Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Estadística

Una propuesta novedosa para la estimación de la hora de la muerte a partir de datos post mortem de expresiones de genes

Autor: *Carlota María Prieto Martínez*

Tutores: *Yolanda Larriba González, Miguel Alejandro Fernández Temprano*

Resumen

El estudio del patrón de expresiones de los genes puede ser un indicador importante del estado de las células y de problemas de salud. Los genes circadianos, asociados al ciclo sueño-vigilia, presentan un patrón de expresión rítmico u oscilatorio, y regulan funciones biológicas básicas como la respiración o digestión. Alteraciones en los patrones de expresión de estos genes están relacionadas con patologías. Debido al riesgo o coste que supone la obtención de este tipo de datos, es habitual trabajar con datos de expresión post mortem para los que el momento en el que se tomaron las muestras es desconocido.

Este trabajo propone una metodología novedosa, basada en inferencia con restricciones, para la estimación del orden temporal de expresiones de genes post mortem y su análisis a partir del ajuste de modelos paramétricos (Cosinor, FMM) y no paramétricos de señal oscilatoria. Por último, propone medidas de error y bondad de ajuste para comparar y validar los métodos utilizados.

Se obtienen resultados de interés tanto desde el punto de vista metodológico (mejor comportamiento del nuevo orden temporal propuesto) como biológico (propuesta de posibles nuevos genes cíclicos).

Abstract

Gene expression pattern analysis is a marker of the cell states and diseases. Circadian genes, related to day-night cycle, display rhythmic or up-down-up patterns and govern basic biological functions. Pattern analysis is key to identify pathologies. However, in practice, gene expression data are not easy to collect since it may suppose a risk for health or could be expensive. Hence, post-mortem expression data, for which the moment at which samples were taken is unknown, are commonly used in practice.

This work proposes a novel methodology, based or order restricted inference, to estimate the temporal order among postmortem gene expression data. Moreover, non-parametric and parametric models (Cosinor, FMM) for oscillatory signals are fitted to analyze these data. Finally, error and goodness of fit measures are used to compare and validate the methods employed.

Interesting results are obtained from the methodological point of view (better behaviour of the new proposed temporal order) and from the biological (proposal of possible new cyclic genes).

Contenido

1. Introducción

La bioestadística ha adquirido un lugar relevante en los últimos años gracias a los continuos avances en diversas áreas y campos biomédicos. Esta ciencia es una rama de la estadística que se ocupa de los problemas planteados dentro de las ciencias de la vida, como la biología o la medicina. Tan relevante es la bioestadística, que gran parte de la evidencia en salud está construida en base a ésta. (Castro, 2019)

El crecimiento de los métodos cuantitativos en las ciencias biomédicas ha hecho de esta disciplina un elemento clave en muchas áreas entre las que están los ensayos clínicos y la cronobiología, área en la que se estudian los fenómenos fisiológicos cíclicos, o ritmos biológicos, en los seres vivos. Para este estudio es fundamental la medición de las expresiones de los genes en los diferentes tejidos de los organismos vivos. Los recientes avances en biotecnología están permitiendo realizar esta medición de la expresión de los genes de forma más segura, más sencilla y barata, produciendo un mayor rendimiento de la información (Shetty, 2020). La información obtenida con estas nuevas técnicas tiene gran potencial para el estudio de su expresión y de los sistemas regulatorios. Un paso clave en estos estudios es detectar genes involucrados en diversos procesos cuya expresión presenta patrones característicos. El análisis de este tipo de datos, desde el punto de vista estadístico supone un reto, no es trivial. (Colleen, A., & Doherty , 2010)

Un gen es una unidad molecular que codifica un producto funcional específico, por ejemplo, una proteína. También son los responsables de transmitir información a la descendencia del organismo. Llamamos expresión de un gen al proceso mediante el cual la información codificada en un gen se utiliza para dirigir el montaje de una molécula de proteína. Este proceso está estrictamente regulado y permite que una célula responda a cambios de su entorno, controla la síntesis de proteínas y estas a su vez los distintos procesos biológicos. (J., y otros, 2004)

El nivel de expresión de los genes es dinámico y los patrones que estos siguen son fundamentales para entender los procesos biológicos, desde la inflamación hasta el envejecimiento. El estudio de los patrones de las expresiones de los genes se empieza a ver muy útil para diagnosticar enfermedades como el cáncer. (Roth, 2002)

En concreto, en este trabajo nos centraremos en los genes cuya expresión sigue un ritmo circadiano, conocidos como genes circadianos. El ritmo circadiano es un ciclo natural de cambios físicos, mentales y de comportamiento que experimenta un individuo en un ciclo de 24 horas. Este ritmo permite a un organismo adaptar su fisiología con anticipación entre la noche y el día provocando oscilaciones en un conjunto diverso de procesos biológicos y así poder preparar al individuo para responder a condiciones ambientales predecibles (Zhang, Lahens, Ballance, Hughes, & Hogenesch, 2014); lo que se traduce en un patrón de expresión oscilatorio en el que hay un único máximo y un único mínimo en el ciclo de expresión, que se denominará patrón up-down-up, coincidiendo el máximo de la expresión con el momento del día (ciclo) en el que el gen realiza la acción biológica. Este tipo de señales que se conocen como señal circular y cuya definición formal se establecerá más adelante, subyacen en la gran mayoría de ritmos biológicos.

La expresión de los genes es muy relevante puesto que es un indicador del estado de las células y se puede asociar con la aparición de enfermedades como la neurodegeneración, depresión, trastornos metabólicos, etc. (Liu, Gershon, & Kelsoe, 2017) Los ritmos circadianos anormales también pueden estar relacionados con la obesidad, la diabetes, la depresión, el trastorno bipolar, el trastorno afectivo estacional y los trastornos del sueño. (Zhang, Lahens, Ballance, Hughes, & Hogenesch, 2014)

Por tanto, el análisis de los patrones de las expresiones de los genes circadianos puede suponer un gran avance para la detección de algunas enfermedades. Sin embargo, aun cuando los procesos para su obtención han mejorado sensiblemente en los últimos tiempos, la obtención de los mismos sigue siendo complicada y supone un riesgo para la salud. (Valadares, Gorki, Liebold , & Hoenicka, 2017). Por ello es habitual que el análisis de expresiones de genes se efectúe a partir de expresiones post mortem, es decir, de personas ya fallecidas. Pero en este orden de circunstancias es relevante tener en cuenta que la muerte clínica, paro cardíaco, no implica el paro de las funciones biológicas y por tanto de la expresión de los genes. Esto supone, en la práctica, que la estimación del momento de la muerte sea desconocida o provenga de estimaciones imprecisas. Como ejemplo de esta cuestión, en el panel izquierdo de la Figura 1 se muestra la representación, ordenado según el orden dado por el momento de fallecimiento estimado (ToD), de la expresión del gen NPRL2, gen típicamente circadiano, que, al contrario de lo esperado, presenta un patrón bastante alejado de ser rítmico. Por el contrario, en el panel derecho se muestra otra estimación del orden temporal que se ajusta de forma más adecuada al ritmo circadiano esperado para el gen NPRL2. Por lo tanto, parece claro que esa estimación hecha de forma directa a partir del momento estimado del fallecimiento puede ser mejorable y que se requiere de un mejor procedimiento de estimación de orden temporal, previo al análisis de expresiones circadianas a partir de modelos de señales oscilatorias.

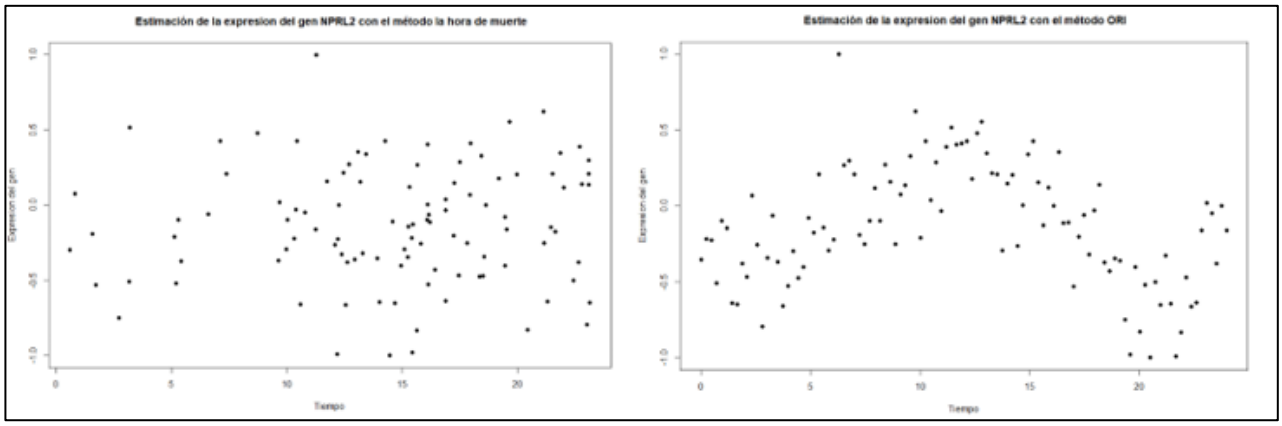


Figura 1: Comparación de dos órdenes temporales para el mismo gen NPRL2. En el panel izquierdo la estimación ha sido realizada con ToD. En el panel derecho la estimación ha sido realizada con el método ORI que se propone en este trabajo.

Así pues, uno de los objetivos de este trabajo consiste en analizar y comparar distintos métodos de estimación del orden temporal y modelos para señales oscilatorias para el análisis de datos de expresión de genes post mortem.

Hay que señalar también que, una vez ordenados los datos, existen diferentes modelos para ajustar una señal rítmica a los datos. Los modelos que se van a considerar en este trabajo, y que se describen con detalle en la sección de metodología, son el clásico modelo Cosinor (Cornelissen, 2014) y otros dos modelos más flexibles, un modelo no paramétrico específicamente diseñado en (Larriba , Rueda , Fernández , & Peddada , 2019) para el ajuste de una señal up-down-up, y un modelo paramétrico, definido en (Rueda , Larriba, & Peddada , 2019), que permite, a diferencia de Cosinor, el ajuste de modelos asimétricos y cuyos parámetros son fácilmente interpretables. Estos modelos, juntamente con las correspondientes medidas de bondad de ajuste, permitirán valorar el desempeño de los órdenes propuestos en este trabajo a la hora de describir los datos analizados.

Los datos que se van a considerar en este trabajo, y que se describirán posteriormente con más detalle, corresponden a 104 individuos sanos y 46 individuos esquizofrénicos y se han obtenido de CommonMind Consortium (CMC). Estos datos se han considerado también en (Seney , y otros, 2019) donde se estudian los posibles cambios en la ritmicidad de algunos genes debidos a la esquizofrenia. Dado que en ese trabajo se considera como punto de partida el orden dado por el momento estimado de fallecimiento, orden que se pretende mejorar, otro de los objetivos que se persigue en este TFG es el de confirmar, o no, las conclusiones obtenidas en ese otro estudio en lo que se refiere a los cambios de ritmicidad producidos por la enfermedad.

1.1 Objetivos

Los objetivos de este trabajo, que se han esbozado durante la descripción anterior, pueden enumerarse de la forma siguiente:

- Estimación del orden temporal para conjuntos de datos de expresiones de genes en los que este orden es desconocido por tratarse de datos de expresiones post mortem, mediante metodologías alternativas a la estimación directa de la hora de muerte.
- Ajuste de las expresiones de los genes en base a patrones rítmicos mediante modelos paramétricos y no paramétricos.
- Valoración global de los órdenes estimados y los ajustes obtenidos en los modelos anteriores.
- Análisis de los resultados obtenidos mediante la metodología anterior en un conjunto de datos recientemente utilizado en la literatura para estudiar la influencia de la esquizofrenia en los patrones rítmicos. Valoración de las hipótesis establecidas en la literatura a la luz de esta nueva metodología.
- Búsqueda de posibles nuevos genes con patrones rítmicos no observados en metodologías previas.

1.2 Estructura del documento

La memoria de este trabajo fin de grado se compone de los siguientes capítulos:

Metodología: En este capítulo se desarrolla la metodología utilizada en este trabajo. Primeramente, se describen los métodos de estimación del orden temporal, a continuación, los modelos que se utilizan para los ajustes de los datos y finalmente las medidas utilizadas para la valoración de los ajustes.

Datos: Se describen los conjuntos de datos utilizados, de dónde se han obtenido y los tratamientos previos que se han efectuado en los mismos.

Resultados: En este capítulo se describen, comparan y valoran los resultados obtenidos de acuerdo a los diferentes órdenes y modelos empleados en el trabajo. Asimismo, se valoran las hipótesis establecidas en la literatura sobre la influencia de la esquizofrenia en el carácter rítmico de los genes y se ofrecen hipótesis sobre nuevos posibles genes rítmicos sugeridos por la metodología.

Conclusiones: Se resumen los resultados obtenidos y se sugieren posibles líneas de desarrollo futuro a partir de dichos resultados.

Anexos: Se incluye el código R utilizado para la elaboración del trabajo.

1.3 Asignaturas relacionadas

La relación entre las técnicas utilizadas en este trabajo y las diferentes asignaturas del grado es la siguiente:

- Minería de datos: en ella se estudia el tratamiento previo que se debe realizar a los datos.
- Computación estadística: esta asignatura aporta una base sólida de R, lenguaje utilizado para la realización de este trabajo.
- Modelos Lineales y Modelos Estadísticos Avanzados: en estas asignaturas se imparten las bases para la comprensión de modelos estadísticos.
- Modelos de Investigación Operativa y Algoritmos y Computación: en estas asignaturas se estudia y se dan soluciones al TSP (Problema del viajante).

1.4 Ampliación de materia

Para poder desarrollar este trabajo fin de grado ha sido necesario el estudio de los siguientes contenidos adicionales a lo estudiado en el grado.

- Modelos de señal oscilatoria: estos modelos han sido utilizados para realizar el ajuste de las expresiones de los genes.
- Método ORI: método propuesto para la estimación del orden temporal.

2. Metodología

En este capítulo se presenta la definición de señal circular, los métodos de estimación de orden temporal, los modelos de señales oscilatorias ajustados y las métricas aplicadas en este trabajo.

Sea $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})'$ con $j = 1 \dots p$ los datos de expresión del gen j para los instantes de tiempo $i = 1 \dots n$, siendo p el número de genes y n el número de instantes de tiempo. Sea X el conjunto de datos de los p genes.

Se dice que x_j sigue un modelo de señal circular si

$$x_j = \mu + \varepsilon_j \tag{1}$$

donde μ es una señal circular tal y como se define a continuación.

Definición 1: Definición de señal circular. (Caso U<L)

Una señal μ es una señal circular si y solo si

$$\mu \in C = \bigcup_{L,U} C_{LU} \tag{2}$$

donde $L = argmin_{1\dots n}\, \mu_i$, $U = argmax_{1\dots n}\, \mu_i$, $L, U \in \{1, \dots, n\}$ y

$C_{LU} = \{\mu \in \mathbb{R}^n : \mu_1 \leq \cdots \leq \mu_U \geq \cdots \geq \mu_L \leq \cdots \leq \mu_n \leq \mu_1\}$.

Definición 2: Definición de orden circular

Se dice que una señal $\boldsymbol{\phi}$ sigue un orden Circular en el espacio Circular si y solo si

$$\phi \in C_0 = \{\phi \in [0, 2\pi)^n : \phi_1 \preceq \cdots \preceq \phi_n \preceq \phi_1\} \tag{3}$$

Donde $\preccurlyeq$ se lee como “es seguido por”. En este caso decimos que ϕ sigue un orden circular.

La siguiente fórmula presenta la equivalencia entre una señal circular en el espacio euclídeo y una señal circular en el espacio circular:

$$\phi_{ij} = T_{LU}(x_{ij}) = \begin{cases} \arcsin(x_{ij}) - \frac{\pi}{2} & \text{, cuando } i \in \{1, \dots, U\} \cup \{L, \dots, n\} \\ \frac{\pi}{2} - \arcsin(x_{ij}) & \text{, en otro caso} \end{cases} \quad (4)$$

Las señales circulares presentan patrones de expresión oscilatorio, up-down-up, es decir, son señales que crecen monótonamente hasta μ_U, después decrecen hasta μ_L para volver a crecer de nuevo. Esto se puede observar en la Figura 2. En el panel izquierdo tenemos una señal circular en el espacio euclídeo con un patrón up-down-up. En el panel derecho tenemos su equivalencia en el espacio circular, donde la señal $\boldsymbol{\phi}$ sigue un orden circular.

Los resultados anteriores se pueden escribir de forma análoga para el caso L<U.

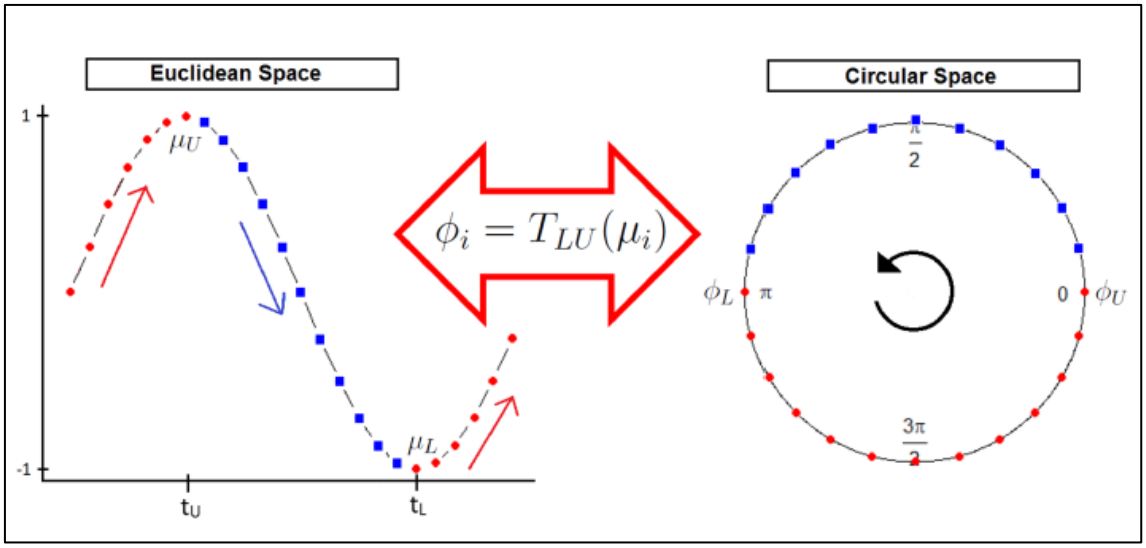


Figura 2: Imagen que muestra la equivalencia entre una señal oscilatoria en el espacio euclídeo y una señal circular en el espacio circular. Imagen obtenida de (Larriba , Rueda , Fernández , & Peddada , 2019)

2.1 Métodos de estimación del orden temporal

Sea $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})'$ con $j = 1 \dots p$ los datos de expresión del gen j para los instantes de tiempo $i = 1 \dots n$, siendo p el número de genes y n el número de instantes de tiempo. Sea X el conjunto de datos de los p genes, el problema que se quiere resolver es encontrar la permutación óptima de los instantes de tiempo, u orden temporal entre las muestras, que esté más próximo al conjunto de genes. Esta permutación óptima será por tanto aquella en la que los genes presenten la mayor ritmicidad posible. En este trabajo se consideran dos métodos para resolver este problema: el primero propone un orden temporal estimado a partir

de información adicional ToD (Time of Death, hora de muerte) y el segundo estima el orden minimizando una función de coste ORI (Order Restricted Inference, Inferencia con restricciones de orden)

2.1.1 Método naive: orden según el ToD

Esta primera estimación del orden se realiza de forma directa. Los individuos son ordenados de forma creciente según su hora de muerte estimada. Véase (Seney , y otros, 2019)

2.1.2 Método ORI (Inferencia con Restricciones de Orden)

La estimación de este segundo orden está basada en Inferencia con Restricciones de Orden. Este método da solución al problema de la estimación del orden temporal a partir del orden circular óptimo entre los instantes de tiempo del conjunto de datos. Es importante señalar que esta estimación se puede realizar con el conjunto completo de los datos o un subconjunto de ellos.

Sea Π el conjunto de todos los órdenes circulares. Para cada orden circular $o \in \Pi$ hay una señal up-down-up tal que $\mu \in C_0$. Para un orden circular dado definimos la distancia entre o y X :

$$d(o, D) = \sum_{j=1}^{P} \sum_{i=1}^{n} v_k \, (X_{ij} - X^*_{o\,ij})^2 \tag{5}$$

donde

$$X_j^* = \arg \min_{Z \in C} \sum_{i=1}^{n} (X_i - Z_i)^2 \tag{6}$$

y v_j es un peso positivo asociado con el j elemento del conjunto de datos.

