tx	Strong transcription
TxWk	Weak transcription
EnhG	Genic enhancers
Enh	Enhancer
ZNF/Rpts	ZNF genes + repeats
Het	Heterochromatin
TssBiv	Bivalent/poised TSS
BivFlnk	Flanking bivalent TSS/Enh
EnhBiv	Bivalent enhancer
ReprPC	Repressed Polycomb
ReprPCWk	Weak repressed Polycomb

The downloaded files contain the location indices of the regulatory region in the human genome. Based on the indices, we extracted the corresponding sequences and mapped them into a genomic signal by the Voss method. In this work, we keep both representations of the DNA, sequences and the genomic signals.

The Voss method is a tetradimensional graphic of the DNA sequences that represent in each dimension a nucleotide and value the presence $x[n]=1$ and absence $x[n]=0$ of the respectively nucleotide. For example, the genomic signal of the sequence "GTCAGTCGTAA" is:

A=[00010000011], C=[00100010000],
G=[10001001000], T=[01000100100].

Symbolic representation

We classify sequences into 14 groups, where each group contains the sequences with one of the regulatory functions from Table 1. As asserted in the Introduction, a DNA sequence can be symbolically represented as a chain of four letters (A, T, C, and G). In this representation, a word of length k or k-mer is an arbitrary subsequence that contain k consecutive nucleotides. It is easy to see that the number of words in a sequence of length l is equal to l-k+1. Hereinafter, we will call these words the k-lexicon of the sequence.

According to this approach, we calculate the k-lexica for each sequence for the **k** values from 3 to 16 nucleotide towards to identify relevant lexica in the non-coding regulatory sequences. Then we calculate the relative frequency for each word in each of the k-lexica to obtain the probability distribution of the lexicon and order the frequencies in the descending order.

The obtained distribution is the so-called Zipf's law distribution (Figure 2). The Zipf's law is a power law that describes many types of data studied in the physical and social sciences, among them the language [21], and states, for instance, that the frequency of any word is inversely proportional to its rank in the frequency table. In the specific case of the language, the Zipf's law is a measure of the complexity/degeneracy of the language and an expression of the least effort principle of the vocabulary [22, 23]. We will adopt the equation of the Zipf's law distribution in the following form:

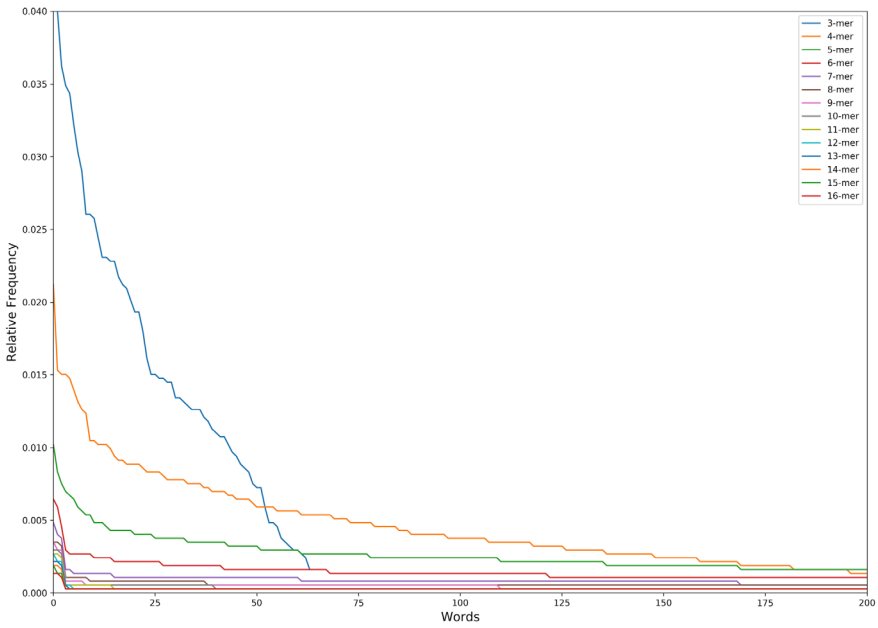


FIGURE 2. The Zipf's Law distribution of the first 200 words (except 3 nucleotides that has a maximum of 64 words lexicon) in the n-lexicon from 3 nucleotides to 16 nucleotides.

$$g(w) = \frac{A}{r(w)^b} \quad (1)$$

where w is a word in the lexicon; $r(w)$ the absolute frequency of the word; A a constant; b the value of the exponent characterizing the Zipf's distribution. In this work we call b the lexicon complexity, and mean the higher the value of b the higher the complexity/degeneracy of the vocabulary [24]. After all, $g(w)$ denotes the relative frequency of the word w.

Thereafter, we linearize the Zipf's distribution by dividing each value with its respective inverse. The result distribution is now a distribution with linear behavior, which slope is the lexicon complexity. We do a linear regression by the least square method to calculate the lexicon complexity of each sequence for its vocabularies from 3 nucleotides to 16 nucleotides.

Numerical representation

For each genomic signal, we calculate its periodogram. A periodogram is a technique to obtain the frequency spectrum of a signal, in this case, a genomic signal. This technique enhances the spectrum and fixes it to a certain length, that is important to this work because of the variable lengths in the sequences of this work. We fix all the periodogram to the length of 500 values.

The equation to calculate the periodogram is given by Eq. 2 where $X[n]$ represents the periodogram of the genomic signal, N the number of points, $x[n]$ is the genomic signal, in this work the Voss representation, and f the frequency. An example of a periodogram is shown in Figure 3.

$$\hat{X}[n] = \frac{1}{N}\sum_{n=0}^{N-1} \tilde{x}[n]\, e^{-i2\pi f n} \quad (2)$$

After all, we divide the periodogram into frequency bands that correspond to periodicities in the genomic signal, recalling that the inverse of the frequency is the periodicity. We have then 14 intervals of frequencies (Table 2) that correspond to the same values of length, from k=3 to k=16, that we calculate for the lexicon complexity.

