UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

DETERMINACION DE GENES POTENCIALES EN EL GENOMA DE LA LEVADURA *Saccharomyces cerevisiae* MEDIANTE EL ANALISIS DE MARCOS DE LECTURA ABIERTA (ORFs)

TESIS PROFESIONAL

QUE PARA OBTENER EL TITULO DE

B I O L O G O

P R E S E N T A :

ALFONSO JOSE VILCHIS PELUYERA

DIRECTOR DE TESIS: VICTOR MANUEL VALDES-LOPEZ

1999

VNIVERDAD NACIONAL
AVPNOMA DE
MEXICO

MAT. MARGARITA ELVIRA CHÁVEZ CANO
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo de Tesis:

DETERMINACION DE GENES POTENCIALES EN EL GENOMA DE LA LEVADURA Saccharomyces cerevisiae MEDIANTE EL ANALISIS DE MARCOS DE LECTURA ABIERTA (ORFs)

realizado por ALFONSO JOSE VILCHIS PELUYERA

con número de cuenta 7023059-0 , pasante de la carrera de BIOLOGIA

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	M en C Víctor Manuel Valdés López
Propietario	Dr Roberto Coria Ortega
Propietario	Dr Germinal Cocho Gil
Suplente	Dra Luisa Alvarina Alba Lois
Suplente	M en IBB Claudia Andrea Segal Kischinevzky

FACUL[illegible]NCIAS

Consejo Departamental de BIOLOGIA

Dra. Edna María Suárez Díaz

DEPARTAMENTO
DE BIOLOGIA

AGRADECIMIENTOS

Es para mí un un enorme gusto el hacer patente mi agradecimiento a las siguientes personas, a las cuales me une un afecto muy especial :

Dra. Luisa Alba Lois y M. en C. Víctor Valdés López, quienes siempre han sido para mí, maestros y amigos y gracias a los cuales mi vida ha tomado un nuevo sentido, lleno de futuro y de realizaciones personales y colectivas.
Dra. Alejandra Mainero Del Paso y M. en I. B. B. Claudia Segal Kischinevzky, quienes, desde que estoy en el laboratorio de Biología Molecular, me han ofrecido toda su ayuda, sus conocimientos y su afecto.
A todos los compañeros del laboratorio, Beatriz Rodarte, Epifanio Quiroz, José Romo y Romel Hernández, en los cuales he encontrado amistad, compañerismo y una buena dosis de buen humor y de capacidad.
A los doctores Germinal Cocho Gil y Roberto Coria Ortega, de los Institutos de Física y de Fisiología Celular, respectivamente, por su trabajo de revisión de esta tesis y sus consejos.
A todos los profesores que me han formado desde la infancia y a los cuales debo, en gran parte, mi desarrollo personal.
Es de particular alegría agradecer a la familia de mi esposa, a su mamá (a la cual quiero como a una segunda madre), sus hermanas y hermanos, especialmente a mi cuñado, Ausenclo Domínguez, por el inmenso apoyo que nos han brindado, a mi esposa, a mis hijos y a mí, en momentos cruciales de nuestras vidas.

DEDICATORIA

Dedico esta investigación a mi abuela Pastora, a mi madre Julia y a mi hermano Julio, a los cuales les hubiera dado una gran alegría tenerla en sus manos. Mi amor, mi agradecimiento, y mi recuerdo para ellos son eternos.

También dedico este trabajo a mi hermana Pastora y a mi hermano Jesús, con todo mi cariño.

De la misma manera, va una dedicatoria a mi cuñada Luz, a mi cuñado Douglas Frankel, y a mis sobrinos Alonso, Ana, Julia, Pablo, Víctor y Jimena.

Es mi mayor emoción dedicar esta tesis a mi esposa LUZ y a mis hijos PASTORA MONTSERRAT e IAN RODRIGO, los cuales significan para mí todo en la vida.