Este problema es resuelto por el siguiente problema de optimización:

$$argmin_{O \in \pi} d(o, D) \tag{7}$$

Este problema es NP-hard ya que existen $\Pi = (\mathrm{n} - 1)!$ órdenes posibles. Por ello se halla una solución aproximada reformulando el Problema del Viajante (TSP) del siguiente modo:

1.- Se representa cada gen en un grafo donde los nodos son los instantes del tiempo que se han de ordenar y las aristas la distancia entre cada par de nodos.

2.- Como se dispone de diferentes genes, se calcula un grafo agregado resultado de la suma de los pesos de las aristas.

3.- Gracias a esto el problema queda reducido a recorrer todos los nodos del grafo en la menor distancia posible, pasando por todos ellos una única vez y empezando y acabando con el mismo nodo, lo que corresponde con el TSP.

4.- Finalmente se pasa del orden circular a una señal up-down-up, gracias al resultado (4).

La Figura 3 muestra de forma esquemática el proceso que se acaba de describir.

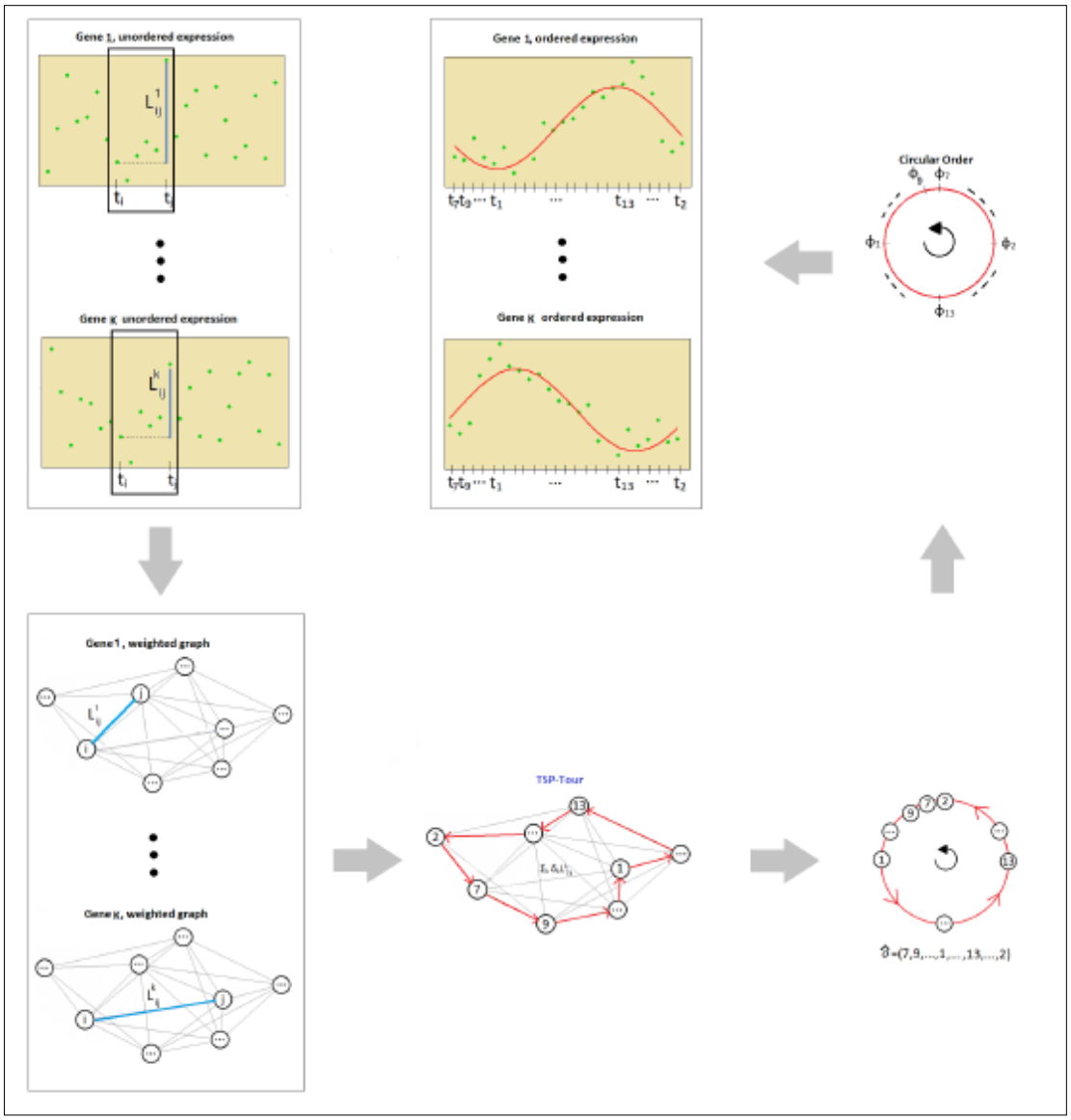


Figura 3: Imagen que representa el algoritmo utilizado para realizar la estimación del orden mediante el método ORI. Imagen obtenida de (Larriba , Rueda , Fernández , & Peddada , 2019)

2.2 Modelos de señal oscilatoria

Una vez se conoce el orden temporal, en esta sección se describen los modelos de señales oscilatorias usados para su validación. Como se comentó en la introducción, se proponen tres modelos, uno no paramétrico y dos paramétricos.

2.2.1 Modelo no paramétrico

Modelo no paramétrico de Regresión Isotónica (IR)

En el modelo no paramétrico la estimación de la señal oscilatoria se calcula mediante mínimos cuadrados

$$X^* = \arg\ \min_{Z \in C} \sum_{i=1}^{n} (X_i - Z_i)^2 \tag{8}$$

Donde *X** es el estimador de IR (Regresión Isotónica) verificando las restricciones dadas en C, véase (2), es decir, X* es una señal oscilatoria de *X* con los C pesos iguales.

Como la derivación de *X** no es trivial se utiliza un algoritmo de computación basado en los resultados teóricos del IR. Véase (Larriba , Rueda , Fernández , & Peddada , 2019):

1.- Encontrar L^* y U^*. Estos valores serán mínimos y máximos locales respectivamente.

2.- Para cada combinación posible de L^* y U^* se realizan las regresiones isotónicas crecientes y decrecientes.

3.- De entre todas estas parejas de candidatos se escoge aquella que cumpliendo (2) obtenga el menor error cuadrático medio.

2.2.2 Modelos paramétricos

Modelo Cosinor

Es el modelo más sencillo que permite la representación de una señal oscilatoria. Este modelo ajusta la onda sinusoidal que más se ajusta a los datos experimentales en tres parámetros: MESOR, amplitud y acrofase.

Definición 3: Definición del Modelo Cosinor

$$X(t) = M + A\,cos\,(\frac{2\pi t}{\tau} + \phi) + e(t) \quad (9)$$

Descripción de los parámetros

- M es el MESOR, es decir, el valor medio de la oscilación. Es el valor alrededor del cual oscila la señal.
- A es la amplitud de la onda. La amplitud es la variación máxima del desplazamiento de la variable en un ciclo.
- ϕ es la acrofase. La acrofase es el tiempo transcurrido entre el momento de referencia y el punto más alto del ciclo.
- τ el periodo. El periodo es la duración de un ciclo.
- $e(t)$ es el término de error que sigue una distribución $N(0\,,\,\sigma^2)$ y es independiente. (Garcés)

En la Figura 4 se pueden ver gráficamente los parámetros del modelo Cosinor.

La mayor limitación de este modelo es que únicamente describe señales simétricas, ya que, en realidad, todos sus parámetros se podrían considerar de localización (M y ϕ) y de escala (A).

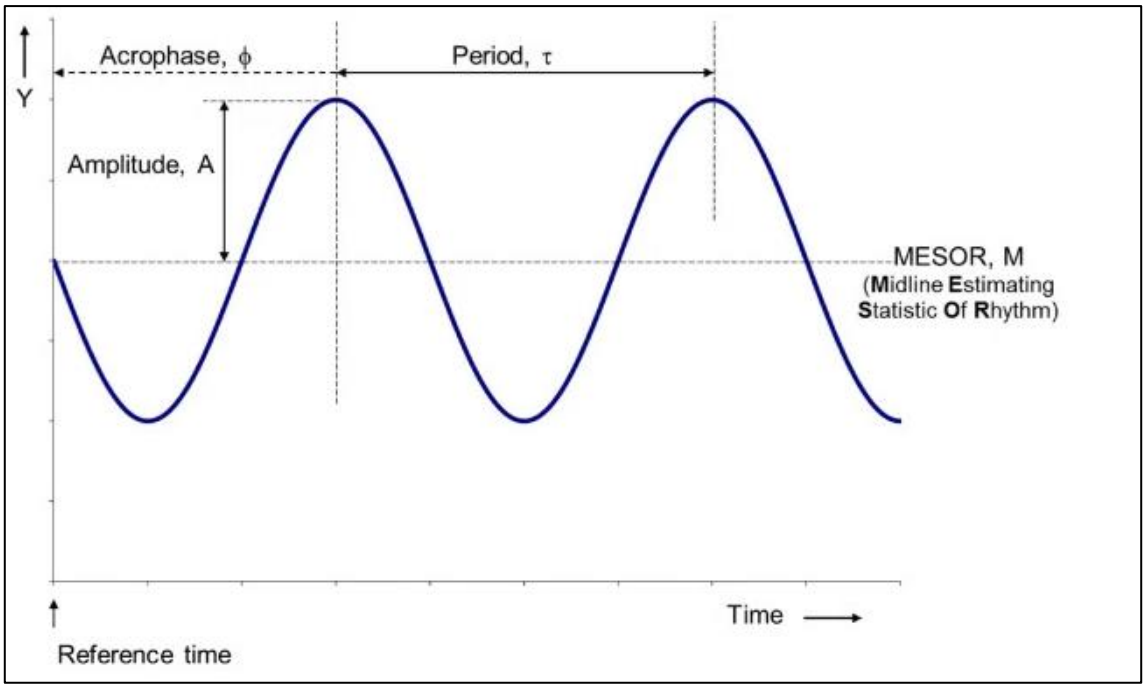


Figura 4: Parámetros del modelo Cosinor. Imagen obtenida de (Cornelissen, 2014)

Modelo FMM

Es un modelo que describe patrones rítmicos en sistemas oscilatorios capaz de describir patrones no sinusoidales. Esto se debe a que emplea una transformación de Möbius como función de enlace, frente la función de enlace lineal que emplea por ejemplo Cosinor. (Rueda , Larriba, & Peddada , 2019)

Definición 4: Definición del modelo FMM

$$X(t) = \mu\,(t) + \ e(t) = M \ + \ A\ cos\ (\ \phi(t))\ + \ e(t) \qquad (10)$$

donde

1.- $M \in \mathbb{R}\,, A \in \mathbb{R}^{+}$

2.- $\phi(t) = \beta + 2atan\ (\omega \tan\ (\frac{t-\alpha}{2}))) \ donde\ \alpha, \beta \in [0, 2\pi]\ ,\ \omega \in [0, 1]$

3.- $e(t)\ \sim\ N(\ 0\ , \sigma^2)$

Descripción de los parámetros del modelo:

- M es el intercept del modelo.
- A es la amplitud de la onda.
- α es un parámetro de localización de fase.
- β es un parámetro de forma relacionado con la asimetría. La onda es simétrica cuando $\beta = 0$ y $\beta = \pi$ y es asimétrica cuando toma un valor intermedio.
- ω es un parámetro de forma asociado al apuntamiento de la señal. Si $\omega = 0$ la onda presenta picos extremos y si $\omega = 1$ la onda es sinusoidal.
- $e(t)$ es el término de error que sigue una distribución $N(0\,, \sigma^2)$ y es independiente.

Los parámetros α, β y ω definen la fase del modelo. En la Figura 5 podemos ver la influencia de β y ω , los parámetros que describen la forma de la señal.

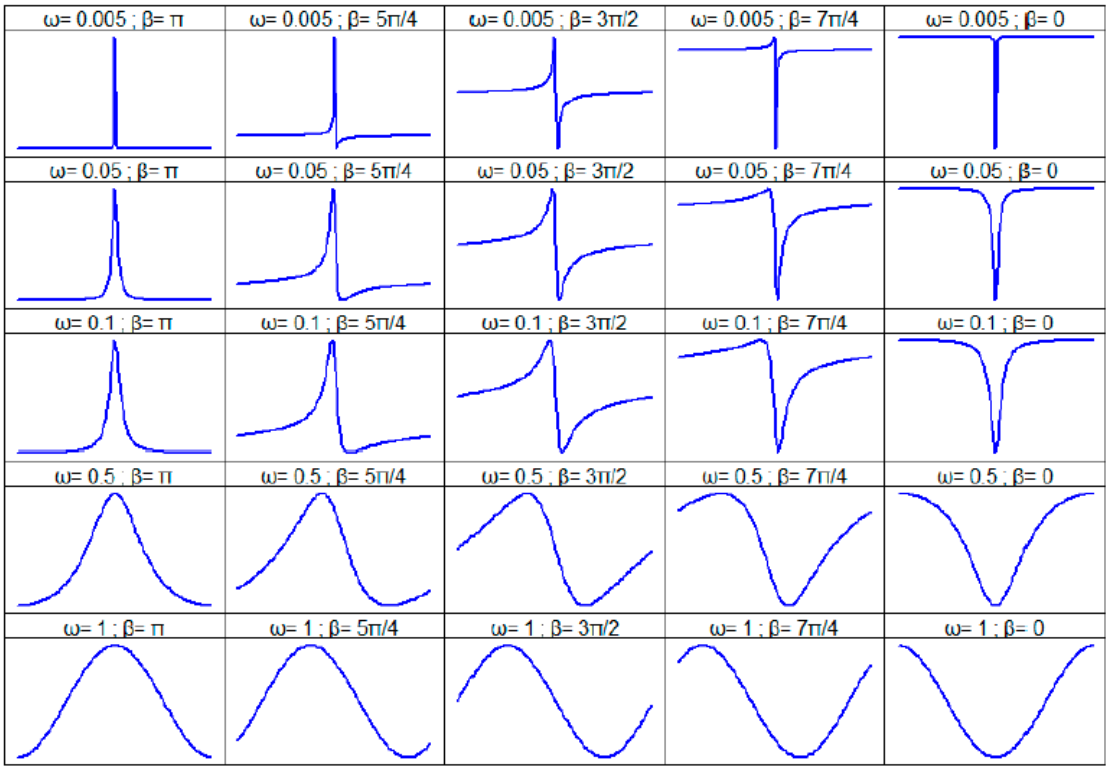


Figura 5: Influencia de los parámetros β ω en el modelo FMM con M=0, A=1 y α =0. Imagen obtenida de (Pérez, 2020)

Otros parámetros importantes del modelo son los picos y sus tiempos que se calculan:

$$t_U = \alpha + 2\text{atan}\left(\frac{1}{\omega}\tan(-\beta/2)\right) \tag{11}$$

$$t_L = \alpha + 2\text{atan}\left(\frac{1}{\omega}\tan\left(\frac{\pi-\beta)}{2}\right)\right) \tag{12}$$

Y los valores de la señal en esos puntos son:

$$Z_U = M + A \tag{13}$$

$$Z_L = M - A \tag{14}$$

Relación entre el modelo Cosinor y el modelo FMM

El modelo Cosinor es un caso particular del modelo FMM. Esto sucede cuando $\omega = 1$ donde $\varphi = \beta - \alpha$. Esto se puede ver de manera directa:

$$\phi(t) = \beta + 2\arctan\left(\omega\tan\left(\frac{t-\alpha}{2}\right)\right) = \beta + 2\arctan\left(\tan\left(\frac{t-\alpha}{2}\right)\right) = t + \beta - \alpha \quad (15)$$

2.3 Medidas de calidad de los modelos

Por último, se presentan las distintas medidas utilizadas para la validación y comparación de los distintos métodos de estimación del orden y modelos de señales oscilatorias ajustados usados en el trabajo.

Medida de error

Para los modelos como medida del error se ha calculado el MSE (Error Cuadrático Medio).

$$\mathrm{MSE_i} = \frac{\sum_{i=1}^{n}(\widehat{X}_i - X_i)^2}{n} \quad (16)$$

$$MSE = \frac{\sum_{i=1}^{P} MSE_i}{p} \quad (17)$$

Donde $\widehat{X}_i$ es el valor ajustado por el modelo para los datos de la expresión del gen i, X_i los datos de la expresión del gen i.

Medida de bondad de ajuste

Y como medida de bondad de ajuste se ha calculado el R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2}{\sum_{i=1}^{n}(Xi - \bar{X})^2} \quad (18)$$

Donde $\widehat{X}_i$ es el valor ajustado por el modelo para los datos de la expresión del gen i, X_i los datos de la expresión del gen i, $\bar{X}$ es el valor medio.

3. Datos

Al tratarse de pacientes, los datos están anonimizados, y cada individuo tiene asociado un identificador, de forma que, para conseguir relacionar un individuo con los valores de la expresión de sus genes, conocer la hora de su muerte (ToD) y saber si es control o esquizofrénico se deben cruzar tales identificadores en los ficheros correspondientes. Una vez efectuado este tratamiento se tienen dos conjuntos de datos, con la expresión de los genes y el ToD, uno para individuos controles y otro para individuos esquizofrénicos.

El conjunto de datos de individuos control está formado por 104 observaciones de las cuales, como puede verse en la Tabla 1, el 78% (81) son hombres y el 22% (23) son mujeres. El 83% (86) de la totalidad de las observaciones pertenecen a personas de raza blanca, el 16% (17) a personas de raza negra y el 1% (1) a personas asiáticas. El 59% (61) de las observaciones vienen de PIT (Pittsburgh) y el 41% (43) de MSSM (Mt. Sinai School of Medicine). Estos individuos tienen una edad media de 48.4 años.

Control							
Sexo		Raza			Lugar		Edad
Hombre	Mujer	Blanca	Negra	Asiática	PIT	MSSM	
78% (81)	22% (23)	83% (86)	16% (17)	1% (1)	59% (61)	41% (43)	48.4

Tabla 1: Distribución de los individuos control por sexo, raza, lugar y edad.

Por otro lado, el conjunto de datos de pacientes esquizofrénicos está formado por 46 individuos. Como puede verse en la Tabla 2, el 70% (32) son hombres y el 30% (14) son mujeres. El 74% (34) son de raza blanca y el 26% (12) son de raza negra. El 48% (22) proviene de PIT y el 52% (22) de MSSM. Su edad media es de 50.1 años.

Esquizofrénico							
Sexo		Raza			Lugar		Edad
Hombre	Mujer	Blanca	Negra	Asiática	PIT	MSSM	
70% (32)	30% (14)	74% (34)	26% (12)	0% (0)	48% (22)	52% (22)	50.1

Tabla 2: Distribución de los individuos esquizofrénicos por sexo, raza, lugar y edad

Todos estos individuos han sido seleccionados en base a tres criterios recogidos en el trabajo (Seney , y otros, 2019):

1. Sujetos para los que la estimación del ToD es conocida y su deceso se desencadenó de forma repentina.
2. Sujetos menores de 65 años.
3. Sujetos con un intervalo post mortem inferior a las 30 horas.

De todos estos individuos se recoge la expresión de 13914 genes en su hora de muerte. En la práctica se trabaja con dos matrices: de individuos sanos o control, $M_{13914x104}$ con 13914 filas de genes y 104 columnas de individuos y de individuos esquizofrénicos $SCZ_{13914x46}$ con 13914 filas de genes y 46 columnas de individuos.

Tratamiento de los datos

Se precisa de dos rescalados sobre los datos para permitir la comparación simultánea de modelos y órdenes, así como para adecuarse a los requisitos de los modelos ajustados.

El primero de ellos se realiza sobre los datos de las expresiones de los genes para poder compararlos, es un escalado al intervalo $[-1, 1]$. La transformación utilizada para este escalado en el eje de las ordenadas es la siguiente:

$$datoEscalado = \frac{\mathrm{y} - \mathrm{ymin}}{\mathrm{ymax} - \mathrm{ymin}} \tag{19}$$

En segundo lugar, se realiza un escalado de los tiempos ToD, es decir, en el eje de abscisas. Los datos originales vienen dados en el intervalo $[-6, 18]$, tiempos habituales en la escala Zeitgeber. Sin embargo, tanto el modelo Cosinor como el modelo FMM requieren que estos tiempos vengan dados en el intervalo $[0, 2\pi]$.