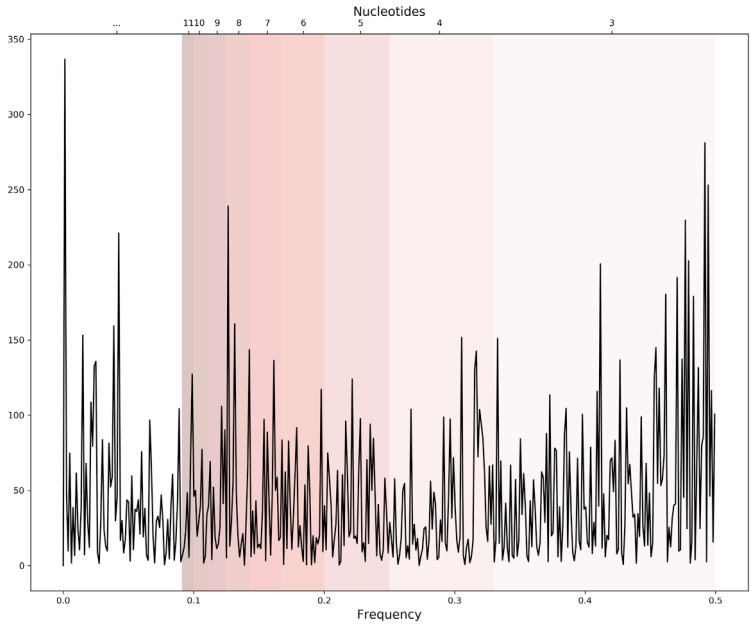


FIGURE 3. **An example of a periodogram of one genomic signal. Bands of frequencies represent the pattern information of a length in the sequence correspondent to certain nucleotides i.e. 0.33-049 frequencies represent the patterns of 3 nucleotides length.**

TABLE 2. **Intervals of frequency and their corresponding periodicities.**

Periodicity (nt)	Interval of frequencies (Hz)
1	0.330 – 0.500
2	0.250 – 0.330
3	0.200 – 0.250
4	0.167 - 0.200
5	0.143 – 0.167
6	0.125 – 0.143
7	0.111 – 0.125
8	0.010 – 0.111
9	0.091 – 0.010
10	0.083 – 0.091
11	0.077 – 0.083
12	0.071 – 0.077
13	0.067 – 0.071
14	0.062 – 0.067
15	0.067 – 0.071
16	0.062 – 0.067

Then, we calculate the average energy of each interval to evaluate the average potential capacity to encode information in the evaluated pattern length [25], adopting the following average energy equation

$$E_k = \frac{1}{N_k} \sum_{n=f_o}^{f_u} \hat{X}[n] \qquad (3)$$

where E_k is the energy ascribed to the periodicity o pattern length k; f_o is the initial frequency of the interval; f_u the upper frequency of the interval; $X[n]$ is the periodogram of the genomic signal; and N_k the number of points in the periodogram corresponding to pattern length k frequencies.

RESULTS AND DISCUSSION

The 14 regulatory regions used in this work correspond to 1'009,178 sequences, distributed as follows: 363,513 from cell line H1 cells; 249,377 from cell line Primary T CD8⁺ naive cells; and 396,288 from cell line to Brain hippocampus middle. The lengths of the sequences vary from 200 to 2,000 nucleotides.

For each sequence, we calculate 14 average energy, and, respectively, 14 lexicon complexity values corresponding to the patterns of lengths between 3 and 16

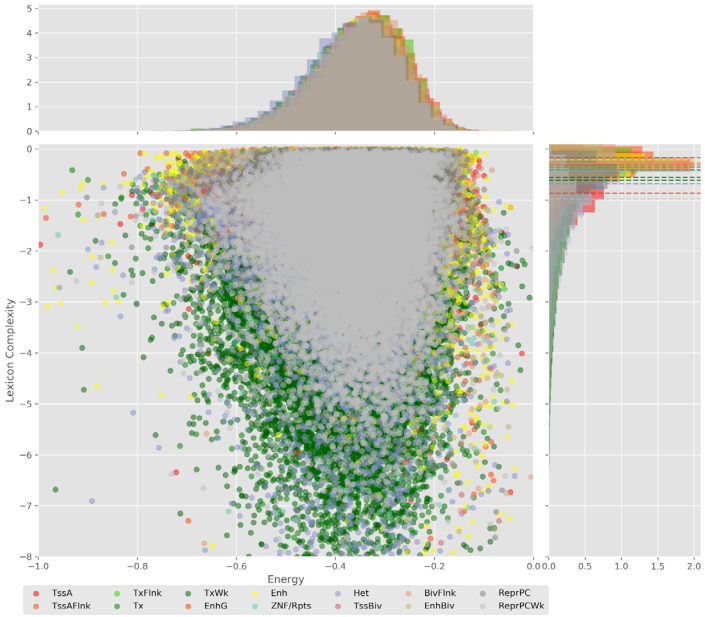


FIGURE 4. Energy against the 6-lexicon complexity (lower left) of the DNA sequences for he 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.

nucleotides. The average energy and the lexicon complexity are indicators that point out to the potential capacity, and respectively, information encoding quality/degeneracy in DNA sequences [26-30].

Figure 4 displays the potential capacity against quality dynamics for the regulatory sequences of the cell line H1 cells and their k-lexicons with k=6. The histograms of the energy (upper left, Figure 4) can be interpreted the way that the potential to encode information at the length of 6 nucleotides tends to be similar in the sequences of any regulatory region. The same behaviour can be observed in the rest of the lengths (see figures S1-S13) that likely indicates that there is no difference in codifying information for any word length.

However, the lexicon complexity for k=6 (right, Figure 4) behaves in a different way for different regulatory regions. The dotted lines represent the common value of lexicon complexity in the sequences of the 14 regulatory regions. This feature suggests that there is a difference, at least for this length, between the regulatory regions in the information encoding quality. Let us note that biological information encoding quality in DNA sequences could be interpreted as a degree of degeneracy.