INDICE

RESUMEN

Recientemente se ha reportado la secuencia completa del genoma de la levadura *Saccharomyces cerevisiae*. Se ha calculado que en este genoma, de mas de 12 mega-bases, se localizan alrededor de 6,000 genes. Los criterios en la asignación de regiones codificantes se han basado en similitudes con genes reportados (genes anotados), presencia de regiones de lectura abierta (ORFs) con capacidad codificante de más de 100 residuos de aminoácidos y utilización preferencial de codones de acuerdo al dialecto del código genético de esta levadura (genes potenciales). Sin embargo, se ha notado una discrepancia en la distribución de tamaño entre los genes anotados y los genes potenciales. En estos últimos aparece un exceso de ORFs de tamaño pequeño (≈ 100 codones). Esta discrepancia se debe a que en una secuencia de nucleótidos al azar (no codificante), estadísticamente existe la probabilidad de que aparezcan ORFs de ese tamaño. El problema que hemos abordado es el de intentar discriminar entre los ORFs pequeños generados al azar de aquellos potencialmente codificantes. Metodológicamente hemos utilizado la característica de distribución asimétrica de purinas/pirimidinas que es distintiva de genes tanto procariontes como eucariontes. Como resultado, se logró la identificación diferencial de ORFs pequeños con potencial codificante y sin potencial codificante dentro de las categorías de ORFs cuestionables e hipotéticos, los cuales no habían sido previamente definidos en el genoma de la levadura.

1- INTRODUCCION.

El estudio genético de la levadura *Saccharomyces cerevisiae* comenzó en el año de 1935 con el descubrimiento de las fases haploide y diploide en el ciclo de vida de este microorganismo; estas investigaciones las llevó a cabo Ojvind Winge en los laboratorios Carlsberg, en la ciudad de Copenhage (Mortimer, 1993). Este descubrimiento, a su vez, dió origen a una serie de investigaciones sobre los sistemas de intercambio sexual en la levadura y, de forma independiente, al descubrimiento del fenómeno de heterotalismo o sistema de polaridad en el apareamiento de esporas por parte de Carl y Gertrude Lindegren en 1943 (Roman, 1986).

Las razones para la elección de *Saccharomyces cerevisiae* como organismo modelo para estudios genéticos han sido las siguientes: simplicidad de cultivo , corto tiempo de generación, caracterización biquímica avanzada, facilidad de identificación de sistemas celulares haploides y diploides y sencillez en la manipulación genética para el aislamiento de mutantes así como para el intercambio cromosómico (Oshima, 1993).

Fue durante la década de 1970 que el estudio genético y bioquímico de la levadura se profundizó con el desarrollo de nuevas técnicas tales como la clonación molecular de DNA ribosómico (Petes y Botstein, 1977), la construcción de las primeras bibliotecas genómicas de levadura (Ratzkin y Carbon,1977) y el desarrollo de métodos de transformación de células de levadura con plásmidos conteniendo genes de la propia levadura (Beggs, 1978; Hsiao y Carbon, 1979). Por esa época, también comenzaron a desarrollarse los primeros experimentos de aislamiento de secuencias de los centrómeros, utilizando la técnica de hibridación de secuencias sobrelapadas, conocida como "chromosome walking" (Chinault y Carbon, 1979), así como la construcción de los primeros cromosomas artificiales o YACs (Yeast Artificial Chromosomes) (Clarke y Carbon, 1980; Murray y Szostak, 1983;

Carbon, 1993). Con estas técnicas esenciales, el estudio de la genética de la levadura a nivel molecular avanzó rápidamente durante los años 80s, incorporando, durante esta década, nuevos métodos de análisis, tales como la transformación permanente por plásmidos recombinantes (Struhl, 1983) y la disrupción de genes (Rothstein, 1983).

A partir de 1989 y con la automatización de las técnicas de secuenciación de nucleótidos, diversos laboratorios europeos, norteamericanos y japoneses, unieron sus esfuerzos en un gran Megaproyecto de secuenciación del genoma de *Saccharomyces cerevisiae*, dirigido por André Goffeau, de la Universidad Católica de Lovaina, Bélgica (Goffeau, 1996), el cual fue completado a principios de 1996 y puesto a disposición de la comunidad científica internacional el 24 de abril de ese año en la Red Mundial Internet en dos direcciones:
http://genome-www.stanford.edu/*Saccharomyces*.
http://speedy.mips.biochem.mpg.de/mips/yeast-genome-www.htmlx.
Las dos direcciones anteriores corresponden al Genome Database (SGD) de la Universidad de Stanford, en California, y al Martinsried Institute for Protein Sequences (MIPS), dependiente del Instituto Max Planck, en Munich, Alemania, respectivamente. Con la secuenciación del genoma completo de la levadura se abrió una nueva etapa en el estudio de la genética molecular, ya que representó el conocimiento de la primera secuencia íntegra del genoma de un organismo eucarionte y que, además, fue puesta a disposición pública en diversos formatos y con información complementaria específica.