Para este segundo escalado se ha aplicado la siguiente fórmula:

$$datoEscalado = 2\pi \frac{\mathrm{x} - \mathrm{xmin}}{\mathrm{xmax} - \mathrm{xmin}} = \frac{\mathrm{x} + 6}{\frac{12}{\pi}} \tag{20}$$

Nótese que el orden ORI que se ha descrito anteriormente no proporciona un instante de tiempo exacto para cada uno de los individuos, sino que solamente proporciona una ordenación entre los mismos. En consecuencia, para poder

estimar los modelos paramétricos, se establecen dos conjuntos equiespaciados de datos en el intervalo $[0, 2\pi]$, teniendo en cuenta los tamaños muestrales de cada conjunto de datos. Concretamente, para los individuos control se toman 104 valores equiespaciados en el intervalo $[0, 2\pi]$, mientras que para los esquizofrénicos se toman 46 en ese mismo intervalo.

4. Resultados

Una vez obtenidos los datos de los individuos se comienza con su procesamiento y análisis. En esta sección mostraremos los principales resultados recogidos en la sección de objetivos del trabajo.

Generación de los órdenes

A continuación, se describe cómo se han obtenido las estimaciones de los órdenes ToD y ORI propuestos en el trabajo.

Para calcular el orden ToD se ordenan los individuos según su hora de muerte de forma ascendente.

A partir del método ORI generaremos dos órdenes. El primero de ellos se estimará a partir de un subconjunto de genes, que llamaremos *genes core*. Se dice que un gen es *core* si las proteínas que sintetiza son esenciales en la generación y regulación de ritmos circadianos. Los *genes core*, presentan patrones de expresión rítmicos. Para el segundo de los órdenes se emplean todos los genes de los datos.

Para los individuos control los *genes core* utilizados para calcular el orden son: ARNTL, NPAS2, CLOCK, NFIL3, CRY1, NR1D1, BHLHE41, NR1D2, DBP, CIART, PER1, PER3, TEF, HLF, CRY2, PER2; Esta selección contiene *genes core* muy conocidos en la biología, está basado en (Wu, 2020).

Sin embargo, para los individuos esquizofrénicos el subconjunto de genes utilizado es diferente ya que la biología circadiana en esta patología es distinta, en el sentido de que genes rítmicos en pacientes sanos pueden dejar de serlo en pacientes esquizofrénicos y genes que presentan ritmicidad en esquizofrénicos pueden dejar de presentarla en sanos. Para estos el conjunto de genes *core* seleccionado es: CIART, WNT10B, LAMB3, OPRL1, CYB561, HDAC8, NIM1K, DUBR, KRT17P1, EBP, PGBD2, CTSK, ZBTB22, NPRL2, IFT122, NFATC4, RNF112, VOPP1, NIT1, USF1. Esta selección de genes se ha hecho en base a los genes que presentan más ritmicidad para los individuos esquizofrénicos de acuerdo con (Seney , y otros, 2019)

Finalmente, se calcula tanto para individuos control como para esquizofrénicos el orden ORI con todos los genes.

Ajuste de los modelos

Una vez obtenidos los órdenes se procede al análisis de los datos de expresión de los genes con el modelo no paramétrico (NP), el modelo Cosinor y el modelo FMM.

Para ello, debe tenerse en cuenta que el modelo Cosinor y FMM necesitan que los instantes temporales estén en el intervalo $[0, 2\pi]$. En el caso del orden ToD deberemos escalar las horas con la fórmula (20).

Para los órdenes ORI se requiere que los vectores de instantes temporales mencionados al final de la sección anterior estén también en el intervalo $[0, 2\pi]$.

4.1 Resultados globales

En esta sección se van a considerar los resultados globales obtenidos para cada uno de los tres órdenes y modelos considerados, teniendo en cuenta todos los genes disponibles, tanto para el grupo de control como para los pacientes de esquizofrenia. Para las valoraciones que se hacen en este apartado se ha considerado como medida de comparación el MSE puesto que permite una mejor valoración de la globalidad de los resultados.

Las Tablas 3 y 4 contienen los valores globales de MSE para los conjuntos de datos de control y de pacientes esquizofrénicos respectivamente. En cada una de las tablas están los valores para los tres órdenes considerados (ORI, ORI Reducido y ToD) y los tres modelos que se han ajustado con esos órdenes (NP, FMM y Cosinor). Puede verse en las tablas que en todos los casos el modelo NP tiene un MSE menor que los otros dos modelos. Esto es consecuencia de que este modelo no paramétrico impone menos restricciones a los ajustes. Además, para el modelo FMM siempre se obtienen valores de MSE menores que para el modelo Cosinor como consecuencia de que el modelo Cosinor es un caso particular del modelo FMM. Es claro entonces que estos resultados eran esperables como consecuencia de la definición de cada uno de los modelos.

Control MSE			
	ORI	**ORI Reducido**	**ToD**
NP	0.0645	0.0827	0.1164
FMM	0.0937	0.1118	0.1343
Cosinor	0.1056	0.1217	0.1455

Tabla 3 : MSE para los datos control para los tres órdenes valorados y los tres modelos considerados.

Esquizofrénicos MSE			
	ORI	**ORI Reducido**	**ToD**
NP	0.0636	0.0937	0.1247
FMM	0.1052	0.1329	0.1564
Cosinor	0.1251	0.1599	0.1864

Tabla 4: MSE para los datos de esquizofrénicos para los tres órdenes valorados y los tres modelos considerados.

Una segunda observación, de más interés que la anterior, es que el orden ToD es globalmente peor, en lo que se refiere al MSE, que los otros dos como puede observarse para ambos conjuntos de datos y para todos los modelos. Este resultado también era esperable puesto que los órdenes ORI se han construido con el objetivo de minimizar una función de error relacionada con el MSE del modelo NP. De todos hay que notar que, a diferencia de lo que ocurre con los modelos, no había seguridad total en este punto puesto que no se resuelve el problema directo de minimización sino uno relacionado con él mediante el TSP.

También es interesante comentar las diferencias observadas en los resultados de los órdenes ORI y ORI Reducido. Se ve en las tablas como el MSE para el orden ORI es inferior al del ORI Reducido. Esto puede ser consecuencia de que el orden ORI está establecido a partir del conjunto completo de genes mientras que ORI Reducido está generado a partir de los *genes core,* cuyo carácter rítmico es bien conocido en la literatura. Esto quiere decir que el orden ORI no es necesariamente mejor que el ORI Reducido puesto que al utilizar todos los genes en la generación del orden se puede estar produciendo un efecto de "sobreajuste" y como consecuencia resultar que los *genes core*, con patrones de expresión claramente rítmicos, no aparezcan como tal en el orden ORI. Este es un posible efecto que se estudiará posteriormente para decidir entre ambos órdenes ORI.

Por último, se observa que en general los resultados para los pacientes esquizofrénicos son, en términos de MSE, peores que para los controles. Esta observación parece estar en concordancia con las hipótesis establecidas en (Seney , y otros, 2019) que sugieren una pérdida de ritmicidad en los pacientes esquizofrénicos.

4.2 Resultados para los *genes core*

En esta sección se describen los resultados obtenidos para los *genes core* de los dos conjuntos de datos utilizados (control y esquizofrénicos) para los distintos órdenes y modelos considerados en este trabajo. Para la comparación de los modelos se ha utilizado el R^2 como medida de bondad de ajuste de cada uno de los modelos considerados ya que, de este modo, se actúa como, por ejemplo, en los modelos de regresión y las comparaciones son más intuitivas.

La Tabla 5 contiene los resultados relativos a los *genes core* para el grupo de control. Los genes están ordenados en orden decreciente considerando el valor obtenido para R^2 en el orden ORI Reducido para el modelo FMM. Puede observarse, en primer lugar, que los valores de R^2 correspondientes al orden ORI Reducido son mejores (excepto para CRY2 y NPAS2) que los correspondientes al orden ORI. Esta observación confirma que el orden ORI Reducido ajusta mejor los *genes core* que, como ya se ha comentado, son

aquellos cuya ritmicidad es conocida en la literatura y nos permite afirmar que parece conveniente utilizar el orden ORI Reducido en el resto de las conclusiones en lugar del orden ORI generado utilizando todos los genes disponibles en el conjunto de datos.

En cuanto al orden ToD como ya se comentó en los resultados globales, salvo un par de excepciones muy puntuales (genes PER2 y ARNTL para el modelo Cosinor), los resultados de los ajustes son peores que los obtenidos por el orden ORI Reducido.

Genes Core Control R^2									
	ORI			ORI Reducido			ToD		
Gen	NP	FMM	Cosinor	NP	FMM	Cosinor	NP	FMM	Cosinor
NR1D1	0,625	0,417	0,392	0,871	0,684	0,653	0,391	0,225	0,208
CIART	0,177	0,074	0,030	0,827	0,676	0,541	0,585	0,494	0,479
PER1	0,637	0,431	0,420	0,850	0,654	0,613	0,438	0,288	0,282
PER3	0,588	0,423	0,337	0,793	0,620	0,523	0,413	0,275	0,251
DBP	0,561	0,386	0,379	0,746	0,596	0,562	0,387	0,267	0,236
BHLHE41	0,620	0,448	0,253	0,759	0,552	0,530	0,301	0,122	0,054
NR1D2	0,549	0,332	0,231	0,687	0,454	0,350	0,333	0,178	0,167
CRY2	0,652	0,438	0,384	0,643	0,431	0,341	0,387	0,196	0,187
CRY1	0,402	0,223	0,152	0,560	0,389	0,355	0,420	0,276	0,225
NPAS2	0,748	0,537	0,462	0,534	0,361	0,132	0,204	0,069	0,010
PER2	0,521	0,338	0,307	0,546	0,331	0,169	0,413	0,262	0,216
HLF	0,541	0,325	0,239	0,548	0,328	0,309	0,199	0,086	0,020
TEF	0,382	0,204	0,184	0,506	0,327	0,303	0,396	0,217	0,174
NFIL3	0,434	0,241	0,069	0,609	0,303	0,195	0,390	0,252	0,181
CLOCK	0,756	0,522	0,396	0,540	0,290	0,267	0,205	0,114	0,018
ARNTL	0,295	0,137	0,007	0,476	0,289	0,178	0,417	0,242	0,236

Tabla 5: Valores de R^2 para los ajustes de los genes core en el grupo de control para los diferentes órdenes y modelos considerados.

A continuación, se presentan tres ejemplos de *genes core* del grupo de control para los que se ve con cierta claridad como el orden ORI Reducido proporciona ajustes mejores que el orden ToD utilizado en (Seney , y otros, 2019). El ajuste proporcionado por el modelo NP está en verde, el de Cosinor en rojo y el que arroja FMM en azul.

Es destacable mencionar que los tres genes que se incluyen como ejemplo, PER3, CRY2 y CLOCK, juegan un papel decisivo en la estructura molecular del reloj circadiano en mamíferos, ubicado en la base del núcleo supraquiasmático. A nivel molecular, las interacciones entre estos genes permiten "orquestar"/sincronizar y regular las distintas funciones biológicas que se suceden a lo largo del día en los distintos tejidos y órganos (Panda , y otros, 2002).

Gen CLOCK

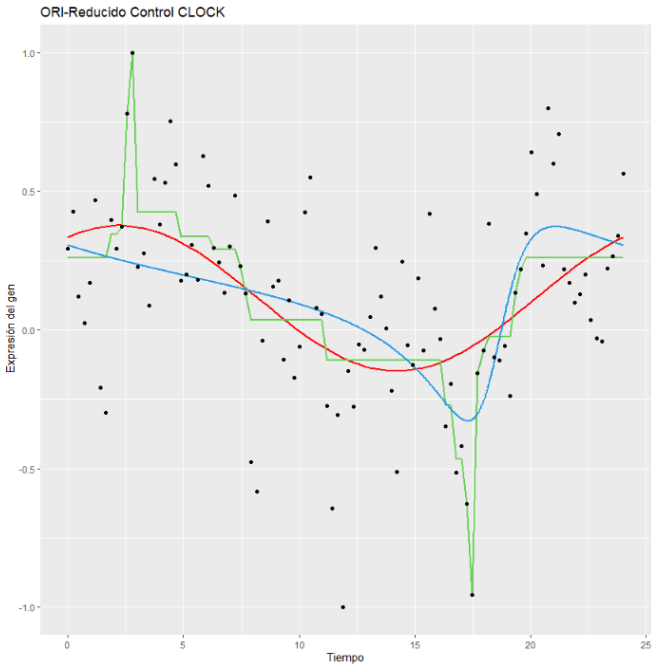


Figura 6: Gen CLOCK con el orden ORI Reducido.

ORI Reducido

Modelo	R^2
NP	0,540
FMM	0,290
Cosinor	0,267

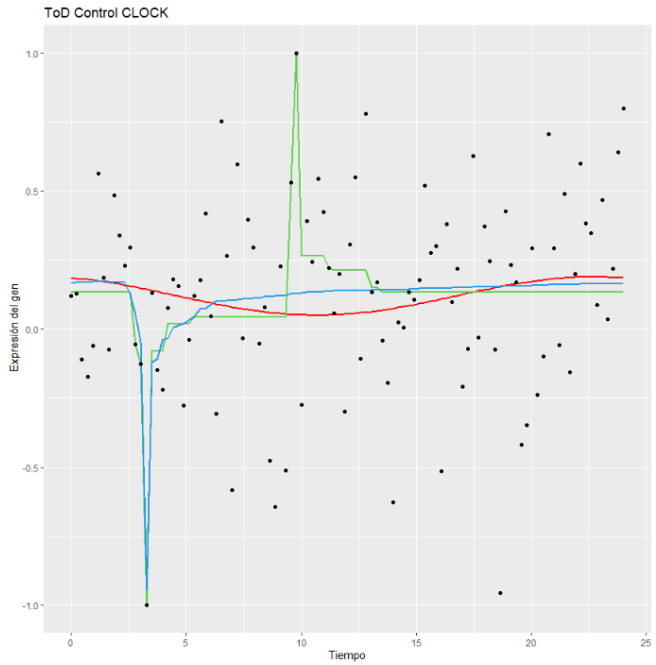


Figura 7 : Gen CLOCK con el orden ToD.

ToD

Modelo	R^2
NP	0,205
FMM	0,114
Cosinor	0,018

Gen PER3

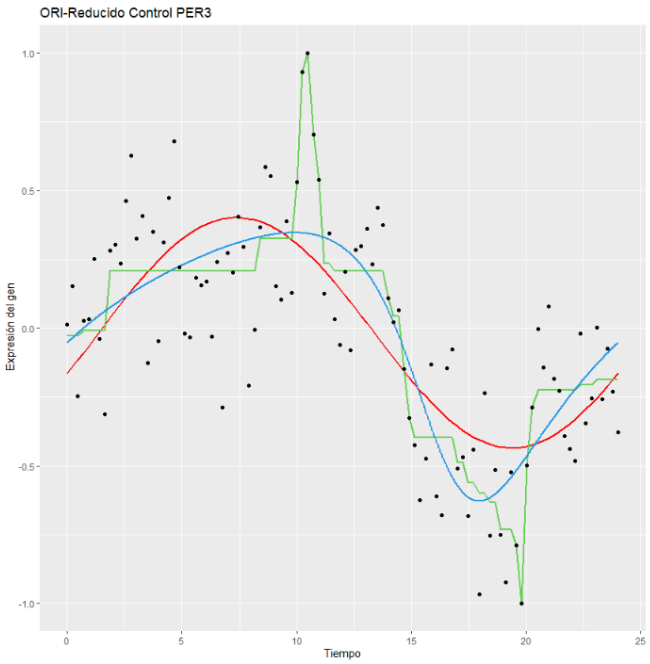


Figura 8: Gen PER3 con el orden ORI Reducido.

ORI Reducido

Modelo	R^2
NP	0,793
FMM	0,620
Cosinor	0,523

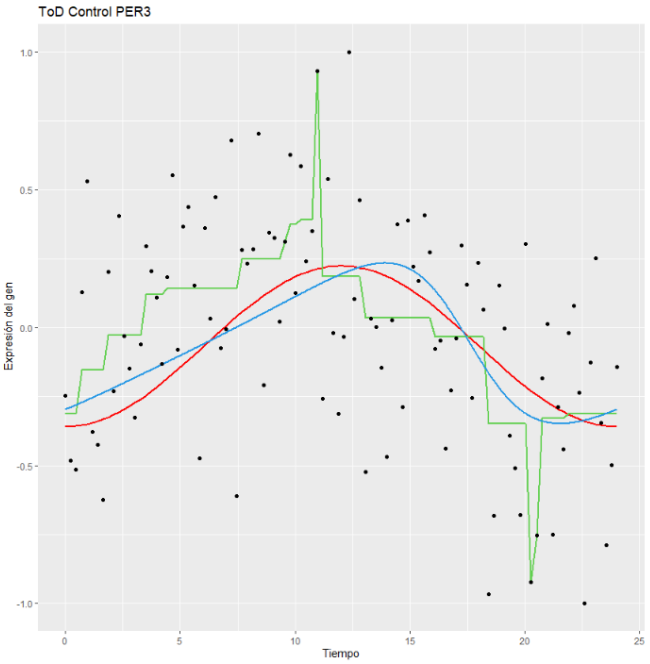


Figura 9: Gen PER3 con el orden ToD.

ToD

Modelo	R^2
NP	0,413
FMM	0,275
Cosinor	0,251

Gen CRY2

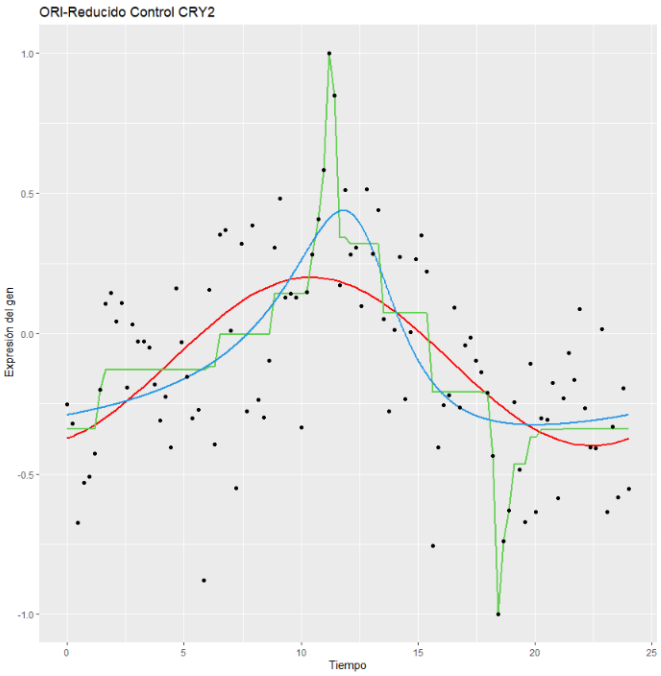


Figura 10: Gen CRY2 con el orden ORI Reducido.

ORI Reducido

Modelo	R^2
NP	0,643
FMM	0,431
Cosinor	0,341

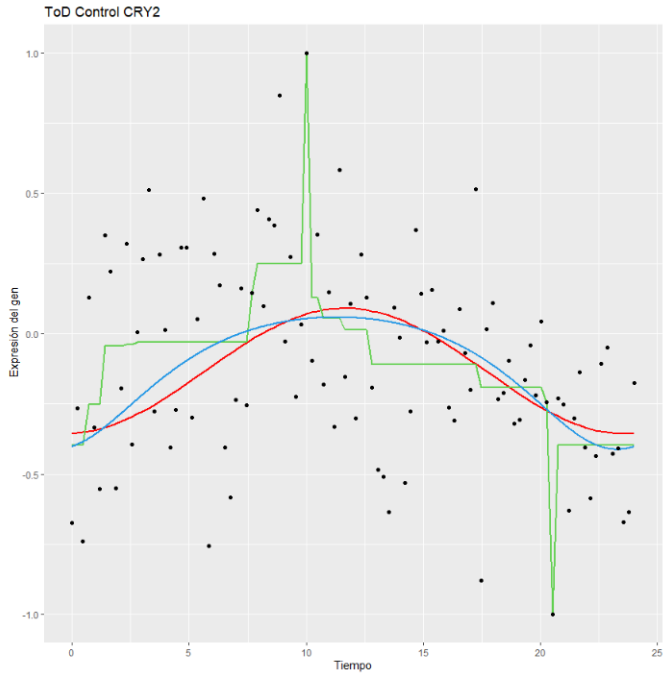


Figura 11: Gen CRY2 con el orden ToD.

ToD

Modelo	R^2
NP	0,387
FMM	0,196
Cosinor	0,187

La Tabla 6 contiene los resultados relativos a los *genes core* para el grupo de pacientes de esquizofrenia. Al igual que en la Tabla 5 los genes están ordenados en orden decreciente considerando el valor obtenido para R2 en el orden ORI Reducido para el modelo FMM. Las conclusiones en cuanto a la preferencia

entre órdenes son similares a las obtenidas en el grupo de control, lo que reafirma la preferencia por el orden ORI Reducido.