Degeneracy is a biological phenomenon that means the ability of elements that are structurally dissimilar to perform the same function or yield the same output [26, 27]. In this work, this notion represents the ability of multiple sequences with a certain potential capacity to codify a unique biological function, a regulatory function.

The different degeneracy values in the regulatory regions indicate that the diversity of the words in their respective lexicons is different depending on the region, and meaning, as well as the numbers of words (signs) that encode such regulatory function (significant). As for their codes, i. e. the relationships between corresponding sets of signs and significant, they may

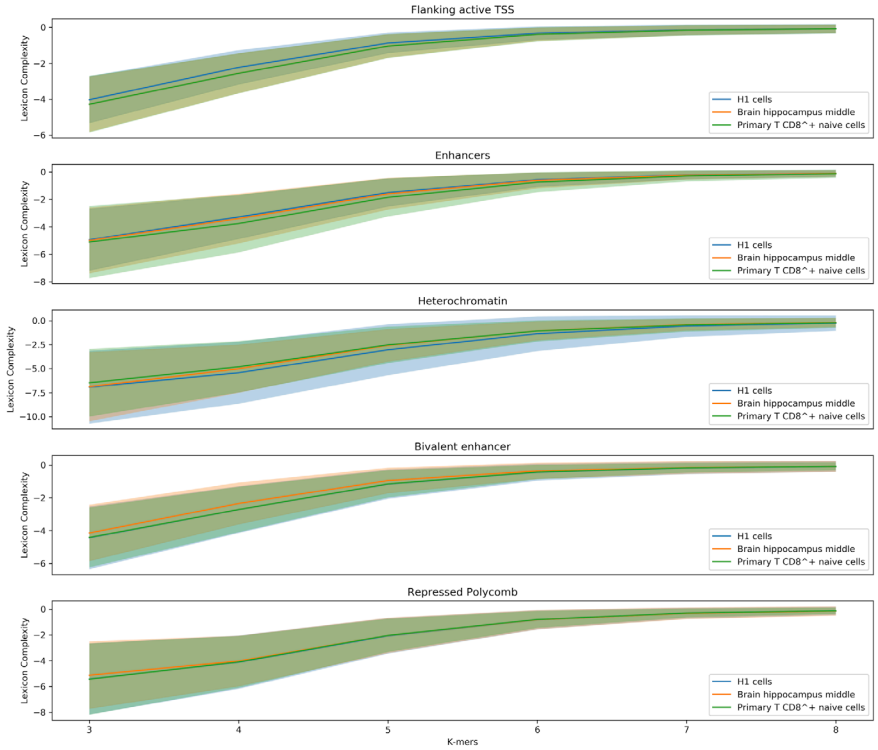


FIGURE 5. Distribution of the lexicon complexity for the regulatory regions Flanking active TSS, Enhancer, Heterochromatin, Bivalent enhancer and Repressed Polycomb in the cell lines H1 cells, Brain hippocampus middle, and Primary T CD8+ naive cells. The bold line represents the mean value of the lexicon complexity of the sequence in the respective cell line for each regulatory region, while the shadow areas represent the standard deviation of the lexicon complexities values.

be different, too. Nevertheless, in this work we don't explore the specific words for each of the single regulatory function lexicons, and, thus we describe the code of each regulatory region as a whole.

The similar behavior, i. e. different degeneracy values in different regulatory regions, can be observed for k=4, 5, and 7 nucleotides, meanwhile, for the other lengths, the degree of degeneracy tends to be similar for all the regulatory regions. This interesting fact can be explained in the following way: a small word length (1, 2, and 3 nucleotides) enables a brief lexicon and the number of regulatory regions coded would be correspondingly small, while a bigger word length (above 7), leads to a wider vocabulary, and, thus, to an enormous amount of energy needed to maintain the code. On the one hand, it contradicts the less effort principle of nature; on the other hand, a very specific code obtained with the number of signs approximately equal to the number of significances would be easier to "hack" what is a risk for the robustness of an organism.

Although the evidence of different degrees of degeneracy refers to the feasibility of a code for the regulatory sequences, it does not yet indicate that the sequence itself plays a role in the context of the regulatory machinery that determines the cell lines.

In order to explore the influence of the sequence's composition, we are comparing lexicon complexities of the five regulatory regions (Flanking active TSS, Enhancer, Heterochromatin, Bivalent enhancer, and Repressed Polycomb) that the Epigenomic Roadmap uses to propose the lineage of the cell lines.

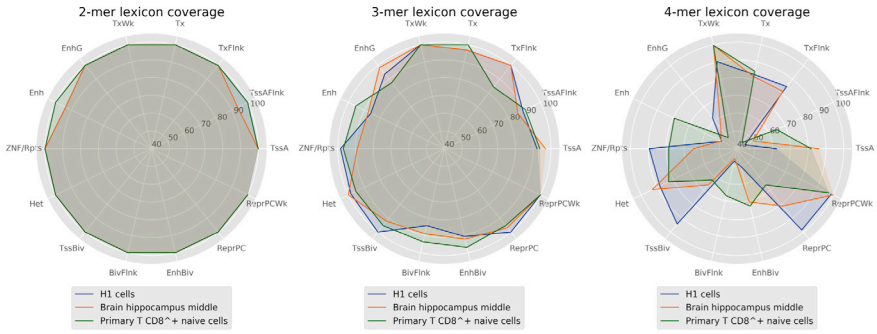


FIGURE 6. Percentage of words covered by the lexicons between 2 and 4 lengths by regulatory regions.

Figure 5 shows value distributions from 3 to 8 lexicon complexities for the three cell lines used in this work. These results present differences between the degree of degeneracy of the regulatory regions of the different cell lines. This is an interesting finding hinting that the sequences could act to establish the regulatory machinery that determines a cell line.