Como resultado de la secuenciación de su genoma, *Saccharomyces cerevisiae* se consolidó como uno de los sistemas experimentales modelos más completos del dominio eucarionte, considerándolo como un "recurso indispensable para el análisis detallado de la función genética así como de la arquitectura del genoma" (Clayton et al, 1997).

Una vez secuenciado el genoma, el siguiente nivel involucra el análisis de esta información desde diferentes enfoques; un primer enfoque consiste en el análisis de la secuencia misma, tratando de identificar señales informativas tales como promotores, regiones de activación o secuencias consenso de intrones (Staden, 1990), así como búsquedas de contenido, es decir, el análisis para discriminar entre secuencias con probabilidad de codificación y secuencias no codificantes (Gribskob et al, 1984; Shepherd, 1990; Staden, 1990; Gelfand,1990; Snyder y Stormo 1993-1995). Un segundo nivel analítico estaría representado por la búsqueda de similitudes entre una secuencia dada del genoma de la levadura y otras secuencias registradas en las bases de datos internacionales; este nivel de análisis, al caracterizar la secuencia problema en función de su similitud con otra(s) secuencia(s), constituye una de las estrategias primarias en la identificación de nuevos genes (Doolittle, 1987; 1990; Das et al, 1997). Un tercer nivel de análisis -y el definitivo- lo constituyen las diversas estrategias experimentales entre las cuales figuran la interrupción funcional de una región genética determinada y su caracterización fenotípica resultante, así como los análisis por control metabólico utilizando desafíos fisiológicos tales como choque térmico, estrés por frío, choque osmótico o diferentes limitaciones nutricionales (Oliver,1996; 1997; Goffeau et al, 1996; Fromont-Racine et al 1997). Aunada a las técnicas anteriores, recientemente se ha desarrollado una nueva metodología denominada Análisis Serial de la Expresión del Genoma (SAGE, en inglés), la cual utiliza sistemas de aislamiento de transcritos (RNA mensajeros) para identificar las regiones genómicas con actividad transcripcional (Velculescu et al, 1995; 1997); esta caracterización del conjunto de genes expresados durante el ciclo de vida de la levadura, el así llamado transcriptoma, ha aportado recientemente una extraordinaria capacidad de análisis de los patrones de expresión de los genes en la levadura, así como un nivel de comparación cuantitativo único de los genes expresados en una gran variedad de condiciones fisiológicas, metabólicas y de desarrollo (Velculescu et al , 1995). La identificación del conjunto de proteínas sintetizadas por la levadura, lo cual constituye el proteoma, es otro de los objetivos derivados de la secuenciación del

genoma; para la caracterización del proteoma se están utilizando dos técnicas fundamentales: la electroforesis bidimensional y la espectrometría de masas (Boucherie et al, 1996; Shevchenko et al, 1996; Fey et al, 1997; Garrels et al,1997).

Los tres niveles de análisis arriba mencionados representan las estrategias metodológicas hoy en día más utilizadas en la caracterización del genoma de *Saccharomyces cerevisiae*, así como de los genomas total o parcialmente secuenciados de otros organismos (Nowak, 1995; Goffeau et al, 1996; Koonin et al, 1996; Fromont-Racine et al, 1997; Oliver, 1997; Lashkari et al, 1997).

¿QUE SE CONOCE Y QUE NO SE CONOCE DEL GENOMA DE *Saccharomyces cerevisiae*?

Durante muchos años, el número exacto del complemento cromosómico de la levadura y el tamaño de su genoma permaneció sin resolver, debido a las bajas resoluciones de los métodos de separación y análisis empleados. Las estimaciones del tamaño del genoma de la levadura a partir de la cinética de reasociación del DNA y del número de cromosomas estimado por estudios genéticos, sugerían que los cromosomas más pequeños de la levadura tenían solamente unos pocos cientos de kilobases. Esta situación cambió hacia 1983 con el desarrollo de la técnica de electroforesis de campo pulsátil (Olson, 1991); una de sus primeras aplicaciones consistió en la separación del complemento cromosómico de la levadura, y para 1985, el cariotipo electroforético completo de este microorganismo fue publicado (Carle y Olson, 1985). Por este método fueron resueltos los 16 cromosomas de *Saccharomyces cerevisiae* así como también fue determinado el tamaño real del genoma haploide, estimado en aproximadamente 13,000 kilobases, sin incluir el genoma mitocondrial, el cual ha sido estimado en 78,500 nucleótidos (Goffeau et al, 1996). Estas primeras determinaciones del genoma de la levadura se realizaron utilizando la cepa S288C, la cual carece de algunas de las familias de genes que se encuentran en las cepas de levaduras empleadas en la industria cervecera, tales como las familias MEL (para