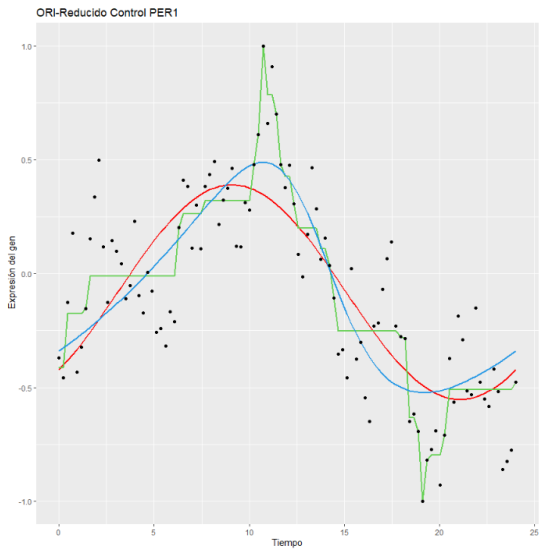
Genes Core SCZ R^2									
	ORI			ORI Reducido			ToD		
Gen	NP	FMM	Cosinor	NP	FMM	Cosinor	NP	FMM	Cosinor
USF1	0,834	0,609	0,580	0,853	0,653	0,417	0,541	0,334	0,237
NPRL2	0,784	0,620	0,607	0,786	0,652	0,506	0,516	0,310	0,252
ZBTB22	0,748	0,444	0,364	0,880	0,644	0,503	0,530	0,326	0,254
LAMB3	0,611	0,385	0,246	0,848	0,615	0,478	0,578	0,318	0,317
OPRL1	0,523	0,257	0,184	0,852	0,615	0,555	0,589	0,411	0,310
KRT17P1	0,659	0,462	0,218	0,957	0,599	0,493	0,691	0,503	0,268
NIM1K	0,660	0,441	0,433	0,751	0,587	0,397	0,503	0,343	0,282
EBP	0,789	0,626	0,582	0,808	0,584	0,425	0,526	0,386	0,260
WNT10B	0,496	0,279	0,133	0,858	0,580	0,483	0,651	0,348	0,327
CIART	0,559	0,294	0,216	0,807	0,566	0,471	0,699	0,523	0,467
CYB561	0,521	0,287	0,121	0,752	0,563	0,315	0,602	0,395	0,295
RNF112	0,464	0,225	0,138	0,744	0,560	0,406	0,455	0,258	0,249
NIT1	0,616	0,421	0,399	0,705	0,502	0,443	0,438	0,329	0,241
IFT122	0,584	0,351	0,331	0,662	0,481	0,432	0,439	0,267	0,250
HDAC8	0,558	0,296	0,202	0,668	0,445	0,226	0,522	0,300	0,287
DUBR	0,704	0,491	0,332	0,679	0,437	0,232	0,615	0,372	0,280
VOPP1	0,348	0,221	0,043	0,628	0,422	0,330	0,504	0,335	0,242
CTSK	0,640	0,354	0,302	0,616	0,320	0,252	0,543	0,275	0,254
NFATC4	0,388	0,196	0,120	0,534	0,310	0,168	0,519	0,301	0,249
PGBD2	0,280	0,159	0,029	0,314	0,213	0,048	0,545	0,279	0,257

Tabla 6: Valores de R^2 para los ajustes de los genes core en el grupo de pacientes de esquizofrenia para los diferentes órdenes y modelos considerados.

A continuación, se ofrecen los ajustes de algunos *genes core* en el grupo de control, a modo de ejemplo para poder valorar la hipótesis de pérdida de ritmicidad en los pacientes de esquizofrenia en algunos genes que se hace en (Seney , y otros, 2019).

Gen PER1
Control

ORI Reducido



ToD

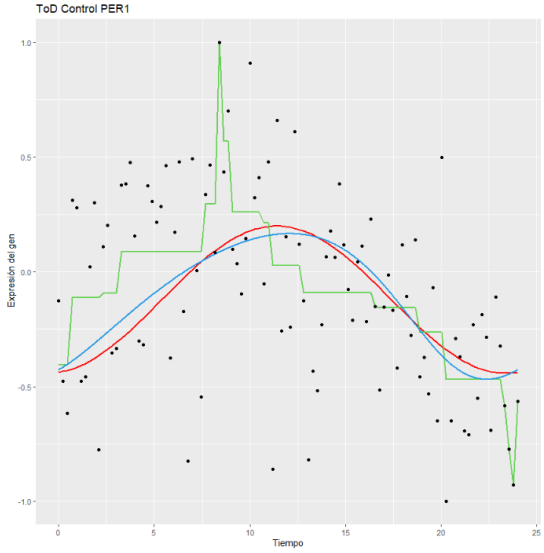
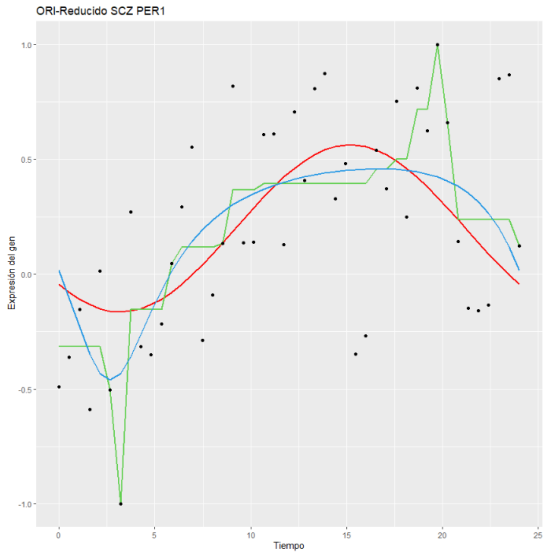


Figura 12: Gen PER con el orden ORI Reducido y con el orden ToD para individuos control.

Esquizofrenia

ORI Reducido



ToD

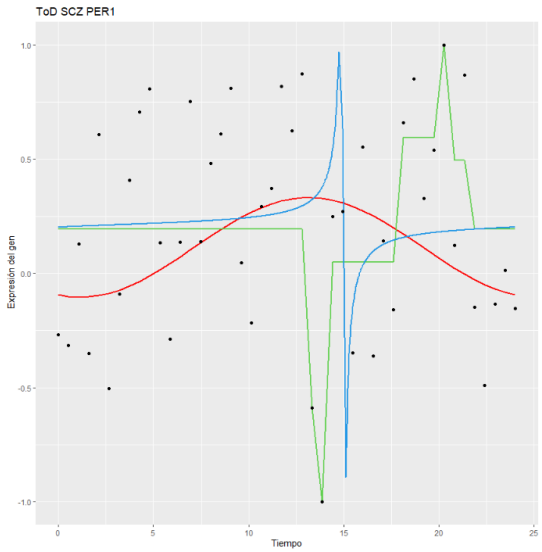


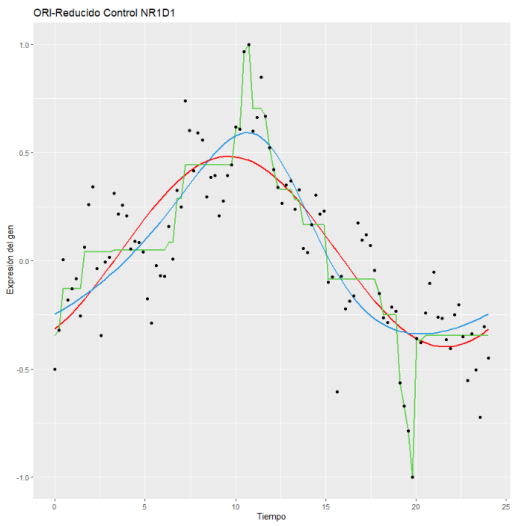
Figura 13: Gen PER con el orden ORI Reducido y con el orden ToD para individuos esquizofrénicos.

Para este gen parece haber pérdida de ritmicidad en los pacientes de esquizofrenia si se utiliza el ToD y el modelo Cosinor, como se hace en (Seney , y otros, 2019) puesto que el valor de R^2 baja desde 0.282 a 0.079. Sin embargo, si se considera el orden ORI Reducido los R^2 correspondientes al modelo

Cosinor son 0.613 y 0.294 y los correspondientes al modelo FMM son 0.654 y 0.373, para control y esquizofrénicos respectivamente.

Gen NR1D1
Control

ORI Reducido



ToD

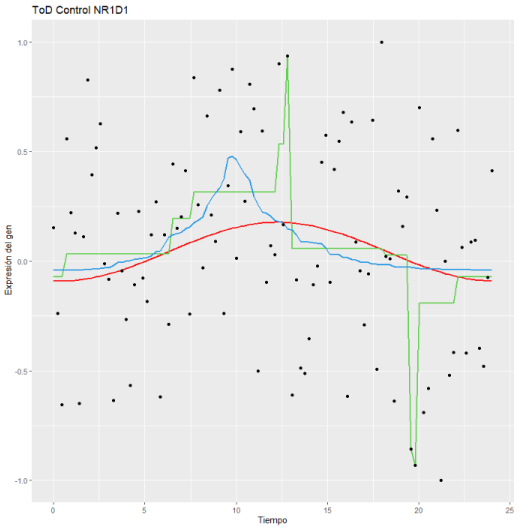
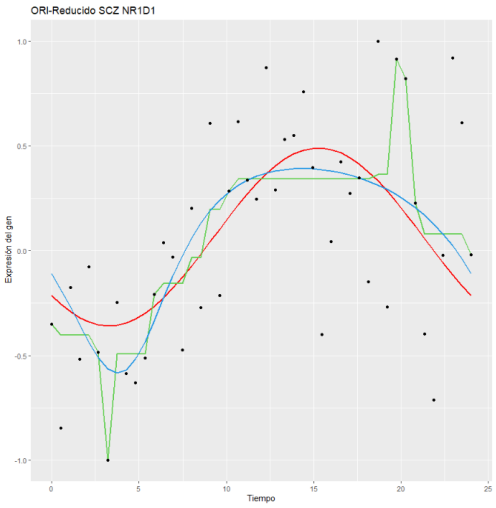


Figura 14: Gen NR1D1 con el orden ORI Reducido y con el orden ToD para individuos control

Esquizofrenia

ORI Reducido



ToD

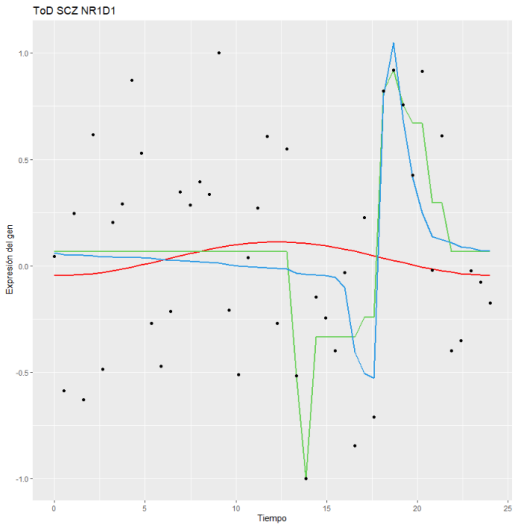


Figura 15: Gen NR1D1 con el orden ORI Reducido y con el orden ToD para individuos esquizofrénicos.

Al igual que en el gen anterior, para este gen parece haber pérdida de ritmicidad en los pacientes de esquizofrenia, si se utiliza el ToD y el modelo Cosinor, como se hace en (Seney , y otros, 2019) puesto que el valor de R^2 baja desde 0.208 a

0.009. Sin embargo, si se considera el orden ORI Reducido los R^2 correspondientes al modelo Cosinor son 0.653 y 0.346 y los correspondientes al modelo FMM son 0.684 y 0.400, para control y esquizofrénicos respectivamente.

Estas observaciones hacen pensar que la hipótesis de pérdida de ritmicidad debería ser estudiada con más detenimiento puesto que, como se observa en estos genes, esa presunta pérdida de ritmicidad puede ser debida a una deficiente estimación del orden de las observaciones.

Los siguientes gráficos permiten valorar los cambios que aparecen en los parámetros del modelo FMM de los *genes core* del conjunto de control, cuando se considera el orden ORI Reducido al comparar este conjunto de genes entre los pacientes control y los pacientes con esquizofrenia. La Tabla 7 contiene los códigos utilizados en los gráficos siguientes para identificar los *genes core* del grupo de control.

Equivalencia gen – letra Genes core control	
Gen	**Letra**
ARNTL	A
NPAS2	B
CLOCK	C
NFIL3	D
CRY1	E
NR1D1	F
BHLHE41	G
NR1D2	H
DBP	I
CIART	J
PER1	K
PER3	L
TEF	M
HLF	N
CRY2	O
PER2	P

Tabla 7: Identificación de los genes core del grupo de control

En la Figura 16 se consideran los parámetros ω y β del modelo FMM. Se observan dos grupos de genes en el grupo de esquizofrenia. Para unos el nuevo modelo FMM tiene un valor de ω cercano a 0 y un valor de β alto y para el otro los valores de β son notablemente más bajos que los del grupo control. En otras palabras, parece haberse producido un desplazamiento hacia los ejes del gráfico.

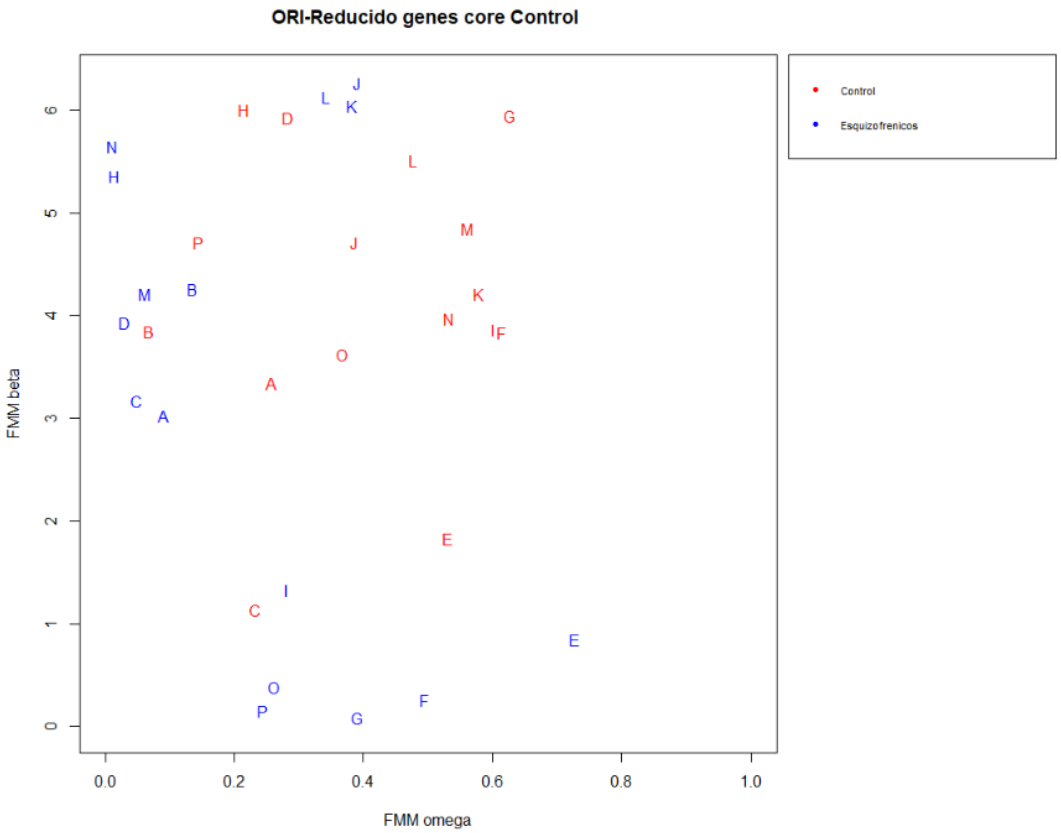


Figura 16: Parámetros ω y β del modelo FMM para los genes core del grupo de control en el conjunto de control y en el de pacientes con esquizofrenia.

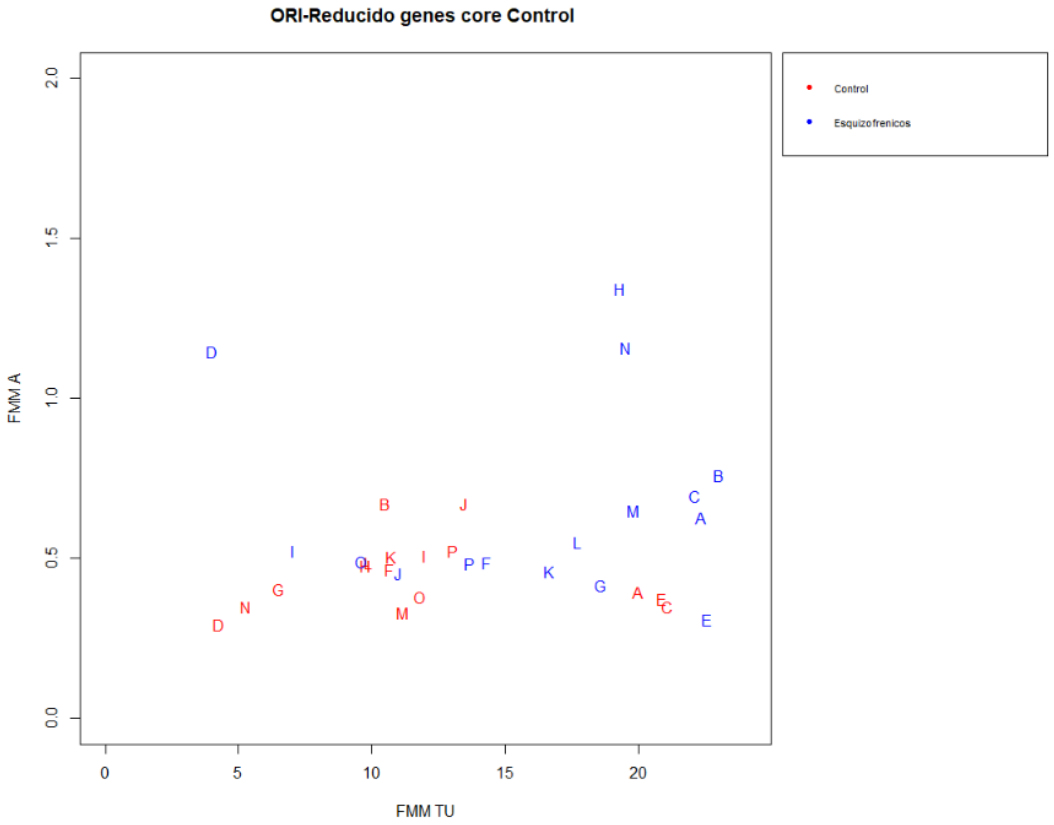


Figura 17: Parámetros A del modelo FMM y valor de t_U estimado con este mismo modelo para los genes core del grupo de control en el conjunto de control y en el de pacientes con esquizofrenia.

La Figura 17 muestra los valores del parámetro A (amplitud) del modelo FMM contra los valores estimados con este mismo modelo para t_U (momento de máxima expresión del gen) para las mismas condiciones descritas para la Figura 16. Se observa como los valores de A son, en general, mayores o iguales para el conjunto de pacientes con esquizofrenia y que también cambian los valores estimados para t_U siendo, en general, más bajos para el grupo de los pacientes control.

La Figura 18 permite la comparación entre el grupo de control y el de pacientes con esquizofrenia, de los valores del parámetro ω y de R^2 para el modelo FMM para los *genes core* del conjunto de datos de control. Se observa como los valores de R^2 son en general más bajos para los pacientes de esquizofrenia.