Moreover, another notable result is the decreasing deviation of the lexicon complexity when the length of the studied words increases. As stated earlier, a larger length of words enables producing very specific words, meanings, leading to a specialized lexicon. In this context, it is natural thinking that two sequences have to be similar in their biological functionality when sharing a highly specific word.

However, it may be a coincidence and the vocabulary may still not be sufficiently robust for encoding a biological function. Otherwise, in the case of a shorter word length, the generated vocabulary would be narrow and the set of shared words between sequences may contain the whole vocabulary (Figure 6).

This likely leads to an ambiguous code and a highly probable regulatory function. At the same time, lexicons with medium words lengths (4 to 6 nucleotides) provide enough word diversity/degeneracy degree relationship to support a robust code that may encode the regulatory function sequences and determine a cell line.

CONCLUSION

The role that non-coding regions plays in DNA sequences is fuzzy due to the diversity and apparent randomness of the sequences. This leads to the notion that these regions are a quiescent part of the genome. However, consortiums as ENCODE, and Epigenomic Roadmap have identified genome regulatory functions in the environment of this part. At the same time, these consortiums do not explore the role of sequences' composition in the determination of the corresponding regulatory function.

Our results show important differences between the lexica of sequences of regulatory regions. While the potential capacity to encode the biological function is similar for any word length, the suitable range of word lengths is between 4 and 7 nucleotides in order of providing sufficient diversity to support the robustness of a code. This is feasible since the degree of degeneracy in these lexica is high enough for the code not to be ambiguous or highly specialized, i. e. the code is robust enough and, hence, not easy to "hack". Furthermore, a broader study could identify the specific words, syntax, and the code that establishes the regulatory function in a sequence, and consequently, determines the cell line to be developed, i. a. taking into account the aspect of the noise immunity of the code [31].

ACKNOWLEDGMENTS

This work was supported by the grant project CB-256465 CONACyT Basic Science 2015.

REFERENCES

[1] Goldman AD, Landweber LF. What is a Genome? Doolittle WF, editor. PLOS Genet [Internet]. 2016 Jul 21;12(7):e1006181. Available from: http://dx.plos.org/10.1371/journal.pgen.1006181

[2] Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the human genome. Nature [Internet]. 2004 Oct 21;431(7011):931-45. Available from: http://www.nature.com/doifinder/10.1038/nature03001

[3] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature [Internet]. 2012 Sep 5;489(7414):57-74. Available from: *http://www.nature.com/doifinder/10.1038/nature11247*

[4] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature [Internet]. 2015 Feb 19;518(7539):317-30. Available from: *http://www.ncbi.nlm.nih.gov/pubmed/25693563*

[5] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types (supplementary). Nature [Internet]. 2011 May 5;473(7345):43-9. Available from: http://www.nature.com/doifinder/10.1038/nature09906

[6] Strahl BD, Allis CD. The language of covalent histone modifications. Nature [Internet]. 2000 Jan 6;403(6765):41-5. Available from: http://www.nature.com/doifinder/10.1038/47412

[7] Jenuwein T, Allis CD. Translating the histone code. Science (80-) [Internet]. 2001;293(5532):1074-80. Available from: *http://www.ncbi.nlm.nih.gov/pubmed/11498575*

[8] Turner BM. Defining an epigenetic code. Nat Cell Biol [Internet]. 2007;9(1):2-6. Available from: http://www.nature.com/doifinder/10.1038/ncb0107-2

[9] Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet [Internet]. 2016 Jun 27;17(8):487-500. Available from: http://dx.doi.org/10.1038/nrg.2016.59

[10] Gardner KE, Allis CD, Strahl BD. Operating on chromatin, a colorful language where context matters. J Mol Biol [Internet]. 2011;409(1):36-46. Available from: http://dx.doi.org/10.1016/j.jmb.2011.01.040

[11] Cohanim AB, Kashi Y, Trifonov EN. Three sequence rules for chromatin. J Biomol Struct Dyn [Internet]. 2006;23(5):559-66. Available from: https://www.tandfonline.com/doi/abs/10.1080/07391102.2006.10507081

[12] Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. Nature [Internet]. 2006;442(August):772-8. Available from: *http://www.ncbi.nlm.nih.gov/pubmed/16862119*

[13] Audit B, Vaillant C, Arnéodo A, D'Aubenton-Carafa Y, Thermes C. Wavelet Analysis of DNA Bending Profiles reveals Structural Constraints on the Evolution of Genomic Sequences. J Biol Phys [Internet]. 2004;30(1):33-81. Available from: http://link.springer.com/10.1023/B:JOBP.0000016438.86794.8e

[14] Salih B, Tripathi V, Trifonov EN. Visible periodicity of strong nucleosome DNA sequences. J Biomol Struct Dyn [Internet]. 2015 Jan 2;33(1):1-9. Available from: http://www.tandfonline.com/doi/abs/10.1080/07391102.2013.855143?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3Dpubmed

[15] Audit B, Vaillant C, Arneodo A, D'Aubenton-Carafa Y, Thermes C. Long-range Correlations between DNA Bending Sites: Relation to the Structure and Dynamics of Nucleosomes. J Mol Biol [Internet]. 2002 Mar;316(4):903-18. Available from: *http://www.ncbi.nlm.nih.gov/pubmed/11884131*

[16] Barbieri M. Code Biology [Internet]. Cham: Springer International Publishing; 2015. Available from: http://link.springer.com/10.1007/978-3-319-14535-8

[17] Barbieri M. What is code biology? Biosystems [Internet]. 2018 Feb;164:1-10. Available from: http://dx.doi.org/10.1016/j.biosystems.2017.10.005

[18] Barbieri M. The Code Paradigm. In: Code Biology [Internet]. Cham: Springer International Publishing; 2015. p. 19-34. Available from: *https://books.google.de/books/about/Code_Biology.html?id=rNp5BgAAQBAJ&pgis=1*