degradación de melobiosa), la familia MAL (para degradación de maltosa) y SUC (utilización de sacarosa); no obstante, la estimación antes mencionada se considera como la más cercana al tamaño real del genoma de la cepa silvestre (wild type) de la levadura.

De esta manera, el cariotipo electroforético aportó un mapa inicial de baja resolución del genoma, con la gran ventaja técnica de que al comenzarse a clonar fragmentos del DNA de la levadura, éstos se pudieron asignar con precisión a un cromosoma determinado, utilizando la hibridación por transferencia de geles. Con esta técnica se pudo ir asignando genes a cromosomas específicos y detallando el mapa genético hasta un nivel de resolución nunca antes logrado; al mismo tiempo, la electroforesis de campo pulsátil simplificó el análisis de los rearreglos cromosómicos tales como translocaciones y deleciones; incluso la aneuploidía (duplicación desigual de cromosomas), la cual es un fenómeno no frecuente en levaduras, puede ser detectada por inspección y comparación de las intensidades relativas de las bandas en el gel (Olson, 1991).

Al principio del proyecto de Secuenciación del Genoma de *Saccharomyces*, su mapa genético tenía asignados alrededor de 1,000 genes, ya fueran codificantes para proteínas o para RNA; el análisis inicial de la secuencia completa del genoma definió 5885 marcos de lectura abierta (open reading frames, ORFs), entre los cuales figuraban genes ya caracterizados así como proteínas hipotéticas aún sin caracterizar; además, se detectaron 140 genes codificantes para RNA ribosómico en un amplio arreglo en tándem en el cromosoma XII, 275 genes codificantes para RNA de transferencia (pertenecientes a 43 familias) y 40 genes codificantes para pequeños RNAs de localización nuclear; tanto los genes de RNA de transferencia como estos últimos se encuentran diseminados a lo largo de los 16 cromosomas (Goffeau et al 1996).

La densidad génica en la levadura, calculada a partir del número de ORFs y del tamaño del genoma es aproximadamente de un gene por cada 2 kilobases; en

comparación, en el genoma de *Eschecrichia coli* se presenta un gene por cada 1000 pares de bases, mientras que en *Caenorhabditis elegans*, un nemátodo, aparece un gen por cada 6 kilobases; en el genoma humano la densidad génica es aproximadamente de un gene por cada 30 kilobases (Nowak, 1995; Koonin et al, 1996). La alta compactación del genoma de *Saccharomyces* se ha relacionado con la escasez de intrones, pues mientras la levadura de fisión *Schizosaccharomyces pombe* muestra una densidad génica de un gene por cada 2.3 kilobases y el cuarenta por ciento de sus genes contiene intrones, en *Saccharomyces cerevisiae* solamente existen 4% de genes conteniendo intrones, y entre ellos, el 50% corresponden a genes codificantes para proteínas ribosómicas (Rodríguez-Medina y Rymond, 1994).

Otra característica notable del genoma de *Saccharomyces cerevisiae* consiste en el hecho de que 55 agrupaciones de genes se encuentran duplicadas entre los 16 cromosomas. Estas regiones duplicadas representan más del 30% del genoma total (Clayton et al, 1997). Las proteínas derivadas de el (los) evento(s) de duplicación representan el 13% de todas las proteínas de la levadura -el también llamado proteoma- e incluyen factores de transcripción, cinasas, miosinas, ciclinas, feromonas, así como genes involucrados en el metabolismo aeróbico y anaeróbico y genes codificantes para proteínas de membrana (Wolfe y Shields,1997).