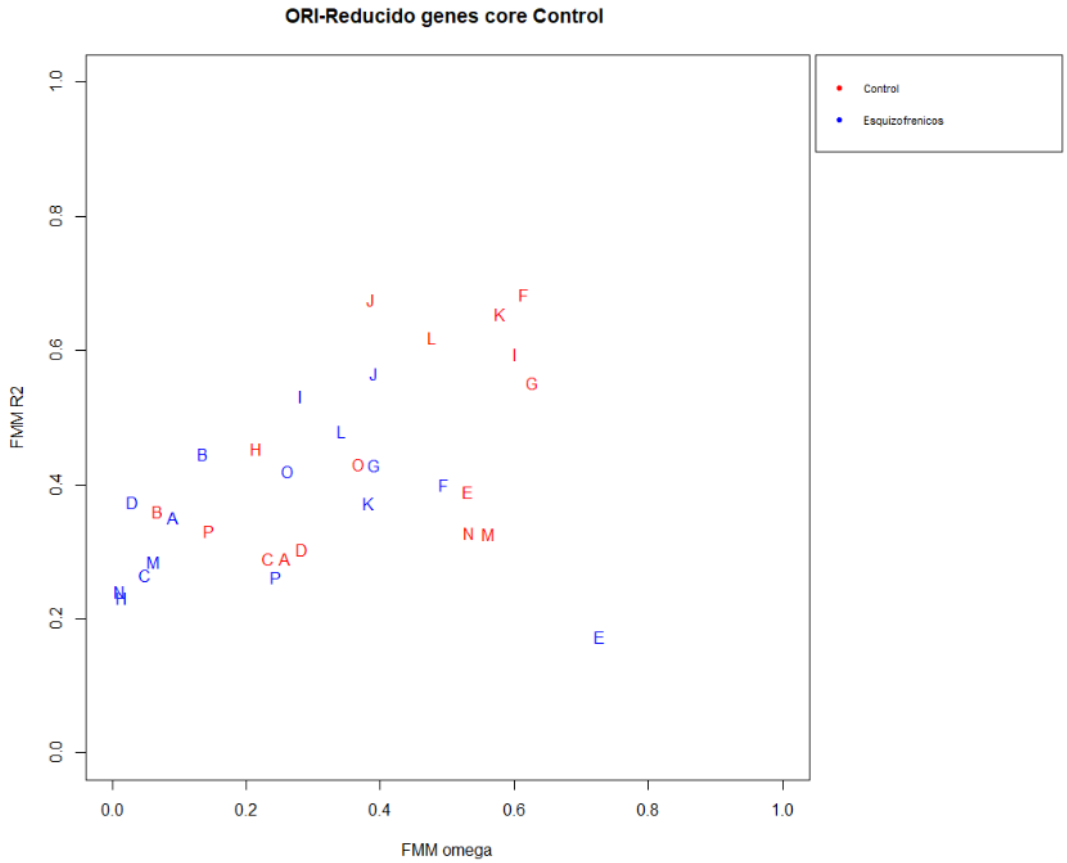


Figura 18: Parámetro ω y valor de R^2 del modelo FMM para los genes core *del grupo de control en el conjunto de control y en el de pacientes con esquizofrenia.*

Finalmente, la Figura 19 permite la comparación entre el grupo de control y los pacientes esquizofrénicos de los valores de R^2 del conjunto de los *genes core* del grupo de control. Se observa una disminución en general de este valor en el grupo de los pacientes esquizofrénicos ya que la mayor parte de los valores están por encima de la diagonal del gráfico. Esta disminución es esperable puesto que los genes que se están considerando son los *genes core* del grupo de control.

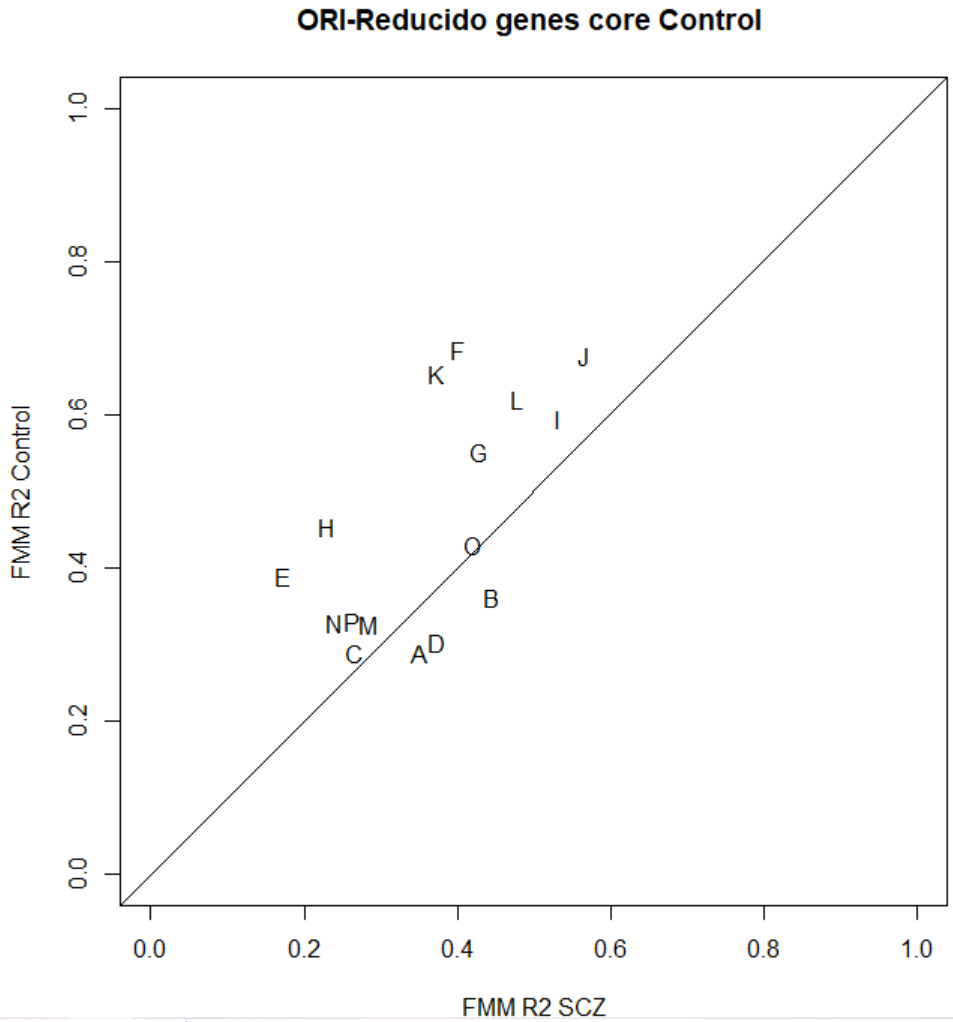


Figura 19: Valores de R^2 del modelo FMM para los genes core del grupo de control en el conjunto de control y en el de pacientes con esquizofrenia.

A continuación, se ofrecen el mismo tipo de gráficos anteriores para los *genes core* del grupo de esquizofrenia con el objetivo de valorar los cambios que aparecen en estos genes entre los grupos de control y de pacientes esquizofrénicos. La Tabla 8 contiene la identificación de los *genes core* del grupo de pacientes esquizofrénicos utilizada en los gráficos que se ofrecen.

Equivalencia gen – letra Genes core esquizofrenia	
Gen	**Letra**
CIART	a
WNT10B	b
LAMB3	c
OPRL1	d
CYB561	e
HDAC8	f
NIM1K	g
DUBR	h
KRT17P1	i
EBP	j
PGBD2	k
CTSK	l
ZBTB22	m
NPRL2	n
IFT122	o
NFATC4	p
RNF112	q
VOPP1	r
NIT1	s
USF1	t

Tabla 8: Identificación de los genes core del grupo de esquizofrenia

En la Figura 20 se consideran los parámetros ω y β del modelo FMM para los *genes core* del conjunto de pacientes con esquizofrenia. En este caso no se observan tendencias claras, con los genes para el grupo de esquizofrenia ocupando en general la parte central del gráfico.

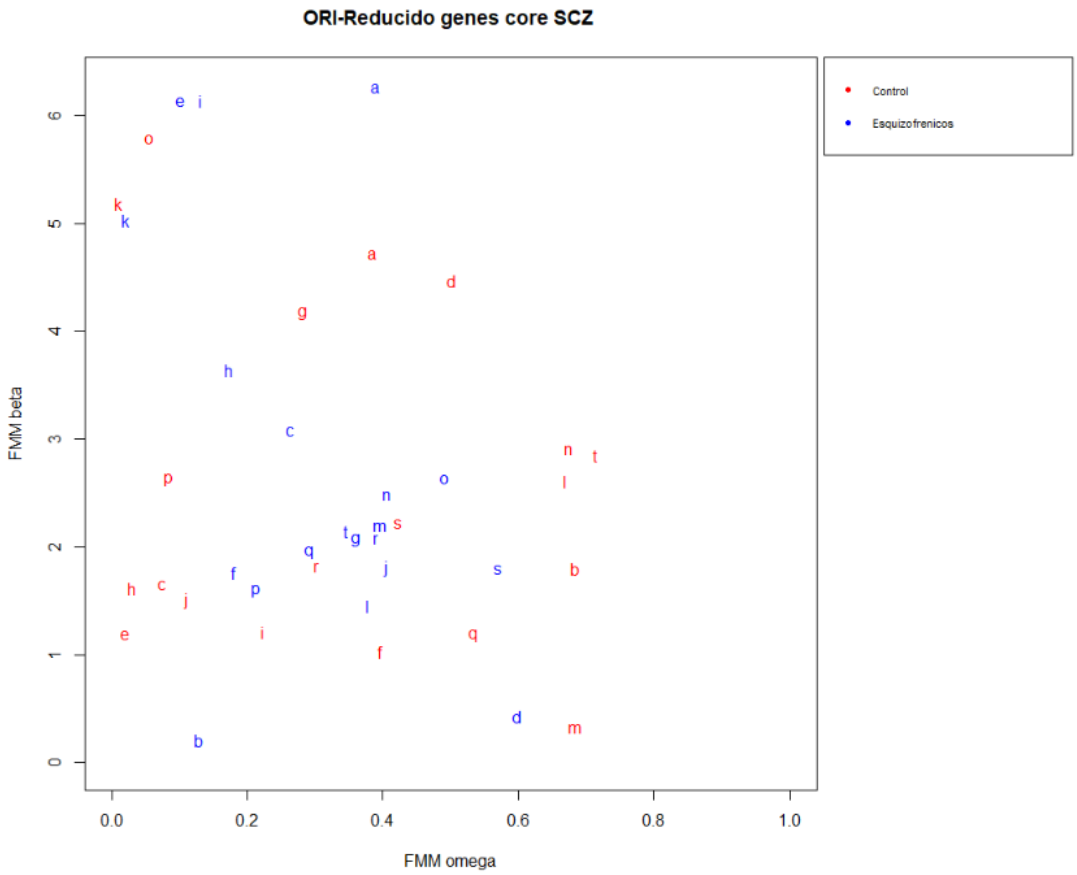


Figura 20: Parámetros ω y β del modelo FMM para los genes core del grupo de esquizofrenia en el conjunto de control y en el de pacientes con esquizofrenia.

La Figura 21 muestra los valores del parámetro A del modelo FMM contra los valores estimados con este mismo modelo para t_U para las mismas condiciones descritas para la Figura 20. Se observa como los valores de A son, en general, más bajos para el conjunto control y que también cambian los valores estimados para t_U siendo, en general, más bajos para el grupo de los pacientes con esquizofrenia.

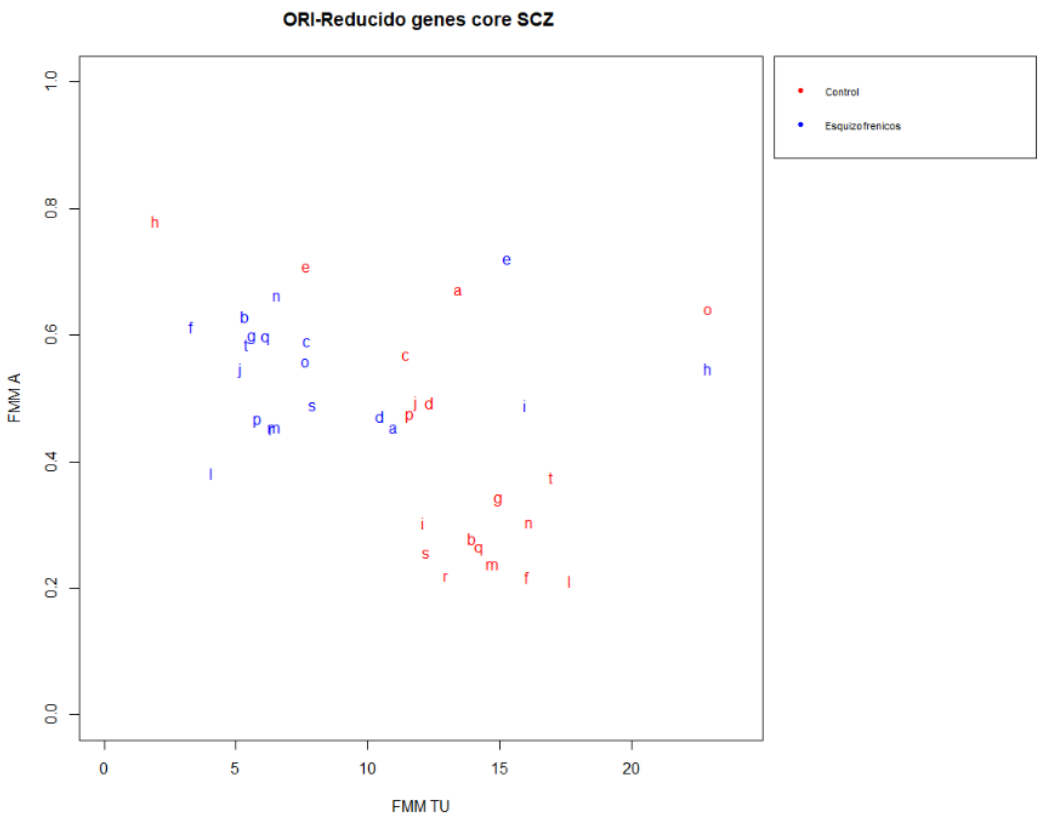


Figura 21: Parámetros A del modelo FMM y valor de t_U estimado con este mismo modelo para los genes core del grupo de esquizofrenia en el conjunto de control y en el de pacientes con esquizofrenia.

La figura 22 permite la comparación entre el grupo de control y el de pacientes con esquizofrenia, de los valores del parámetro ω y de R^2 para el modelo FMM para los *genes core* del conjunto de datos de esquizofrenia. Se observa como los valores de R^2 son en general más bajos para los pacientes del grupo de control. Hay que tener en cuenta que esto es lógico puesto que en este gráfico se están considerando los *genes core* para el grupo de esquizofrenia.

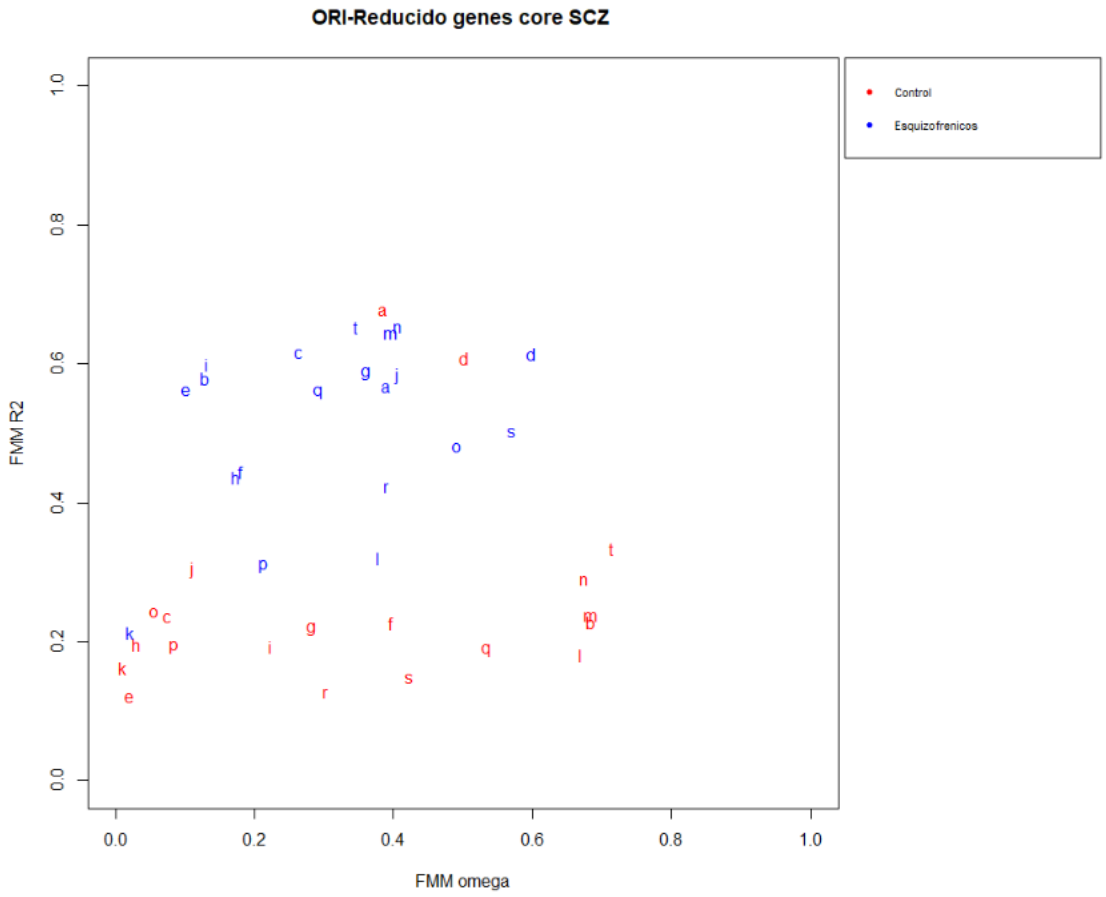


Figura 22: Parámetro ω y valor de R^2 del modelo FMM para los genes core del grupo de esquizofrenia en el conjunto de control y en el de pacientes con esquizofrenia.

.

La Figura 23 permite la comparación entre el grupo de control y los pacientes esquizofrénicos de los valores de R^2 del conjunto de los *genes core* del grupo de esquizofrenia. Se observa una disminución en general de este valor en el grupo control ya que la mayor parte de los valores están por debajo de la diagonal del gráfico. Al igual que en el gráfico anterior, esta disminución es esperable puesto que los genes que se están considerando son los *genes core* del grupo de esquizofrenia.

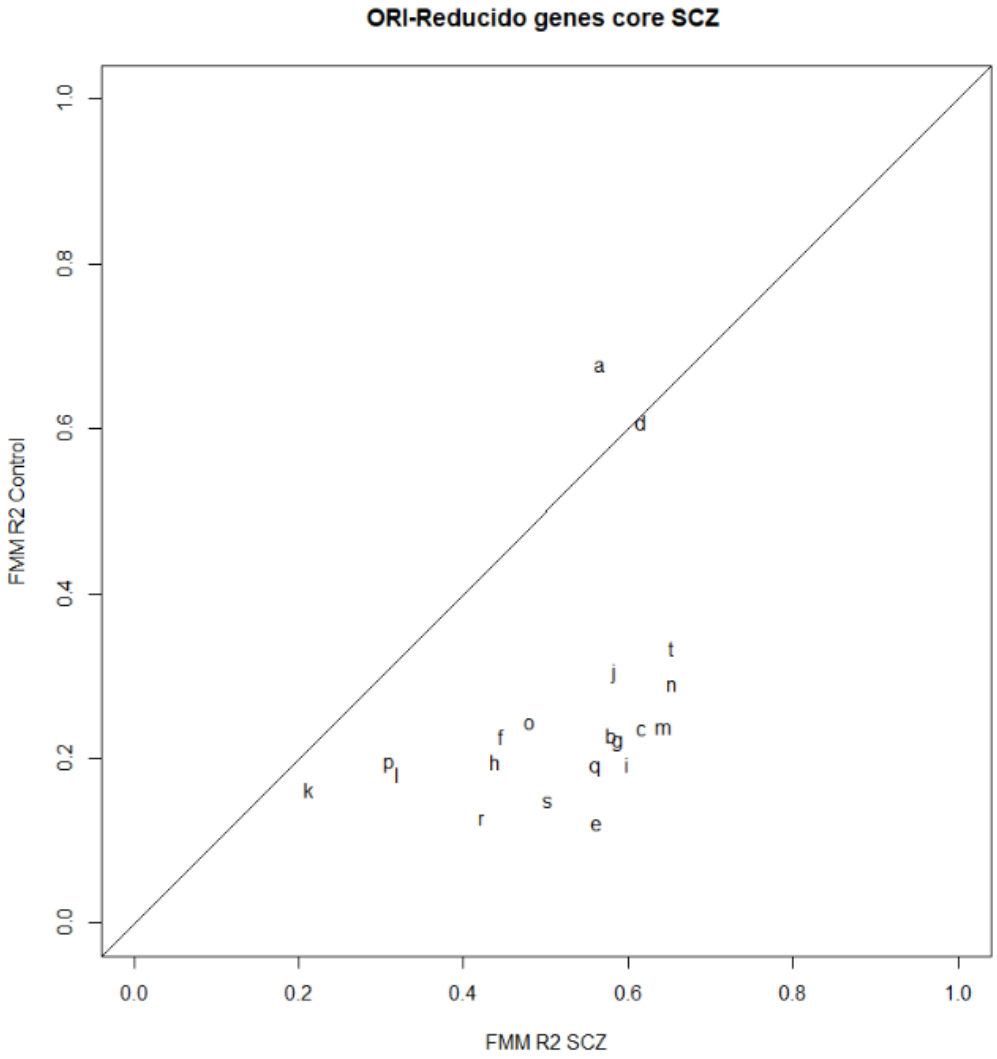


Figura 23: Valores de R^2 del modelo FMM para los genes core del grupo de esquizofrenia en el conjunto de control y en el de pacientes con esquizofrenia.