[19] Hofmeyr J-HS. The first Special Issue on code biology - A bird's-eye view. Biosystems [Internet]. 2018 Feb;164:11-5. Available from: http://dx.doi.org/10.1016/j.biosystems.2017.12.007

[20] Prakash K, Fournier D. Evidence for the implication of the histone code in building the genome structure. Biosystems [Internet]. 2018 Feb;164:49-59. Available from: http://dx.doi.org/10.1016/j.biosystems.2017.11.005

[21] Tsonis AA, Schultz C, Tsonis PA. Zipf's law and the structure and evolution of languages. Complexity [Internet]. 1997 May;2(5):12-3. Available from: http://doi.wiley.com/10.1002/%28SICI%291099-0526%28199705/06%292%3A5%3C12%3A%3AAID-CPLX3%3E3.0.CO%3B2-C

[22] Lestrade S. Unzipping Zipf's law. Cai Z, editor. PLoS One [Internet]. 2017 Aug 9;12(8):e0181987. Available from: http://dx.plos.org/10.1371/journal.pone.0181987

[23] Piantadosi ST. Zipf's word frequency law in natural language: A critical review and future directions. Psychon Bull Rev. 2014;21(5):1112-30.

[24] Balasubrahmanyan VK, Naranan S. Information Theory and Algorithmic Complexity: Applications to Language Discourses and DNA Sequences as Complex Systems Part II: Complexity of DNa Sequences, Analogy with Linguistic Discourses. J Quant Linguist [Internet]. 2000 Aug 1;7(2):153-83. Available from: http://dx.doi.org/10.1076/0929-6174(200008)07:02;1-z;ft153

[25] Ji S. Waves as the Symmetry Principle Underlying Cosmic, Cell, and Human Languages. Information [Internet]. 2017;8(1):24. Available from: *http://www.mdpi.com/2078-2489/8/1/24*

[26] Whitacre J, Bender A. Degeneracy: A design principle for achieving robustness and evolvability. J Theor Biol [Internet]. 2010;263(1):143-53. Available from: http://dx.doi.org/10.1016/j.jtbi.2009.11.008

[27] Mason PH. Degeneracy at Multiple Levels of Complexity. Biol Theory [Internet]. 2010 Sep 14;5(3):277-88. Available from: http://link.springer.com/10.1162/BIOT_a_00041

[28] Gammerman A, Vovk V. Kolmogorov Complexity: Sources, Theory and Applications. Comput J [Internet]. 1999 Apr 1;42(4):252-5. Available from: http://comjnl.oupjournals.org/cgi/doi/10.1093/comjnl/42.4.252

[29] Popov O, Segal DM, Trifonov EN. Linguistic complexity of protein sequences as compared to texts of human languages. Biosystems [Internet]. 1996 Jan;38(1):65-74. Available from: *http://linkinghub.elsevier.com/retrieve/pii/030326479501568X*

[30] Orlov YL, Te Boekhorst R, Abnizova II. Statistical Measures of the Structure of Genomic Sequences: Entropy, Complexity, and Position Information. J Bioinform Comput Biol [Internet]. 2006 Apr;04(02):523-36. Available from: *http://www.ncbi.nlm.nih.gov/pubmed/16819800*

[31] Fimmel E, Strüngmann L. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems [Internet]. 2018 Feb;164:186-98. Available from: **https://doi.org/10.1016/j.biosystems.2017.09.007**

SUPPLEMENTARY FIGURES

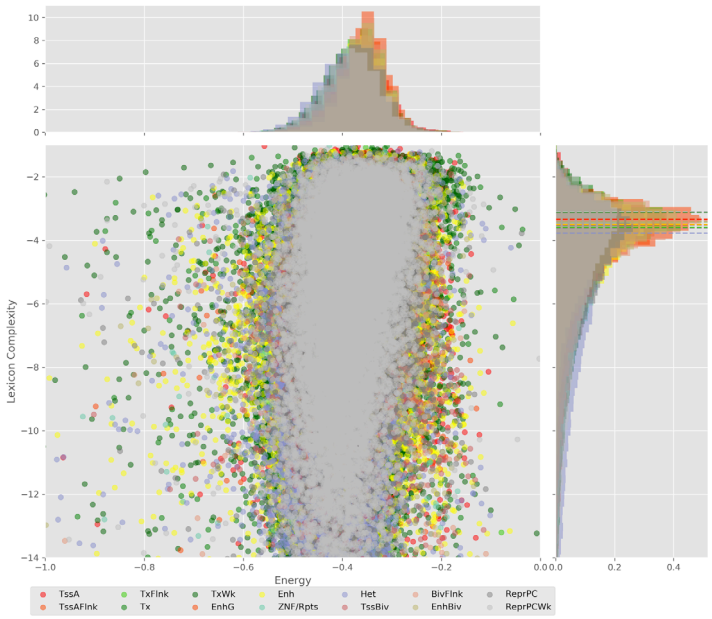


FIGURE S1. Energy against the 3-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.

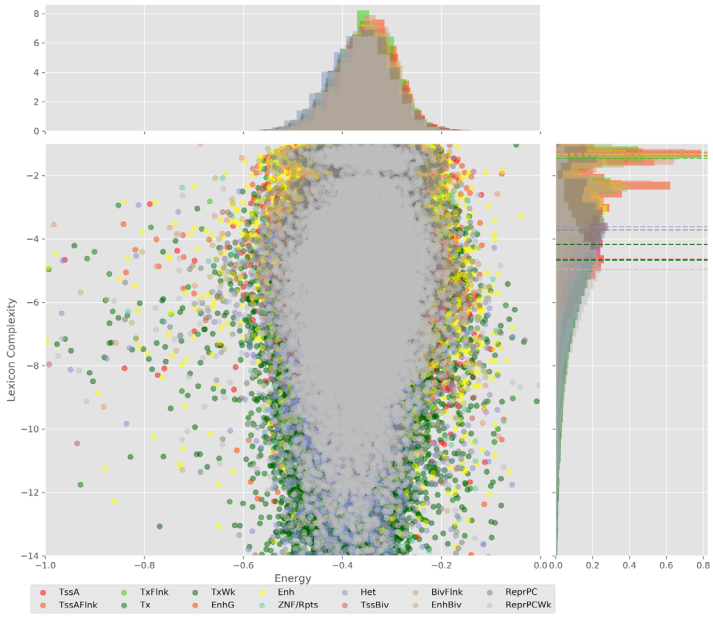


FIGURE S2. Energy against the 4-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.