Un hecho interesante que ha revelado el análisis del genoma de la levadura es la variación en el contenido de guanina más citosina (G+C) a lo largo de los distintos cromosomas; estas variaciones se correlacionan con las variaciones en la densidad génica y con la frecuencia de recombinación, ya que se ha reportado que los dominios ricos en G+C exhiben alta densidad génica y, en el caso de los cromosomas más pequeños -I, III, VI y IX- estas regiones corresponden a áreas de alta frecuencia de recombinación (Goffeau, 1996).

Pasando al análisis del proteoma, es decir, al conjunto de todas las proteínas que una célula sintetiza, el análisis por computadora basado en las similitudes de

secuencias de aminoácidos codificadas por la levadura respecto a las proteínas de función conocida, ha presentado el siguiente panorama: la levadura asigna el 11% de su proteoma a funciones metabólicas generales, 3% a la producción y conservación de energía -incluyendo procesos aeróbicos y anaeróbicos-; 3% a replicación del DNA así como a procesos de reparación y recombinación mediados por la proteína REC; 7% al proceso de transcripción y 6% a traducción .

Además de lo anterior, 430 proteínas están involucradas en transporte intracelular y señalización y 250 proteínas han demostrado tener una función estructural; asimismo, se han identificado cerca de 200 factores de transcripción y 250 proteínas de transporte (Goffeau, 1996). Las estimaciones anteriores se refieren solamente a aquellas proteínas para las cuales se han identificado una o varias similitudes significativas; al integrar el número de proteínas caracterizadas bioquímicamente -aunque no totalmente en su papel fisiológico-, resulta que existen alrededor de 2,300 ORFs en *Saccharomyces cerevisiae* sin función conocida y *sin similitud con otras secuencias de función caracterizada (Clayton et al, 1997).*

Una vez obtenida la secuencia completa de los 16 cromosomas de la levadura, una de las primeras fases del análisis realizado por los equipos responsables de este Megaproyecto, fue el identificar los marcos de lectura abierta -ORFs- presentes en la secuencia completa. Los criterios para asignar un ORF a una determinada secuencia fueron los siguientes:

a) Presencia de un codón de inicio -ATG- y un codón de término -TAA, TAG, TGA- en la misma fase de lectura.

b) Presencia de al menos 100 codones codificantes contiguos, incluyendo el codón inicial ATG.

c) Que un ORF no esté contenido dentro de otro ORF, es decir, se excluyen los ORFs de menor tamaño presentes en un ORF mayor, ambos correspondientes a la misma fase de lectura.

d) ORFs conteniendo menos de 150 codones de extensión y con un valor de CAI menor a 0.110 se consideran ORFs cuestionables. El valor de CAI se refiere al Codón Adaptation Index (Indice de Adaptación de Codones) y representa la medida de utilización preferencial de codones por los genes de la levadura; un alto valor de CAI significa que el ORF - o el gen, en su caso- utiliza de manera preferencial ciertos codones (en el caso de codones sinónimos); este sesgo en la utilización de codones es una característica compartida por los genes con un alto nivel de expresión y es especie-específico. Un bajo valor de CAI, por otra parte, indica, generalmente, un bajo nivel de expresión, así como una utilización de codones que no es la que con mayor frecuencia se detecta en la levadura. El CAI, por lo tanto, es una estimación probabilística del grado de utilización de codones codificantes y del nivel de expresión de un gene, aplicable también, de manera estadística, como un elemento de predicción de la capacidad codificación de un ORF (Sharp y Li ,1987 ; Dujon et al, 1994 ; Li y Luo ,1996). El criterio de considerar como cuestionable a ORFs con valores de CAI menores a 0.110 y con una extensión menor a 150 codones se basa en consideraciones estadísticas ya que existen genes caracterizados los cuales, sin embargo, tienen longitudes menores a 150 codones y muestran valores de CAI inferiores a 0.110 - este límite es un valor empírico derivado de la cuantificación de la utilización preferencial de codones y es, por lo tanto, una cota operacional. Utilizando los criterios antes mencionados, el proyecto de secuenciación del genoma de *Saccharomyces cerevisiae* reveló la existencia de 6,275 ORFs que, teóricamente, podrían codificar proteínas de una longitud mayor de 99 aminoácidos (297 nucleótidos); en estos 6,275 ORFs están incluídos los genes ya caracterizados así como las secuencias sin función conocida. Del número total de ORFs, 390 fueron considerados como cuestionables por las razones anteriormente mencionadas (Goffeau et al, 1996); de esta manera, se postula que