Como últimos análisis, en las Figuras 24 y 25 se ofrecen comparaciones entre los valores t_U obtenidos mediante el modelo Cosinor y el modelo FMM. En la Figura 24 aparecen los resultados correspondientes a los genes y el conjunto de control, mientras que en la 25 se observan los correspondientes al grupo de pacientes esquizofrénicos y los *genes core* de este mismo. Conviene recordar que este valor tiene interés desde el punto de vista biológico puesto que corresponde al momento de máxima expresión del gen que es cuando realiza su función biológica.

En la Figura 24 no se observan grandes diferencias en la estimación de t_U para los genes del grupo de control, pero en la figura 25 se observa que sí que existen diferencias en las estimaciones importantes de este valor en algunos de los genes (los más alejados de la diagonal) para el caso de los *genes core* correspondientes al grupo de esquizofrenia.

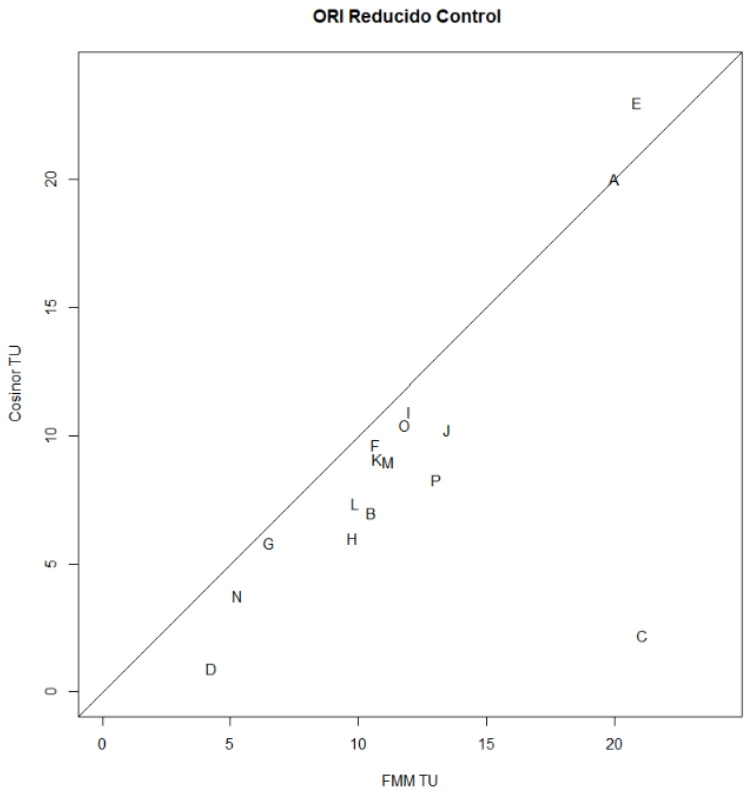


Figura 24: Valores de t_U estimados a partir de los modelos FMM y Cosinor para los genes core del grupo de control en el mismo conjunto de datos.

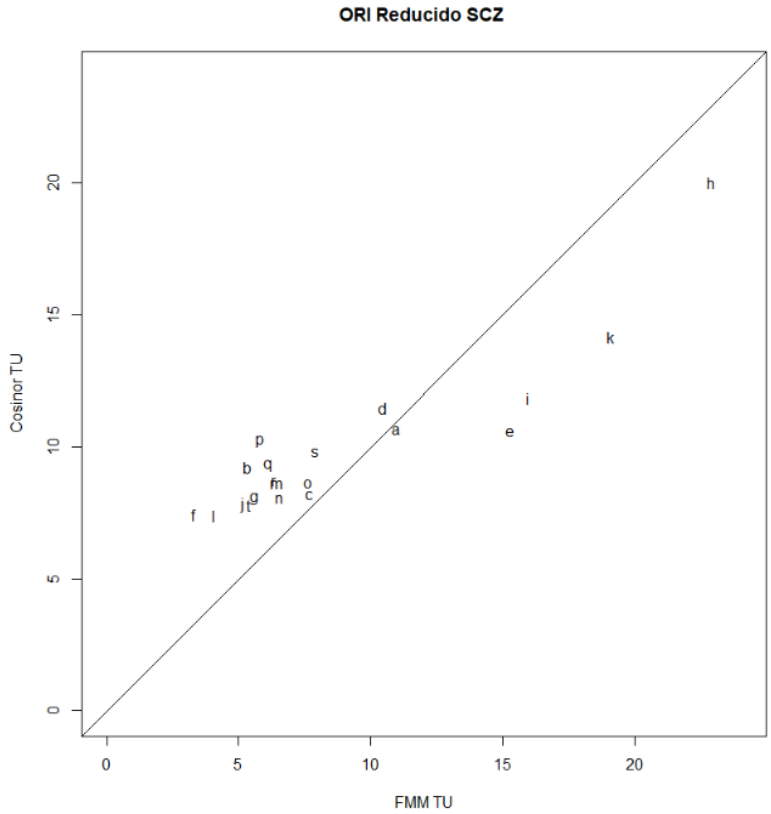


Figura 25: Valores de t_U estimados a partir de los modelos FMM y Cosinor para los genes core del grupo de esquizofrenia en el conjunto de pacientes con esquizofrenia.

4.3 Posibles nuevos genes rítmicos

En esta sección se ofrecen, a modo de ejemplo, los resultados obtenidos para algunos genes que no están en este momento catalogados como rítmicos pero que, con el nuevo procedimiento de estimación de orden considerado en este trabajo y con el modelo FMM parecen tener un patrón claramente rítmico. Pensamos que puede ser interesante, desde el punto de vista biológico, investigar las funciones de los genes de este tipo que aparecen gracias a la metodología considerada en este trabajo.

Gen FOXRED2
Control

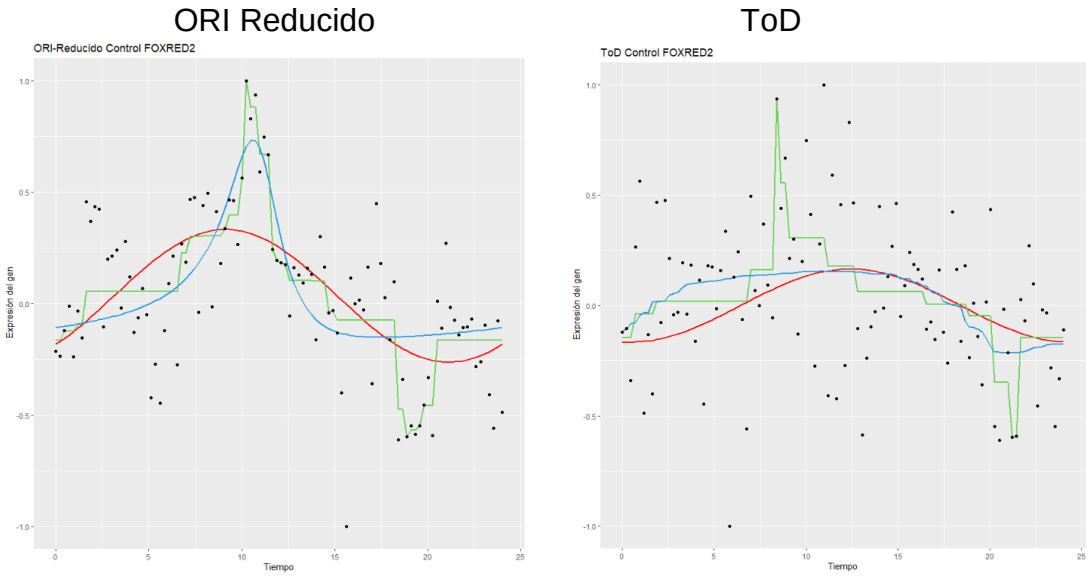


Figura 26: Gen FOXRED2 con el orden ORI Reducido y el orden ToD para individuos control.

Esquizofrenia

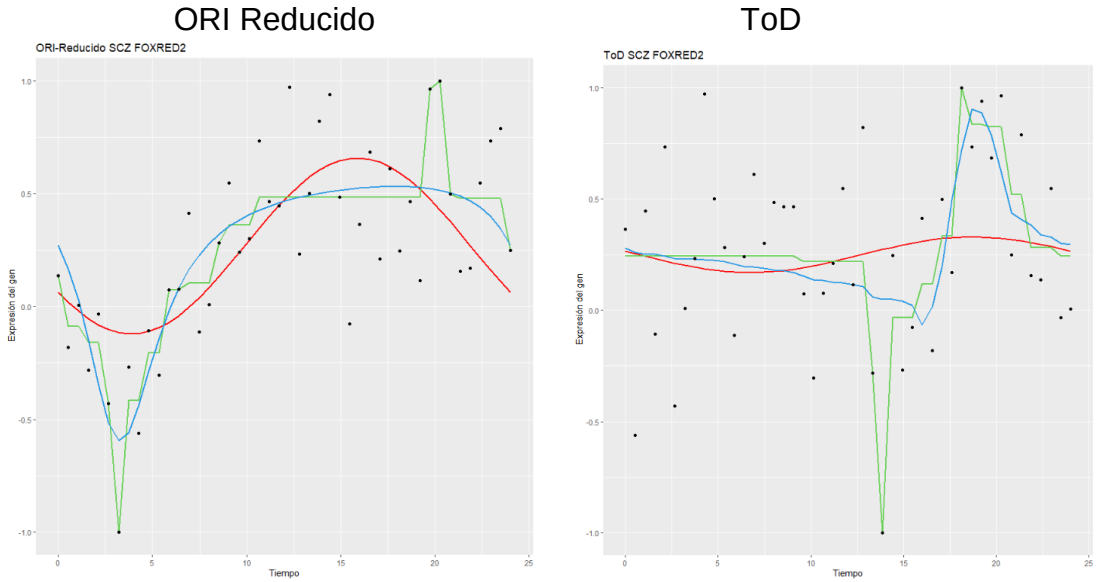
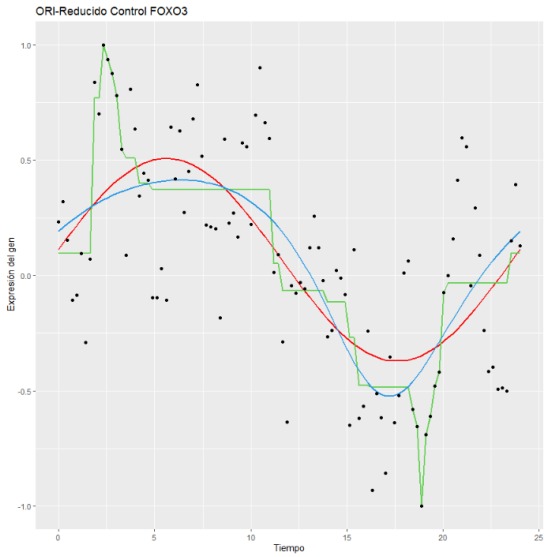


Figura 27: Gen FOXRED2 con el orden ORI Reducido y el orden ToD para individuos esquizofrénicos.

Este es un gen que tanto para el conjunto de datos de control como para el de esquizofrenia gana fuertemente en carácter rítmico si se considera el orden ORI Reducido en lugar del orden ToD. El valor de la medida de bondad de ajuste R^2 del modelo Cosinor es de 0,105 y 0,017 para el orden ToD en los conjuntos de datos de control y de esquizofrénicos respectivamente, mientras que para el orden ORI Reducido el valor de R^2 es de 0,349 y 0,412 para el modelo Cosinor y de 0,465 y 0,604 para el modelo FMM.

Gen FOXO3
Control

ORI Reducido



ToD

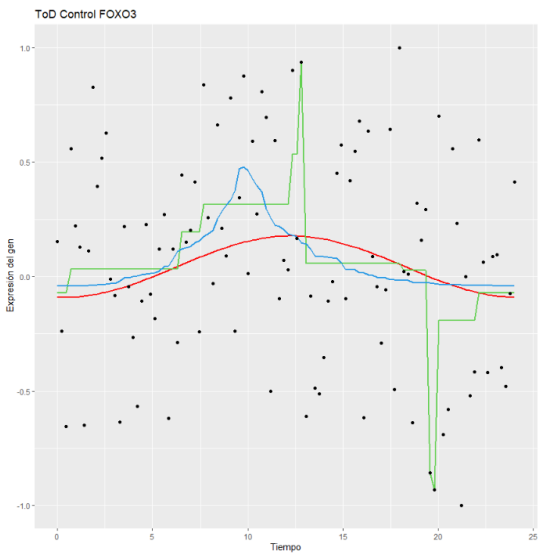
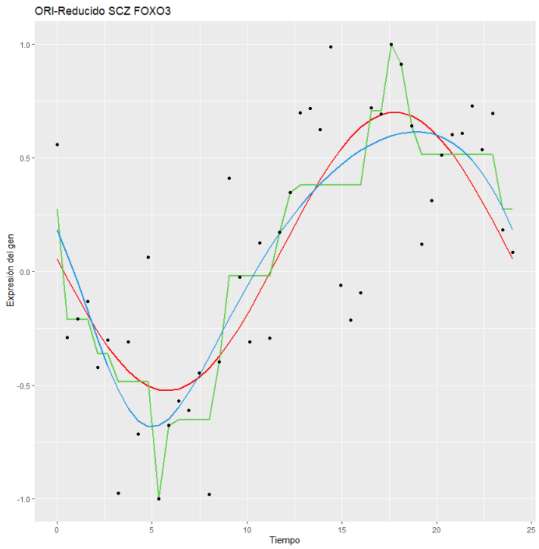


Figura 28: Gen FOXO3 con el orden ORI Reducido y el orden ToD para individuos control.

Esquizofrenia

ORI Reducido



ToD

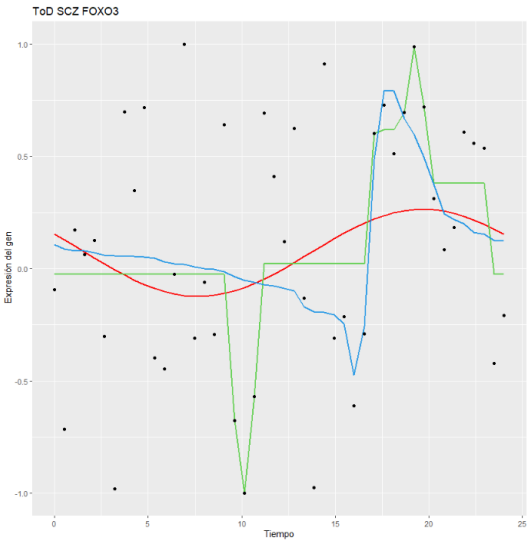
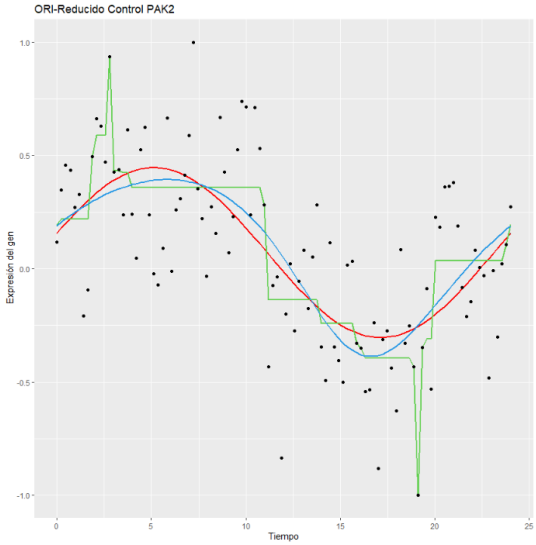


Figura 29: Gen FOXO3 con el orden ORI Reducido y el orden ToD para individuos esquizofrénicos.

La situación con este gen es similar al del anterior, es decir, que tanto para el conjunto de datos de control como para el de esquizofrenia gana fuertemente en carácter rítmico si se considera el orden ORI Reducido en lugar del orden ToD. El valor de la medida de bondad de ajuste R^2 del modelo Cosinor es de 0,041 y 0,057 para el orden ToD en los conjuntos de datos de control y de esquizofrénicos respectivamente, mientras que para el orden ORI Reducido el valor de R^2 es de 0,441 y 0,602 para el modelo Cosinor y de 0,466 y 0,643 para el modelo FMM.

Gen PAK2
Control

ORI Reducido



ToD

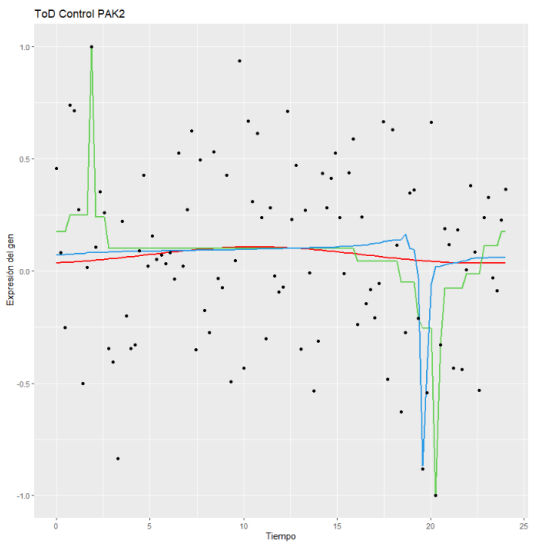
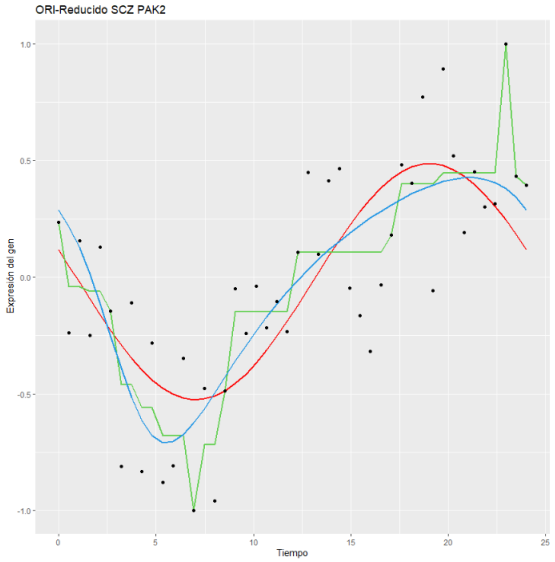


Figura 30: Gen PAK2 con el orden ORI Reducido y el orden ToD para individuos control.

Esquizofrenia

ORI Reducido



ToD

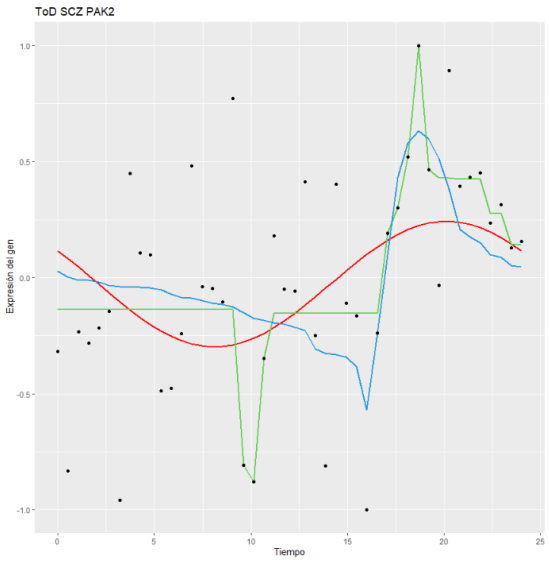


Figura 31: Gen PAK2 con el orden ORI Reducido y el orden ToD para individuos esquizofrénicos.

La situación con este gen es especialmente acentuada para el grupo de control en el que se observa en las figuras un importante incremento del carácter rítmico cuando se considera el orden ORI Reducido en lugar del orden ToD. El valor de la medida de bondad de ajuste R^2 del modelo Cosinor es de 0,004 y 0,148 para el orden ToD en los conjuntos de datos de control y de esquizofrénicos respectivamente, mientras que para el orden ORI Reducido el valor de R^2 es de 0,439 y 0,556 para el modelo Cosinor y de 0,457 y 0,643 para el modelo FMM.

5. Conclusiones

En este trabajo se propone el uso de una metodología novedosa para la estimación del orden temporal y análisis de expresiones de genes post mortem donde, se desconoce, a priori, el momento o instante en el que se toman los datos de expresión.