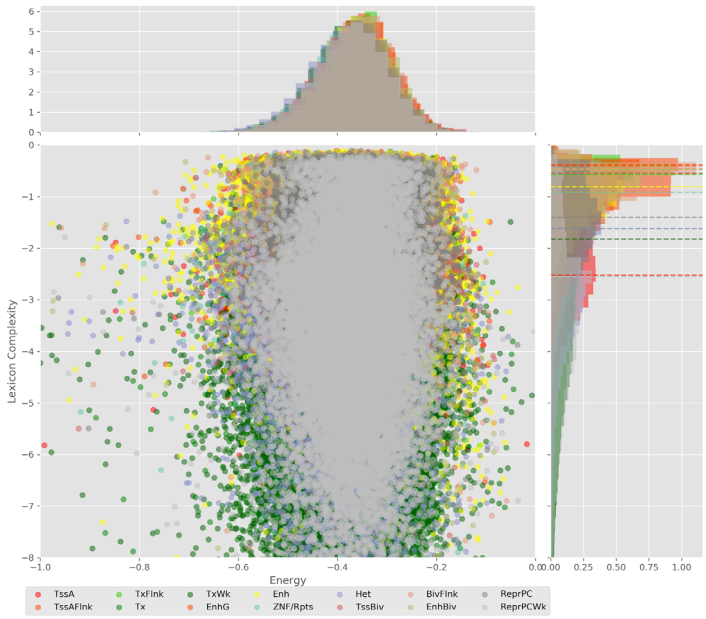


FIGURE S3. **Energy against the 5-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

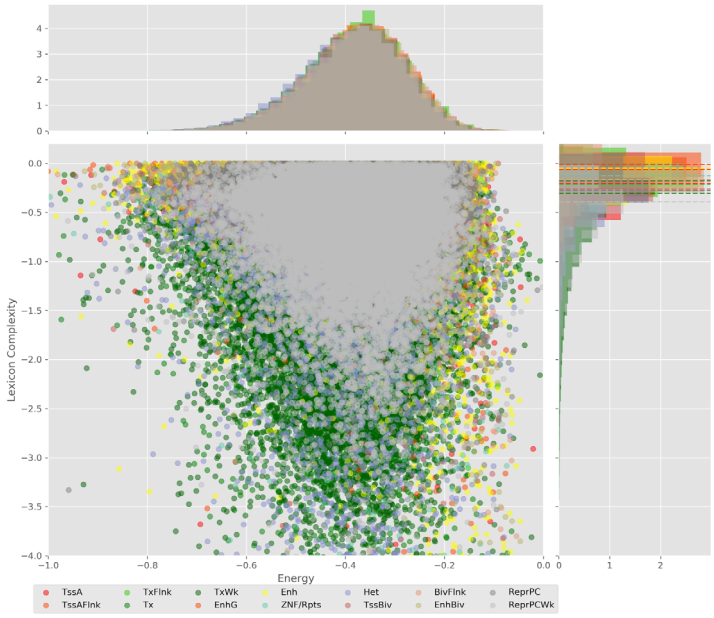


FIGURE S4. **Energy against the 7-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

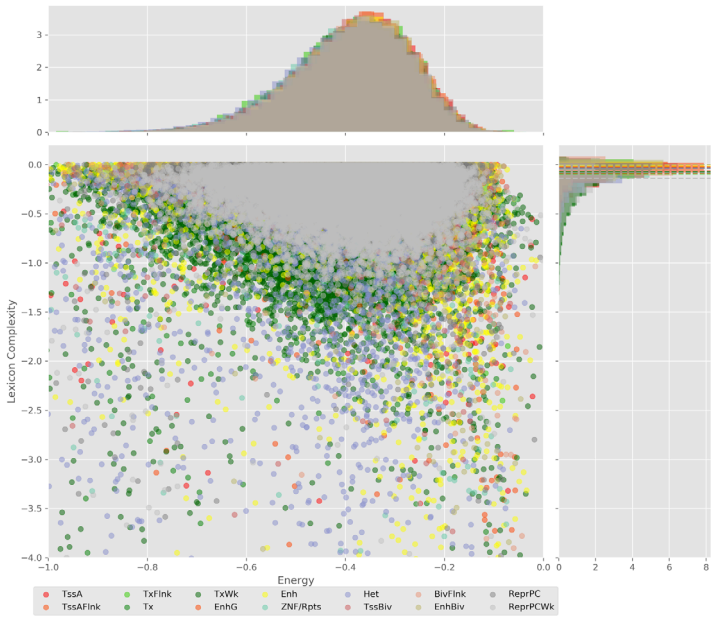


FIGURE S5. **Energy against the 8-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

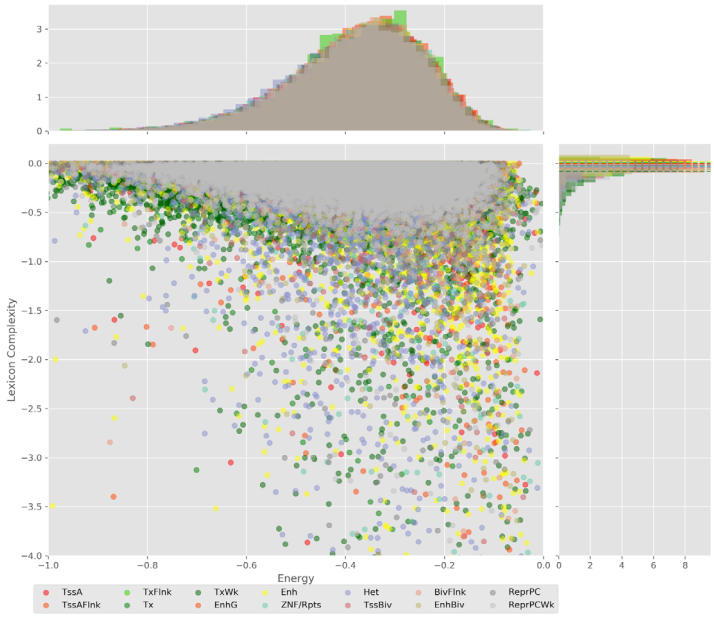


FIGURE S6. **Energy against the 9-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