en la levadura existen aproximadamente 5885 genes codificantes para productos proteicos, aunque actualmente se estiman alrededor de 2,300 ORFs o genes putativos sin función conocida y sin similitud con genes de función caracterizada. Se ha señalado que estos 2,300 genes "huérfanos" son un vasto territorio virgen para los genetistas y los químicos de proteínas y constituyen, actualmente, uno de los programas de investigación más ambiciosos para identificar su función (Clayton, 1997); de hecho, el primer gran paso en esta dirección lo constituye la caracterización del transcriptoma de la levadura; este análisis ha revelado la presencia de 60,633 transcritos correspondientes a 4,665 genes con función conocida y no conocida, y con niveles de expresión variables entre 0.3 y 200 transcritos por gen (Velculescu et al, 1997).

Junto a este programa de caracterización de la expresión del genoma de esta levadura, recientemente se ha integrado un nuevo proyecto de análisis denominado Proyecto de Deleción del Genoma de *Saccharomyces cerevisiae*; este programa de investigación está integrado por ocho laboratorios norteamericanos y ocho laboratorios europeos; su objetivo es generar un conjunto completo de mutantes por deleción con el propósito de asignar una función específica a cada uno de los ORFs de función desconocida a través del análisis fenotípico de las mutaciones generadas (Baudin et al, 1993; Guldener et al, 1996). La estrategia experimental está basada en generar deleciones, utilizando la Reacción en Cadena de la Polimerasa (PCR), con el fin de producir interrupciones completas desde el codón de inicio al de término en un ORF determinado. Como parte del proceso de deleción, cada una de las secuencias es marcada específicamente con un oligómero de 20 nucleótidos. La presencia de estas marcas o etiquetas puede ser detectada vía hibridación a un conjunto de oligonucleótidos, permitiendo que el crecimiento de fenotipos individuales pueda ser analizado en paralelo (Baudin et al, 1993; Wach et al, 1994; Shoemaker et al, 1996).

Ahora bien, haciendo una recapitulación de lo que se conoce del genoma de *Saccharomyces cerevisiae* resulta que, del total de ORFs registrados,(6,275), 390 ORFs (6%) son considerados como cuestionables, por lo que restan 5,885 ORFs con potencial codificante; de estas secuencias, se tienen identificados como genes con función caracterizada 3,025 (5 %), y 2,860 (49%) sin función conocida ni similitud registrada.

Por otro lado, los datos obtenidos de la caracterización del transcriptoma revelan la expresión de 4,665 genes; este número, respecto al total de ORFs con potencial codificante -5,885-, registra una diferencia de 1,220 secuencias cuyo papel se desconoce. Si existen 2,860 secuencias sin función conocida, de las cuales 1,220 parecen no codificar ningún producto -aunque se desconoce cuáles son estas secuencias -, el problema que se plantea estriba en discriminar cuáles de estos 2,860 ORFs tienen potencial codificante, y en torno a los cuales no existe información sobre su rol genético. En este contexto se ubica el objetivo de la presente investigación de tesis, la cual está inserta en el Proyecto de Análisis del Genoma de *Saccharomyces cerevisiae* que se lleva a cabo en el Laboratorio de Biología Molecular en la Facultad de Ciencias de la UNAM bajo la dirección del M. en C. Victor Valdes López.

El problema, en concreto, puede ser planteado de la siguiente manera: ¿cómo se pueden reconocer, entre los 2860 ORFs, los que son codificantes?

Otra forma de exponer este programa de investigación es: ¿cuáles son los elementos informacionales que son indicativos de la capacidad de codificación en una secuencia de nucleótidos?

LA ELECCION DE UNA ESTRATEGIA DE ANALISIS DEL GENOMA.

Existen dos tipos de información que pueden ser usados para la identificación de regiones potencialmente codificantes: la primera consiste en la presencia de

secuencias señal tales como promotores, señales de activación, potenciadores (enhancers) o secuencias consenso en las zonas de procesamiento intrón-exón. El segundo tipo de información consiste en la identificación de pautas o formatos en la seriación de nucleótidos de una región genómica determinada; tal tipo de búsqueda también se denomina búsqueda por contenido y, a diferencia del tipo de información por señal, la búsqueda por contenido utiliza un mayor número de criterios estadísticos (Staden, 1984; 1990; Sharp y Li, 1987; Shepherd, 1990; Gelfand, 1990; Fickett y Tung, 1992; Snyder y Stormo, 1993; Barry et al, 1996).