Las principales contribuciones metodológicas de este trabajo son dos.

Por un lado, la aplicación la metodología ORI, basada en inferencia con restricciones de orden, como alternativa a la estimación clásica del orden temporal de los instantes de muerte de los individuos a partir del ToD, tal y como viene siendo habitual en la literatura, y pese a que como se ha ilustrado en este trabajo, en ocasiones esta estimación es bastante imprecisa al estar influenciada por factores externos como la causa de la muerte o el tiempo que ha transcurrido desde el deceso.

Por otro lado, y una vez que se conoce la estimación del orden temporal entre las muestras, este trabajo propone dos modelos alternativos al modelo Cosinor, que es el modelo más extendido en la literatura para el análisis de este tipo de datos: el modelo no paramétrico NP y el modelo paramétrico FMM. En particular, cabe destacar el papel del modelo FMM en el análisis de señales oscilatorias, ya que a diferencia del modelo Cosinor, aporta la suficiente flexibilidad para adaptare a una gran variedad de patrones rítmicos, incluyendo formas asimétricas.

Para la validación y comparación de estos órdenes y medidas se han utilizado tanto medidas del error como de bondad de ajuste. En términos del error, se concluye que el orden ToD es globalmente peor que los órdenes ORI, independientemente del modelo utilizado. Si focalizamos esta medida en los dos órdenes ORI, se tiene que los valores del error del orden ORI global, que utiliza todos los genes, son sistemáticamente más bajos que los del orden ORI Reducido, obtenido únicamente con los *genes cores*. Sin embargo, se opta finalmente por el orden ORI Reducido como orden de referencia, puesto que en el orden ORI global se puede estar produciendo un efecto de “sobreajuste” y, en consecuencia, que *genes core*, con patrones de expresión claramente rítmicos, no aparezcan como tal en el orden ORI. Las medidas de bondad de ajuste vienen a complementar lo expuesto anteriormente. Se observa que para ambas selecciones de *genes core*, en ambos grupos de pacientes, el orden ORI Reducido es el que presenta valores más altos de estas medidas, y en particular para el modelo FMM. Y en línea con lo expuesto anteriormente, las medidas de bondad de ajuste presentan los valores más bajos para el orden ToD, en cualquiera que sea el modelo empleado, pesa a que este es el método de estimación de orden que se viene utilizando en la literatura.

Desde el punto de vista biológico este trabajo propone varias contribuciones de peso.

En primer lugar, se coincide con (Seney , y otros, 2019) en que el carácter rítmico de las expresiones de genes circadianos en los pacientes controles presentan un carácter rítmico más marcado que en los pacientes esquizofrénicos. Sin embargo, cuestiona la hipótesis de pérdida sistemática o ausencia generalizada de ritmicidad en pacientes esquizofrénicos establecida en este mismo trabajo, ya que este efecto podría deberse a una decisión equivocada en el método de estimación del orden temporal.

En segundo lugar, se evidencia la precisión en la estimación del momento de activación de los genes del modelo FMM frente al Cosinor, principalmente para los *genes core* del grupo control, donde el carácter rítmico es más marcado y la variedad de patrones rítmicos puede ser mayor, incluyendo patrones asimétricos. Cabe destacar la importancia de una estimación precisa de estos instantes, ya que, por ejemplo, conocerlos puede ser clave para mejorar entre otros tratamientos la quimioterapia, en función de la hora del día en que se apliquen.

De forma complementaria, el modelo FMM proporciona una caracterización y comparación de los patrones de expresión de genes en términos de la amplitud, la simetría y el apuntamiento, que es novedosa y de gran utilidad en la práctica.

Por último, se establece una nueva hipótesis de nuevos genes rítmicos como FOXRED2 y FOXO3 no considerados hasta ahora en la literatura.

Las perspectivas de trabajo futuro de este trabajo son evidentes.

A nivel biológico, se requiere de un estudio detallado para corroborar las nuevas hipótesis planteadas, así como un análisis específico de las funciones biológicas en las que los nuevos genes rítmicos están involucrados y del papel que estos juegan en ellas.

A nivel metodológico, se propone una extensión del procedimiento ORI de estimación del orden temporal que permita obtener órdenes temporales no equiespaciados, siendo más fiel a lo que ocurre en la práctica. Por otro lado, también sería interesante un análisis clúster a partir de los parámetros derivados del modelo FMM para comparar los distintos patrones de genes circadianos entre pacientes control y esquizofrénicos.

Referencias

Castro, M. (2019). Bioestadística aplicada en investigación clínica: conceptos básicos. *REVISTA MÉDICA CLÍNICA LAS CONDES*.

Colleen, K., A., S., & Doherty , J. (2010). Circadian Control of Global Gene Expression Patterns. *Annual Review of Genetics*, 419-444.

Cornelissen, G. (2014). Cosinor-based rhythmometry. *Theor Biol Med Model*.

Garcés, C. (s.f.). *Ritmos circadianos y deporte.Estudio de las oscilaciones circadianas del rendimiento y de algunos factores que las afectan.* Universidad de Barcelona.

Griffiths, J., A., Wessler, F., Lewontin, S., Gelbart, R., Suzuki, W., Miller, J. (2004). *An Introduction to Genetic Analysis.* Nueva York: MacMillan Learning.

Larriba , Y., Rueda , C., Fernández , M., & Peddada , S. (2019). Order restricted inference in chronobiology. *Statistics in Medicine*.

Liu, C., Gershon, E., & Kelsoe, J. (2017). From Gene Expression To Disease Association. *European Neuropsychopharmacology*, 416.

Panda , S., Antoch, M., Miller, B., Su, A., Schook, A., Straume, M., . . . Hogenesch , F. (2002). Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, 307-320.

Pérez, A. L. (2020). Implementación de un paquete de software para el. *TFM*.

Roth, C. M. (2002). Quantifying Gene Expression. *Current Issues in Molecular Biology*, 93-100.

Rueda , C., Larriba, Y., & Peddada , S. (2019). Frequency Modulated Möbius Model Accurately Predicts Rhythmic Signals in Biological and Physical Sciences. *Sci Rep*.

Seney , M., Cahill, K., Enwright III, J., RW, L., Huo, Z., Zong, W., . . . McClung , C. (2019). Diurnal rhythms in gene expression in the prefrontal cortex in schizophrenia. *Nature Communications*.

Shetty, M. D. (2020). The Evolution of DNA Extraction Methods. *American Journal of Biomedical Science & Research*.

Valadares, A. C., Gorki, H., Liebold , A., & Hoenicka, M. (2017). Extraction of Total RNA from Calcified Human Heart Valves for Gene Expression Analysis. *The Journal of heart valve disease*, 185-192.

Wu, G. R. (2020). A population-based gene expression signature of molecular clock phase from a single epidermal sample. *Genome medicine*.

Zhang, R. L. (2014). A circadian gene expression atlas in mammals: implications for biology and medicine. *Proceedings of the National Academy of Sciences of the United States of America*, 16219-16224.

Zhang, R., Lahens, N., Ballance, H., Hughes, M., & Hogenesch, J. (2014). A circadian gene expression atlas in mammals:Implications for biology and medicine. *PNAS*.

Anexo A: Código de generación de órdenes

```
###################### Grupo de Control ####################
#Fichero de individuos y genes Filtered_log2CPM.csv
inData<-read.csv(file="C:/Users/carlo/Desktop/TFG
Estadistica/Filtered_log2CPM.csv", header=TRUE, sep=",")[,-1]
inData<-as.matrix(inData)
rownames(inData)<-read.csv(file="C:/Users/carlo/Desktop/TFG
Estadistica/Filtered_log2CPM.csv", header=TRUE, sep=",")[,1]

# Fichero con la hora de muerte data1Seney2019.xlsx
deathData <- read.xlsx(file="C:/Users/carlo/Desktop/TFG
Estadistica/data1Seney2019.xlsx", 1,header=TRUE, sep=",")

#Fichero con todas las caracteristicas CMC_MSSM-Penn-Pitt_Clinicalv2.csv
clinicalData<-read.csv(file="C:/Users/carlo/Desktop/TFG Estadistica/CMC_MSSM-
Penn-Pitt_Clinicalv2.csv", header=TRUE, sep=";")#[,1]
clinicalData<-as.matrix(clinicalData)
controlDeath<- subset(deathData, Dx == "Control")
controlData<-merge(controlDeath, clinicalData, by="Individual_ID")

#Devuelve los indices
indControlSubjects<-
match(controlData[,"DLPFC_RNA_Sequencing_Sample_ID"],colnames(inData))
dataControlSubjects<-inData[,indControlSubjects]
controlData<- as.matrix(controlData)


####################### ORI ############################
ORI_order<-TSP_Euc_v7(datos=dataControlSubjects,dist_type="Euclidea",
                      datos2p=dataControlSubjects,pesosE=FALSE,pesosMSE=TRUE,
                      pesosAdd=FALSE,unPeriodo=TRUE,
                      pesos=rep(1,ncol(dataControlSubjects)),
                      centrar=FALSE,intentos=25,onlyHeuristica2=FALSE)

indORIData<-dataControlSubjects[,ORI_order[[1]]]
dim(dataControlSubjects)
ncol(dataControlSubjects)
time<-rescale(rep(1:104,1),to=c(0, 2 * pi))
periodo<-24


#################### ORI Reducido ##########################
geneNames<-
c("ARNTL","NPAS2","CLOCK","NFIL3","CRY1","NR1D1","BHLHE41","NR1D2","DBP","CIAR
T","PER1","PER3","TEF","HLF","CRY2","PER2")
geneSelectionData<- dataControlSubjects[geneNames,]
ORI_order_Reduced<-TSP_Euc_v7(datos=geneSelectionData,dist_type="Euclidea",
                    datos2p=dataControlSubjects,pesosE=FALSE,pesosMSE=TRUE,
                   pesosAdd=FALSE,unPeriodo=TRUE,
                  pesos=rep(1,ncol(gn)),
```

```
                centrar=FALSE,intentos=15,onlyHeuristica2=FALSE)
indORIDataReduced<-dataControlSubjects[,ORI_order_Reduced[[1]]]
time<-rescale(rep(1:104,1),to=c(0, 2 * pi))
periodo<-24


##################### Orden ZT ╦ ToD ######################
orderZT<- controlData[order(as.numeric(controlData[,"ZeitgeberTime"])
,decreasing=FALSE),]
indOrder<-match(orderZT[ ,"DLPFC_RNA_Sequencing_Sample_ID" ],colnames(inData))
indOrderData<-inData[,indOrder]
timeZT<-as.numeric((orderZT[,"ZeitgeberTime"])


######################  Grupo SCZ ########################
#Fichero de individuos y genes Filtered_log2CPM.csv
inDataSCZ<-read.csv(file="C:/Users/carlo/Desktop/TFG
Estadistica/Filtered_log2CPM.csv", header=TRUE, sep=",")[,-1]
inDataSCZ<-as.matrix(inDataSCZ)
rownames(inDataSCZ)<-read.csv(file="C:/Users/carlo/Desktop/TFG
Estadistica/Filtered_log2CPM.csv", header=TRUE, sep=",")[,1]
#Fichero con todas las caracteristicas CMC_MSSM-Penn-Pitt_Clinicalv2.csv
clinicalDataSCZ<-read.csv(file="C:/Users/carlo/Desktop/TFG
Estadistica/CMC_MSSM-Penn-Pitt_Clinicalv2.csv", header=TRUE, sep=";")#[,1]
dim(clinicalDataSCZ)

# Fichero con la hora de muerte y el ID
ZTSCZ <- read.xlsx(file="C:/Users/carlo/Desktop/TFG
Estadistica/tod_sz_control.xlsx", 1,header=TRUE, sep=",")
dim(ZTSCZ)# 92 5
ZTSCZ<- subset(ZTSCZ, Dx == "SCZ")
dim(ZTSCZ)# 46 5
SCZDeath<- subset(clinicalDataSCZ, Dx == "SCZ")
dim(SCZDeath)# 275 18

#Devuelve los indices
indSCZSubjects<-
match(SCZDeath[,"DLPFC_RNA_Sequencing_Sample_ID"],colnames(inDataSCZ),
na.omit)
dataSCZSubjects<-inDataSCZ[,indSCZSubjects]
dataSCZSubjects<-as.data.frame(dataSCZSubjects)
dataSCZSubjects <-
dataSCZSubjects[,colSums(is.na(dataSCZSubjects))<nrow(dataSCZSubjects)]
dim(dataSCZSubjects)
dataSCZSubjects<-as.matrix(dataSCZSubjects)



############################ ORI  ########################
ORI_order_SCZ<-TSP_Euc_v7(datos=dataSCZSubjects,dist_type="Euclidea",
datos2p=dataSCZSubjects,pesosE=FALSE,pesosMSE=TRUE,
                                pesosAdd=FALSE,unPeriodo=TRUE,
                                pesos=rep(1,ncol(gn)),
centrar=FALSE,intentos=25,onlyHeuristica2=FALSE)
ORI_order_SCZ<-ORI_order_ReducedSCZ
```

```
indORIDataSCZ<-dataSCZSubjects[,ORI_order_SCZ[[1]]]
timeSCZ<-rescale(rep(1:46,1),to=c(0, 2 * pi))


########################### ORI Reducido 20 ################
geneNames<-c("CIART", "WNT10B","LAMB3", "OPRL1", "CYB561", "HDAC8", "NIM1K",
"DUBR", "KRT17P1", "EBP", "PGBD2",
"CTSK","ZBTB22","NPRL2","IFT122","NFATC4","RNF112","VOPP1","NIT1","USF1")
geneSelectionDataSCZ<- dataSCZSubjects[geneNames,]
geneSelectionDataSCZ<-data.matrix(geneSelectionDataSCZ)
ORI_order_ReducedSCZ<-
TSP_Euc_v7(datos=geneSelectionDataSCZ,dist_type="Euclidea",
datos2p=dataSCZSubjects,pesosE=FALSE,pesosMSE=TRUE,
                                pesosAdd=FALSE,unPeriodo=TRUE,
                                pesos=rep(1,ncol(gn)),
centrar=FALSE,intentos=25,onlyHeuristica2=FALSE)
indORIDataReducedSCZ<-dataSCZSubjects[,ORI_order_ReducedSCZ[[1]]]
timeSCZ<-rescale(rep(1:46,1),to=c(0, 2 * pi))


#################### Orden ZT - ToD  ######################
mergeData<- merge(orderZTSCZ,SCZDeath, by="Individual_ID")
mergeData<- mergeData[order(as.numeric(mergeData[,"ZeitgeberTime"])
,decreasing=FALSE),]
indicesOrderSCZ<-match(mergeData[ ,"DLPFC_RNA_Sequencing_Sample_ID"
],colnames(dataSCZSubjects))
indOrderSCZ<-dataSCZSubjects[,indicesOrderSCZ]
timeZTSCZ<-mergeData[,"ZeitgeberTime"]
```