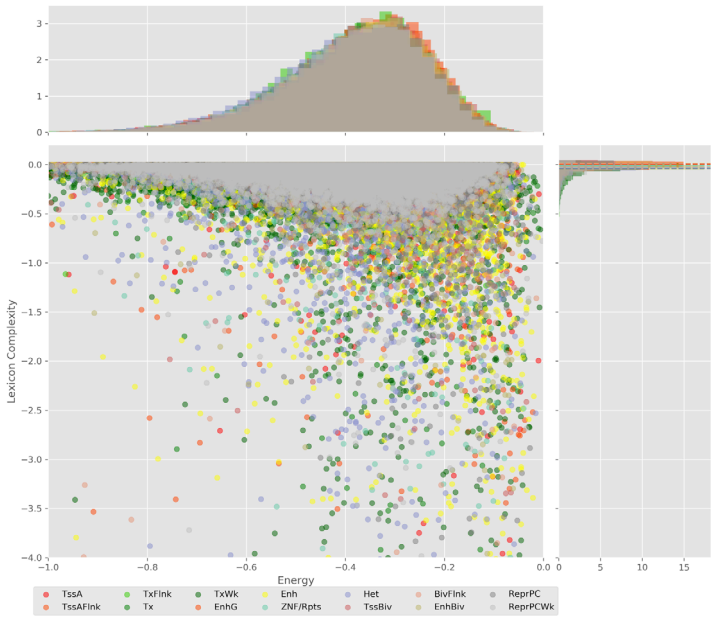


FIGURE S7. **Energy against the 10-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

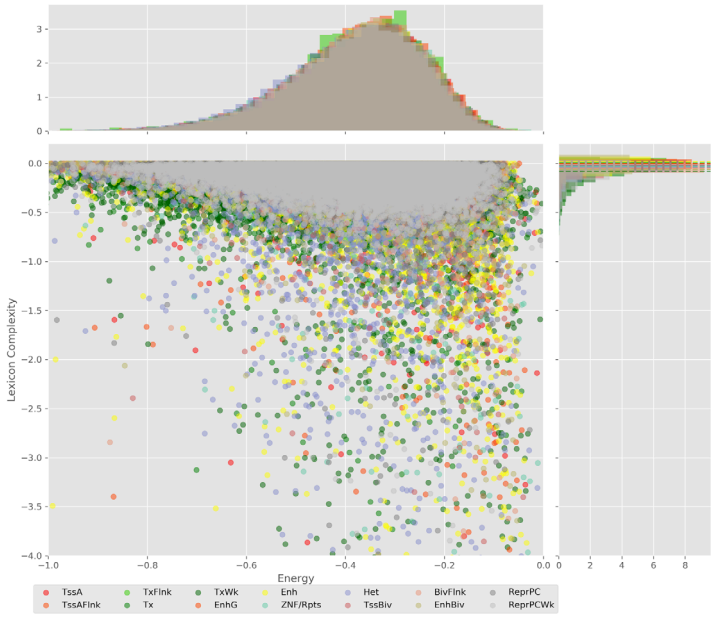


FIGURE S8. **Energy against the 11-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

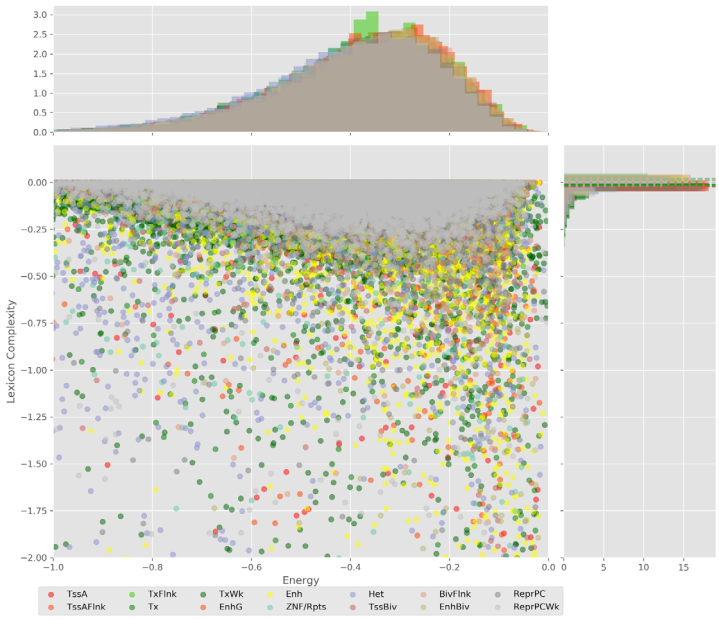


FIGURE S9. **Energy against the 12-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