Dado que en el genoma de la levadura las secuencias consenso de promotores, potenciadores (enhancers) y señales de terminación no están bien caracterizadas, y dada la bajísima presencia de intrones en el genoma -aproximadamente el 4 %- (Goffeau et al, 1996), el desarrollo de métodos de búsqueda por contenido, ha sido de gran utilidad en la identificación de regiones con potencial de codificación (Stormo, 1987; Staden, 1990).

Entre estos últimos métodos, desarrollados para la determinación de regiones genéticas en secuencias de ácidos nucleicos, los de mayor utilidad son de dos tipos: Aquellos que se basan en la cuantificación de la utilización preferencial de codones, esto es, en el uso desigual de codones sinónimos (Ikemura, 1985; Sharp y Li, 1987), y aquellos basados en la búsqueda de asimetrías posicionales de las bases en codones (Shepherd, 1981; 1990); este segundo tipo de método busca una correlación entre nucleótidos en posiciones definidas en tripletes de bases, incluso sin tomar en cuenta la utilización preferencial de codones (Stormo, 1987).

En la actualidad existe un gran número de algoritmos diseñados para la identificación de regiones codificantes en genomas, pero casi todos ellos son variantes de los dos métodos básicos arriba mencionados. Evidentemente, ambas estrategias de análisis comparten un denominador común, y este consiste en la identificación inicial de un marco de lectura abierta -ORF- en una región dada del genoma; esta identificación tiene como finalidad el discriminar entre ORFs

posiblemente codificantes y ORFs generados al azar. Para el genoma de la levadura, la probabilidad de la ocurrencia al azar de un ORF igual o mayor a 100 aminoácidos -300 pares de bases- ha sido estimada en menos del 0.2 % (Oliver et al., 1992); también ha sido calculado que la probabilidad, en una secuencia aleatoria, de aparición de un ORF con una extensión de 50 codones es menor al 10 %. Ya que existen 3 codones de término, en una secuencia al azar de composición de bases uniforme, la longitud promedio de un ORF generado en forma aleatoria es de 21 codones (Staden, 1990). Puesto que la mayoría de las proteínas estudiadas hasta ahora tienen más de 100 aminoácidos de longitud, la búsqueda aislada de ORFs de mayor extensión es una estrategia de análisis estadísticamente válida; lo anterior está basado en que la probabilidad de aparición de un ORF al azar mayor de 100 codones es menor de 1% (Stormo, 1987; Oliver et al, 1992). Por otra parte, la existencia de intrones en los genomas eucariontes disminuye la eficacia de los métodos de búsqueda de ORFs largos en secuencias de DNA, aunque la escasez de intrones en el genoma de la levadura descarta en gran parte esta dificultad. Así pues, la identificación de marcos de lectura abierta es el paso inicial en la búsqueda de una región codificante putativa, aunque el problema principal, una vez identificados los ORFs, aún subsiste: ¿cómo identificar un ORF codificante respecto a otro ORF generado al azar?.

Los criterios utilizados actualmente en búsquedas por contenido para discriminar entre estas regiones se basan en dos factores principales: a) uso desigual de codones en secuencias codificantes; b) frecuencia diferencial de las bases en las 3 posiciones del codón en secuencias codificantes. Esta condición se ha denominado variación periódica de bases.

Respecto a los métodos utilizados para la identificación de ORFs con potencial codificante basados en la presencia diferencial de codones, se han desarrollado diversos algoritmos que emplean índices de utilización diferencial; estos índices, entre los cuales destaca el CAI, han sido útiles en este contexto, particularmente al localizar genes de expresión elevada; sin embargo, las dificultades aparecen en la