Anexo B: Ajuste de los modelos

```
################## GRUPO CONTROL ####################

#################### ORI ############################
indORIData<-dataControlSubjects[,orderORIControl]
errorORI<- calculoError(indORIData ,nrow(indORIData),104, time, 24)#,
"NP_ORI_Control", "Cos_ORI_Control",  "Cos_ORI_Control" )
colnames(errorORI$error)<-
c("gen","ORI_Control_NP_Error","ORI_Control_Cos_Error",
"ORI_Control_FMM_Error")
colnames(errorORI$R2)<-c("gen","ORI_Control_NP_R2","ORI_Control_Cos_R2",
"ORI_Control_FMM_R2")
colnames(errorORI$parametros_Cosinor)<-
c("gen","ORI_Control_Cosinor_M","ORI_Control_Cosinor_A",
"ORI_Control_Cosinor_phi")
colnames(errorORI$parametros_FMM)<-
c("gen","ORI_Control_FMM_M","ORI_Control_FMM_A",
"ORI_Control_FMM_alpha","ORI_Control_FMM_beta","ORI_Control_FMM_omega")
colnames(errorORI$pico_Cosinor)<-
c("gen","ORI_Control_Cosinor_ZL","ORI_Control_Cosinor_ZU",
"ORI_Control_Cosinor_TL", "ORI_Control_Cosinor_TU")
colnames(errorORI$pico_FMM)<-
c("gen","ORI_Control_FMM_ZL","ORI_Control_FMM_ZU", "ORI_Control_FMM_TL",
"ORI_Control_FMM_TU")


###################### ORI Reducido ####################
indORIDataReduced<-dataControlSubjects[,orderORIReducedControl]
errorORIReduced<- calculoError(indORIDataReduced ,nrow(indORIDataReduced),104,
time, 24)#, "NP_ORI_Reducido_Control", "Cos_ORI_Reducido_Control",
"FMM_ORI_Reducido_Control")
colnames(errorORIReduced$error)<-
c("gen","ORI_Reducido_Control_NP_Error","ORI_Reducido_Control_Cos_Error",
"ORI_Reducido_Control_FMM_Error")
colnames(errorORIReduced$R2)<-
c("gen","ORI_Reducido_Control_NP_R2","ORI_Reducido_Control_Cos_R2",
"ORI_Reducido_Control_FMM_R2")
colnames(errorORIReduced$parametros_Cosinor)<-
c("gen","ORI_Reducido_Control_Cosinor_M","ORI_Reducido_Control_Cosinor_A",
"ORI_Reducido_Control_Cosinor_phi")
colnames(errorORIReduced$parametros_FMM)<-
c("gen","ORI_Reducido_Control_FMM_M","ORI_Reducido_Control_FMM_A",
"ORI_Reducido_Control_FMM_alpha","ORI_Reducido_Control_FMM_beta","ORI_Reducido
_Control_FMM_omega")
colnames(errorORIReduced$pico_Cosinor)<-
c("gen","ORI_Reducido_Control_Cosinor_ZL","ORI_Reducido_Control_Cosinor_ZU",
"ORI_Reducido_Control_Cosinor_TL", "ORI_Reducido_Control_Cosinor_TU")
colnames(errorORIReduced$pico_FMM)<-
c("gen","ORI_Reducido_Control_FMM_ZL","ORI_Reducido_Control_FMM_ZU",
"ORI_Reducido_Control_FMM_TL", "ORI_Reducido_Control_FMM_TU")
```

```
###################### ZT - ToD ######################
indOrderData<-inData[,indOrder]
errorZT<- calculoError(indOrderData ,nrow(indOrderData),104, escalado(timeZT),
24)#, "NP_ZT_Control", "Cos_ZT_Control",  "FMM_ZT_Control")
colnames(errorZT$error)<-c("gen","ZT_Control_NP_Error","ZT_Control_Cos_Error",
"ZT_Control_FMM_Error")
colnames(errorZT$R2)<-c("gen","ZT_Control_NP_R2","ZT_Control_Cos_R2",
"ZT_Control_FMM_R2")
colnames(errorZT$parametros_Cosinor)<-
c("gen","ZT_Control_Cosinor_M","ZT_Control_Cosinor_A",
"ZT_Control_Cosinor_phi")
colnames(errorZT$parametros_FMM)<-
c("gen","ZT_Control_FMM_M","ZT_Control_FMM_A",
"ZT_Control_FMM_alpha","ZT_Control_FMM_beta","ZT_Control_FMM_omega")
colnames(errorZT$pico_Cosinor)<-
c("gen","ZT_Control_Cosinor_ZL","ZT_Control_Cosinor_ZU",
"ZT_Control_Cosinor_TL", "ZT_Control_Cosinor_TU")
colnames(errorZT$pico_FMM)<-c("gen","ZT_Control_FMM_ZL","ZT_Control_FMM_ZU",
"ZT_Control_FMM_TL", "ZT_Control_FMM_TU")


################## FICHERO ####################
error_Control<-merge(errorORI$error,
merge(errorORIReduced$error,errorZT$error, by="gen" ),by="gen")
R2_Control<-merge(errorORI$R2, merge(errorORIReduced$R2,errorZT$R2, by="gen"
),by="gen")
Cosinor_Parametros_Control<-merge(errorORI$parametros_Cosinor,
merge(errorORIReduced$parametros_Cosinor,errorZT$parametros_Cosinor, by="gen"
),by="gen")
FMM_Parametros_Control<-merge(errorORI$parametros_FMM,
merge(errorORIReduced$parametros_FMM,errorZT$parametros_FMM, by="gen"
),by="gen")
Cosinor_Pico_Control<-merge(errorORI$pico_Cosinor,
merge(errorORIReduced$pico_Cosinor,errorZT$pico_Cosinor, by="gen" ),by="gen")
FMM_Pico_Control<-merge(errorORI$pico_FMM,
merge(errorORIReduced$pico_FMM,errorZT$pico_FMM, by="gen" ),by="gen")
error_R2_Control<- merge(error_Control,R2_Control,by="gen" )
parametros_Control<-
merge(Cosinor_Parametros_Control,FMM_Parametros_Control,by="gen" )
picos_Control<- merge(Cosinor_Pico_Control,FMM_Pico_Control,by="gen" )
total_Control<- merge(error_R2_Control,merge(parametros_Control,picos_Control
, by="gen" ),by="gen" )


################ GRUPO SCZ  ##################

################## ORI ############################
indORIDataSCZ<-dataSCZSubjects[,orderORISCZ]
errorORISCZ<- calculoError(indORIDataSCZ ,nrow(indORIDataSCZ),46, timeSCZ, 24)
colnames(errorORISCZ$error)<-c("gen","ORI_SCZ_NP_Error","ORI_SCZ_Cos_Error",
"ORI_SCZ_FMM_Error")
colnames(errorORISCZ$R2)<-c("gen","ORI_SCZ_NP_R2","ORI_SCZ_Cos_R2",
"ORI_SCZ_FMM_R2")
colnames(errorORISCZ$parametros_Cosinor)<-
c("gen","ORI_SCZ_Cosinor_M","ORI_SCZ_Cosinor_A", "ORI_SCZ_Cosinor_phi")
```

```
colnames(errorORISCZ$parametros_FMM)<-c("gen","ORI_SCZ_FMM_M","ORI_SCZ_FMM_A",
"ORI_SCZ_FMM_alpha","ORI_SCZ_FMM_beta","ORI_SCZ_FMM_omega")
colnames(errorORISCZ$pico_Cosinor)<-
c("gen","ORI_SCZ_Cosinor_ZL","ORI_SCZ_Cosinor_ZU", "ORI_SCZ_Cosinor_TL",
"ORI_SCZ_Cosinor_TU")
colnames(errorORISCZ$pico_FMM)<-c("gen","ORI_SCZ_FMM_ZL","ORI_SCZ_FMM_ZU",
"ORI_SCZ_FMM_TL", "ORI_SCZ_FMM_TU")


############### ORI Reducido ######################
indORIDataReducedSCZ<-dataSCZSubjects[,orderORIReducedSCZ]
errorORIReducedSCZ<- calculoError(indORIDataReducedSCZ
,nrow(indORIDataReducedSCZ),46, timeSCZ, 24) #, "NP_ORI_Reducido_Control",
"Cos_ORI_Reducido_Control",  "FMM_ORI_Reducido_Control")
colnames(errorORIReducedSCZ$error)<-
c("gen","ORI_Reducido_SCZ_NP_Error","ORI_Reducido_SCZ_Cos_Error",
"ORI_Reducido_SCZ_FMM_Error")
colnames(errorORIReducedSCZ$R2)<-
c("gen","ORI_Reducido_SCZ_NP_R2","ORI_Reducido_SCZ_Cos_R2",
"ORI_Reducido_SCZ_FMM_R2")
colnames(errorORIReducedSCZ$parametros_Cosinor)<-
c("gen","ORI_Reducido_SCZ_Cosinor_M","ORI_Reducido_SCZ_Cosinor_A",
"ORI_Reducido_SCZ_Cosinor_phi")
colnames(errorORIReducedSCZ$parametros_FMM)<-
c("gen","ORI_Reducido_SCZ_FMM_M","ORI_Reducido_SCZ_FMM_A",
"ORI_Reducido_SCZ_FMM_alpha","ORI_Reducido_SCZ_FMM_beta","ORI_Reducido_SCZ_FMM
_omega")
colnames(errorORIReducedSCZ$pico_Cosinor)<-
c("gen","ORI_Reducido_SCZ_Cosinor_ZL","ORI_Reducido_SCZ_Cosinor_ZU",
"ORI_Reducido_SCZ_Cosinor_TL", "ORI_Reducido_SCZ_Cosinor_TU")
colnames(errorORIReducedSCZ$pico_FMM)<-
c("gen","ORI_Reducido_SCZ_FMM_ZL","ORI_Reducido_SCZ_FMM_ZU",
"ORI_Reducido_SCZ_FMM_TL", "ORI_Reducido_SCZ_FMM_TU")

######################## ZT - ToD ######################
indOrderSCZ<-dataSCZSubjects[,indicesOrderSCZ]
timeZTSCZ<-mergeData[,"ZeitgeberTime"]
errorZTSCZ<- calculoError(indOrderSCZ ,nrow(indOrderSCZ),46,
escalado(timeZTSCZ), 24) #, "NP_ZT_Control", "Cos_ZT_Control",
"FMM_ZT_Control")
colnames(errorZTSCZ$error)<-c("gen","ZT_SCZ_NP_Error","ZT_SCZ_Cos_Error",
"ZT_SCZ_FMM_Error")
colnames(errorZTSCZ$R2)<-c("gen","ZT_SCZ_NP_R2","ZT_SCZ_Cos_R2",
"ZT_SCZ_FMM_R2")
colnames(errorZTSCZ$parametros_Cosinor)<-
c("gen","ZT_SCZ_Cosinor_M","ZT_SCZ_Cosinor_A", "ZT_SCZ_Cosinor_phi")
colnames(errorZTSCZ$parametros_FMM)<-c("gen","ZT_SCZ_FMM_M","ZT_SCZ_FMM_A",
"ZT_SCZ_FMM_alpha","ZT_SCZ_FMM_beta","ZT_SCZ_FMM_omega")
colnames(errorZTSCZ$pico_Cosinor)<-
c("gen","ZT_SCZ_Cosinor_ZL","ZT_SCZ_Cosinor_ZU", "ZT_SCZ_Cosinor_TL",
"ZT_SCZ_Cosinor_TU")
colnames(errorZTSCZ$pico_FMM)<-c("gen","ZT_SCZ_FMM_ZL","ZT_Control_FMM_ZU",
"ZT_SCZ_FMM_TL", "ZT_SCZ_FMM_TU")
```

```
######################## FICHERO #################
error_SCZ<-merge(errorORISCZ$error,
merge(errorORIReducedSCZ$error,errorZTSCZ$error, by="gen" ),by="gen")
R2_SCZ<-merge(errorORISCZ$R2, merge(errorORIReducedSCZ$R2,errorZTSCZ$R2,
by="gen" ),by="gen")
Cosinor_Parametros_SCZ<-merge(errorORISCZ$parametros_Cosinor,
merge(errorORIReducedSCZ$parametros_Cosinor,errorZTSCZ$parametros_Cosinor,
by="gen" ),by="gen")
FMM_Parametros_SCZ<-merge(errorORISCZ$parametros_FMM,
merge(errorORIReducedSCZ$parametros_FMM,errorZTSCZ$parametros_FMM, by="gen"
),by="gen")
Cosinor_Pico_SCZ<-merge(errorORISCZ$pico_Cosinor,
merge(errorORIReducedSCZ$pico_Cosinor,errorZTSCZ$pico_Cosinor, by="gen"
),by="gen")
FMM_Pico_SCZ<-merge(errorORISCZ$pico_FMM,
merge(errorORIReducedSCZ$pico_FMM,errorZTSCZ$pico_FMM, by="gen" ),by="gen")
error_R2_SCZ<- merge(error_SCZ,R2_SCZ,by="gen" )
parametros_SCZ<- merge(Cosinor_Parametros_SCZ,FMM_Parametros_SCZ,by="gen" )
picos_SCZ<- merge(Cosinor_Pico_SCZ,FMM_Pico_SCZ,by="gen" )
total_SCZ<- merge(error_R2_SCZ,merge(parametros_SCZ,picos_SCZ , by="gen"
),by="gen" )
```

Anexo C: Funciones y librerías

```
library(Iso)
library(TSP)
library(foreach)
library(iterators)
library(parallel)
library(doParallel)
library(bigstatsr)
library(scales)
library(xlsx)
library(FMM)
library(ggplot2)
library(cosinor2)
library(ggplot2)
library(gridExtra)
library(grid)
library("writexl")
library("grid")
library("ggplotify")



cosinor.lm <- function(formula, period = 12,
                       dato, na.action = na.omit){

  # build time tranformations

  Terms <- terms(formula, specials = c("time", "amp.acro"))

  stopifnot(attr(Terms, "specials")$time != 1)
  varnames <- get_varnames(Terms)
  timevar <- varnames[attr(Terms, "specials")$time - 1]

  xx<-cos(dato[,timevar])
  zz<-sin(dato[,timevar])

  fit<-lm((dato[,1])~xx+zz)
  Mest<-fit$coefficients[1]
  bb<-fit$coefficients[2]
  gg<-fit$coefficients[3]
  phiEst<-atan2(-gg,bb)%%(2*pi)
  Aest<-sqrt(bb^2+gg^2)

  rss<- sum (( dato[,1] - (Mest+bb*cos(2*pi*time/periodo)-
gg*sin(2*pi*time/periodo)))^2)

  #data$rrr <- cos(2 * pi * dato[,timevar] / period)
  #data$sss <- sin(2 * pi * dato[,timevar] / period)

  data$rrr <- xx
  data$sss <- zz

  spec_dex <- unlist(attr(Terms, "special")$amp.acro) - 1
```

```
  mainpart <- c(varnames[c(-spec_dex, - (attr(Terms, "special")$time - 1))],
"rrr", "sss")
  acpart <- paste(sort(rep(varnames[spec_dex], 2)), rep(c("rrr", "sss"),
length(spec_dex)), sep = ":")
  newformula <- as.formula(paste(rownames(attr(Terms, "factors"))[1],
                                 paste(c(mainpart, acpart), collapse = " + "),
sep = " ~ "))




  fit <- lm(newformula, data, na.action = na.action)

  mf <- fit

  r.coef <- c(FALSE, as.logical(attr(mf$terms, "factors")["rrr",]))
  s.coef <- c(FALSE, as.logical(attr(mf$terms, "factors")["sss",]))
  mu.coef <- c(TRUE, ! (as.logical(attr(mf$terms, "factors")["sss",]) |
                          as.logical(attr(mf$terms, "factors")["rrr",])))

  beta.s <- mf$coefficients[s.coef]
  beta.r <- mf$coefficients[r.coef]

  groups.r <- c(beta.r["rrr"], beta.r["rrr"] + beta.r[which(names(beta.r) !=
"rrr")])
  groups.s <- c(beta.s["sss"], beta.s["sss"] + beta.s[which(names(beta.s) !=
"sss")])

  amp <-Aest
  #amp <- sqrt(groups.r^2 + groups.s^2)
  names(amp) <- gsub("rrr", "amp", names(beta.r))

  #acr <-phiEst
  acr <- atan(groups.s / groups.r)
  # print("acrofase")
  # print(atan(groups.s / groups.r))
  names(acr) <-  gsub("sss", "acr", names(beta.s))
  coef <- c(mf$coefficients[mu.coef], amp, acr)
  #coef <- c(Mest, amp, acr)


  structure(list(fit = fit, Call = match.call(), Terms = Terms, coefficients =
coef, period = period, rss=rss), class = "cosinor.lm")

}

get_varnames <- function(Terms){

  spec <- names(attr(Terms, "specials"))
  tname <- attr(Terms, "term.labels")

  dex <- unlist(sapply(spec, function(sp){

    attr(Terms, "specials")[[sp]] - 1
```

```
  }))

  tname2 <- tname
  for(jj in spec){

    gbl <- grep(paste0(jj, "("), tname2, fixed = TRUE)
    init <- length(gbl) > 0
    if( init ){
      jlack <- gsub(paste0(jj, "("), "", tname2, fixed = TRUE)
      tname2[gbl] <- substr(jlack[gbl], 1, nchar(jlack[gbl]) - 1)
    }

  }

  tname2

}

update_covnames <- function(names){

  covnames <- grep("(amp|acr|Intercept)", names, invert = TRUE, value = TRUE)

  lack <- names
  for(n in covnames){
    lack <- gsub(paste0(n, ":"), paste0("[", n, " = 1]:"), lack)
    lack <- gsub(paste0("^", n, "$"), paste0("[", n, " = 1]"), lack)
  }
  lack
}

ggplot.cosinor.lm <- function(object,l, x_str = NULL){

  timeax <- seq(0, object$period, length.out = l)
  covars <- grep("(rrr|sss)", attr(object$fit$terms, "term.labels"), invert =
TRUE, value = TRUE)

  newdata <- data.frame(time = timeax, rrr = cos(2 * pi * timeax /
object$period),
                        sss = sin(2 * pi * timeax / object$period))
  for(j in covars){
    newdata[,j] <- 0
  }
  if(!is.null(x_str)){

    for(d in x_str){

      tdat <- newdata
      tdat[,d] <- 1
      newdata <- rbind(newdata, tdat)

    }
    newdata$levels <- ""
    for(d in x_str){

      newdata$levels <- paste(newdata$levels, paste(d, "=", newdata[,d]))
```

```
    }

  }


  newdata$Y.hat <- predict(object$fit, newdata = newdata)

  if(missing(x_str) || is.null(x_str)){

    ggplot(newdata, aes_string(x = "time", y = "Y.hat")) + geom_line()

  } else {

    ggplot(newdata, aes_string(x = "time", y = "Y.hat", col = "levels")) +
geom_line()

  }
}

grafico<-function(datos, gen, longitud, periodo,t, titulo= "grafica"){

  datosCos<-rescale(datos[gen,], to=c(-1,1))
  dataGen<- rescale(datos[gen,], to=c(-1,1))
  NP<- function1Local(dataGen)
  COS<-funcionCosinor(datosCos,t,periodo)
  print(" phi")
  print(COS[[5]])
  FMM<-fitFMM(vData=dataGen,timePoints=t)
  #t<-rescale(rep(1:longitud,1),to=c(0, 2 * pi))



  data<-rbind(datosCos, t)
  data<-t(data)
  dato<-data
  dato<-as.data.frame(dato)
  cos<-cosinor.lm(dato[,1]~ time(t),period = periodo, dato=dato )

  g<-ggplot.cosinor.lm(cos, l=longitud)+geom_line(lwd=1,color="red") +
ggtitle(titulo)+theme(aspect.ratio=1)+labs(x = "Tiempo", y =" Expresión del
gen")
  g<-g+geom_line(aes(y=NP[[1]]), lwd=1, col=3)#ajuste NP

  print(FMM@M+FMM@A*cos(FMM@beta+2*atan(FMM@omega*tan((t-FMM@alpha)/2))))
  g<-g+geom_line(aes(y=FMM@M+FMM@A*cos(FMM@beta+2*atan(FMM@omega*tan((t-
FMM@alpha)/2)))), col=4,lwd=1)#ajuste NP
  g<-g+geom_point(aes(y=datosCos))#+geom_vline(xintercept = COS[[8]]+6)
  g
  return(g)

}

calculoError<- function(data,nGen,nIndv, time, period, NP="NP", Cos="Cos",
FMM="FMM"  ){
```

```
  error<- matrix(nrow=nGen, ncol = 4)
  ajuste<- matrix(nrow=nGen, ncol = 4)
  parametros_Cosinor<- matrix(nrow=nGen, ncol = 4)
  parametros_FMM<- matrix(nrow=nGen, ncol = 6)
  picos_Cosinor<- matrix(nrow=nGen, ncol = 5)
  picos_FMM<- matrix(nrow=nGen, ncol = 5)

  for (i in 1:nGen){#nrow(indORIDataReducedSCZ)){
    data[i,]<- rescale(data[i,], to=c(-1,1))
    NP<- function1Local(data[i,])
    COS<-funcionCosinor(data[i,],time,period)
    FMM<-fitFMM(vData=data[i,],timePoints=time)
    picoFMM<-getFMMPeaks(FMM)
    a<-ajus(data, i, time, "Ori")

    error[i, ]<- c(rownames(data)[i] ,as.numeric(NP[[2]]),
                   as.numeric(COS[[6]]/nIndv), as.numeric(getSSE(FMM)/nIndv))
    ajuste[i, ]<- c(rownames(data)[i] ,a$NP, a$cos, a$FMM)
    parametros_Cosinor[i,]<-c(rownames(data)[i] ,COS[[2]],COS[[3]], COS[[5]])
    parametros_FMM[i,]<-c(rownames(data)[i] ,FMM@M, FMM@A, FMM@alpha,
FMM@beta, FMM@omega)
    picos_Cosinor[i,]<-c(rownames(data)[i]
,(COS[[2]]+COS[[3]]*cos(pi)),(COS[[2]]+COS[[3]]*cos(0)), COS[[8]], COS[[9]])
    picos_FMM[i,]<-c(rownames(data)[i] ,picoFMM$ZL, picoFMM$ZU,
picoFMM$tpeakL,picoFMM$tpeakU)
  }

  colnames(error)<-c("gen","NpError","CosError", "FMMError")
  colnames(ajuste)<-c("gen","NpR2","CosR2", "FMMR2")
  colnames(parametros_Cosinor)<-c("gen","M","A", "phi")
  colnames(parametros_FMM)<-c("gen","M","A", "alpha", "beta", "omega")
  colnames(picos_Cosinor)<-c("gen","ZL","ZU", "TL", "TU")
  colnames(picos_FMM)<-c("gen","ZL","ZU", "TL", "TU")

  NPError<-sum(as.numeric(error[,2]),na.rm=TRUE)
  CosError<-sum(as.numeric(error[,3]),na.rm=TRUE)
  FMMError<-sum(as.numeric(error[,4]),na.rm=TRUE)

  return(list(error=error, R2=ajuste, parametros_Cosinor=parametros_Cosinor,
parametros_FMM=parametros_FMM, pico_Cosinor=picos_Cosinor,
pico_FMM=picos_FMM))

}

PV <- function(vData,pred){
  meanVData <- mean(vData)
  return(1 - sum((vData-pred)^2)/sum((vData-meanVData)^2))
}



ajusteR2<- function (Ori, OriR, ZT, t, tZT, gen){
```

```
  ajus(Ori, gen, t, "Ori")
  ajus(OriR, gen, t, "Ori Reducido ")
  ajus(ZT, gen, tZT, "ZT ")

}

ajus<- function(data, gen, time, nombre= "Ajuste"){
  datos<-data[gen,]
  dataGen<- rescale(datos, to=c(-1,1))
  NP3<- function1Local(dataGen)
  FMM<-fitFMM(vData=dataGen,timePoints=time)
  cos<- funcionCosinor(dataGen, time, 24)
  return(list(NP=NP3[[7]], cos=cos[[7]], FMM=FMM@R2))

}


escalado<- function(datos){
  esc<- matrix(nrow=1, ncol = length(datos))
  for( i in 1: length(datos)){

    esc[i]<- (datos[i] + 6)/(12/pi)

  }
  esc<-c(esc)
  return(esc)

}


escaladoInverso<- function(datos){
  esc<- matrix(nrow=1, ncol = length(datos))
  for( i in 1: length(datos)){

    esc[i]<- ((datos[i] *12)/pi)

  }
  esc<-c(esc)
  return(esc)

}

generateFMM <-
function(M,A,alpha,beta,omega,from=0,to=2*pi,length.out=100,timePoints=seq(fro
m,to,length=length.out),
                         plot=TRUE,outvalues=TRUE,sigmaNoise=0){

  pl<-nullGrob()
  narg <- max(length(M),length(A),length(alpha),length(beta),length(omega))

  if(length(M)>1){
    warning("M parameter should be a vector of length 1.
            The intercept parameter used in the simulation is the sum of the
elements of the argument M.")
    M <- sum(M)
```

```
  }
  M <- rep(M/narg,length.out=narg)


  A <- rep(A,length.out=narg)
  if(sum(A <= 0) > 0) stop("A parameter must be positive.")

  alpha <- rep(alpha,length.out=narg)
  alpha <- alpha%%(2*pi) # between 0 and 2*pi

  beta <- rep(beta,length.out=narg)
  beta <- beta%%(2*pi) # between 0 and 2*pi

  omega <- rep(omega,length.out=narg)
  if(sum(omega<0)>0 | sum(omega>1)>0) stop("omega parameter must be between 0
and 1.")

  t <- timePoints

  phi <- list()
  for(i in 1:narg){
    phi[[i]] <- beta[i]+2*atan(omega[i]*tan((t-alpha[i])/2))
  }

  ym <- list()
  for(i in 1:narg){
    ym[[i]] <- M[i]+A[i]*cos(phi[[i]])
  }

  y <- rep(0,length(t))
  for(i in 1:narg){
    y <- y + ym[[i]]
  }

  if (sigmaNoise > 0) y <- y + rnorm(length.out,0,sigmaNoise)

  if(plot) {
    type_<-ifelse(sigmaNoise==0,"l","p")
    pl<- plot(t,y,type=type_,lwd=2,col=2,xlab="Time",ylab="Response",
              main=paste("Simulated data from FMM model"))

  }


  if(outvalues) return(list(input=list(M = M[1]*narg,
A=A,alpha=alpha,beta=beta,omega=omega,t=t,y=y, p=pl)))
}
```