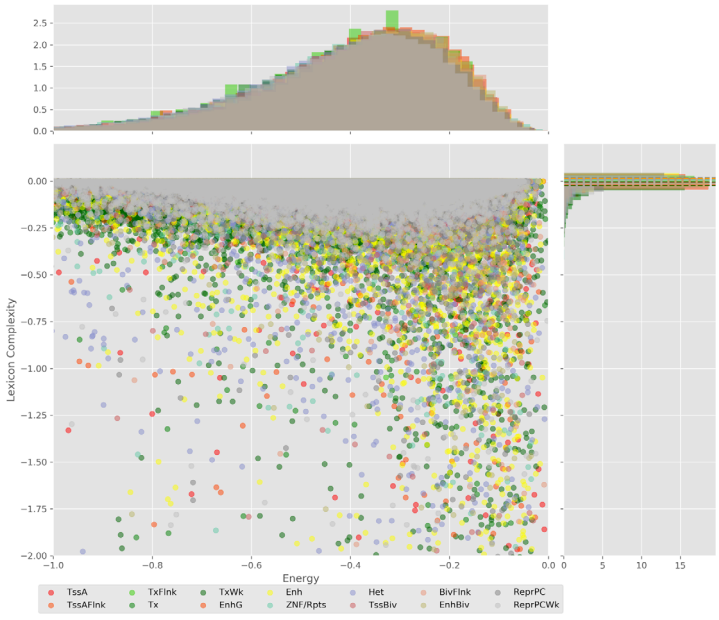


FIGURE S10. **Energy against the 13-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

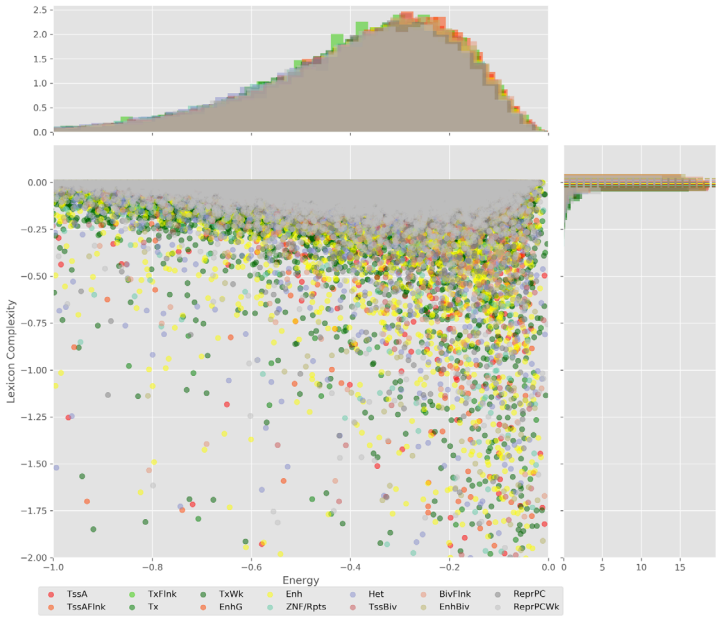


FIGURE S11. **Energy against the 14-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

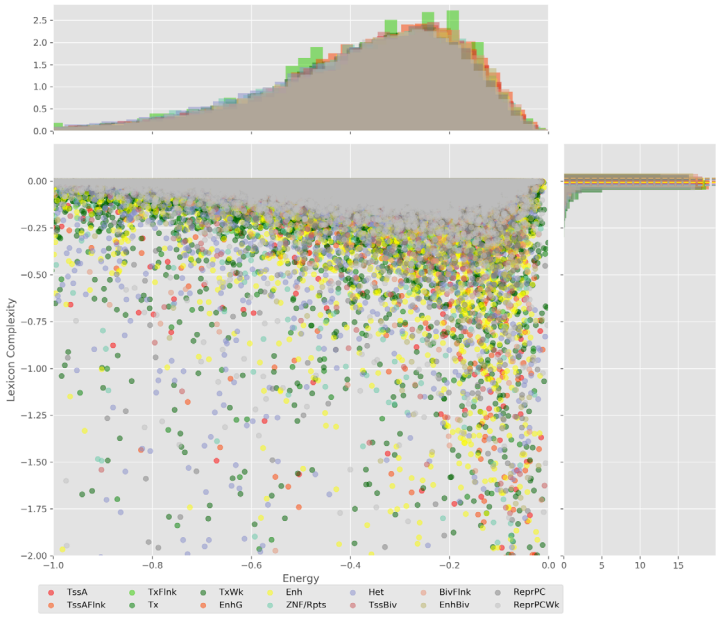


FIGURE S12. **Energy against the 15-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.**

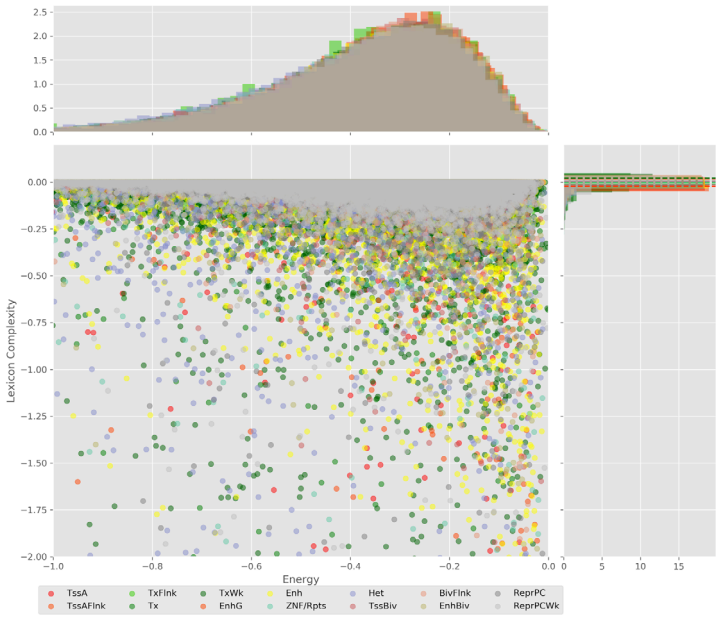


FIGURE S13. Energy against the 16-lexicon complexity (lower left) of the DNA sequences for the 14 regulatory regions of cell line H1 cells. The histograms correspond to the distribution of the energy (upper) and the distribution of the lexicon complexity (right). The dotted lines represent the most common value of the respective regulatory region.