interpretación de secuencias con valores bajos de CAI y de esta manera, mientras un valor alto de CAI es una buena indicación respecto a la existencia de una región codificante en un marco de lectura abierta, un valor bajo de CAI puede indicar: 1) un bajo nivel de expresión de un gen; 2) un gen adquirido por transferencia horizontal (gen heterólogo) y 3) una región no codificante en la que por azar no aparecen codones de término (Sharp y Li, 1987; Dujon et al, 1994). A este respecto es importante señalar que el valor de CAI estimado para una secuencia aleatoria se ubica en 0.17, lo cual señala que valores de CAI más bajos que éste indicarían secuencias no codificantes o con muy pequeña probabilidad de codificación (Sharp y Li, 1987); sin embargo, algunos valores de CAI asignados a genes de *Saccharomyces cerevisiae* son del orden de 0.06, lo cual es un valor casi tres veces menor al asignado a una secuencia aleatoria. Esta es la razón por la cual Sharp y Li, los creadores del índice de adaptación de codones -CAI- son cautelosos al aplicar este criterio al análisis global de genomas, puesto que los niveles diferenciales de expresión genética así como la misma extensión de los ORFs pueden introducir desviaciones importantes en los valores de CAI y conducir, por lo tanto, a interpretaciones equivocadas respecto a la capacidad de codificación de una secuencia (Sharp y Cowe, 1991; Oliver et al, 1992; Dujon et al, 1994; Li y Luo, 1996).

Respecto al tamaño de los ORFs y su relación con su potencial de codificación, los investigadores involucrados en el Proyecto de Secuenciación del Genoma de *Saccharomyces cerevisiae* definieron un umbral mínimo de 100 codones arriba del cual un marco de lectura abierta es considerado como una secuencia codificante putativa (Dujon, et al, 1994). Para evaluar la posibilidad de que algunos ORFs predichos pudieran haberse generado al azar, estos investigadores generaron una secuencia aleatoria del mismo tamaño y composición de bases que la secuencia del cromosoma 11 de la levadura. Esta secuencia aleatoria generó 37 ORFs de tamaño comprendido entre 101 y 150 codones con valores promedio de CAI de 0.10, es decir, valores por debajo del umbral de 0.17 definido por Sharp y Li para

valores de CAI relacionados con secuencias al azar; por esta razón, secuencias con longitudes menores de 150 codones y mostrando valores de CAI menores a 0.110 fueron consideradas por los autores del proyecto de secuenciación como ORFs cuestionables. De hecho, los mismos autores señalan que la existencia de ORFs como genes funcionales ha sido sistemáticamente examinada, utilizando, como criterios de posible identificación, su índice de adaptación de codones junto con su tamaño; no obstante esto, los autores reconocen que no existe una correlación general entre el valor de CAI y el tamaño del ORF (Oliver, et al, 1992; Dujon, et al, 1994; Mewes, et al, 1997). Actualmente se tienen identificados 85 genes con una longitud menor a 100 codones con una función proteíca asignada (MIPS, 1998); a este respecto, Das et al (1997) llamaron la atención al hecho de que la cantidad de ORFs registrados con una longitud entre 100 y 110 codones es tres veces superior al número de ORFs de longitudes mayores, pero que los ORFs de 100 codones corresponden casi en su totalidad al grupo en el que no se ha reportado similitud con otras secuencias; lo anterior significa que las secuencias de ORFs con funciones asignadas corresponden a las más estudiadas y por lo tanto están representadas en forma desproporcionada en las bases de datos, generando un sesgo en la identificación de sus similares en la levadura. Este hecho señala la necesidad de analizar con más cuidado la población de ORFs de tamaño reducido en la cual puedan, potencialmente, descubrirse nuevas proteínas de las que aun está por caracterizarse su función bioquímica y biológica (Oliver, 1996; Barry et al, 1996; Das et al, 1997; Clayton et al, 1997; Andrade et al, 1997; Olivas et al, 1997; Rudd et al, 1998).

Como prueba de lo anterior, es necesario mencionar que el gene más pequeño conocido consta tan sólo de 6 pares de bases, codificantes para un minipéptido de 2 aminoácidos, el dipéptido formil-metionina-isoleucina, el cual está involucrado en mecanismos de toxicidad celular; a este gene se le ha denominado bar y está considerado el minigen más pequeño (Valadez et al, 1997); también se han reportado minigenes de 5 y 7 codones con funciones de resistencia a antibióticos,

tales como el pentapéptido E (15 codones) y la microcina C7, un heptapéptido lineal que inhibe la síntesis de proteínas en Enterobacterias (González Pastor et al, 1994). En *Saccharomyces cerevisiae*, los péptidos más pequeños descubiertos hasta ahora corresponden a proteínas ribosomales y son codificadas por genes de 25 codones (MIPS, 1998).

Toda esta evidencia apunta al hecho de que el pequeño tamaño de un ORF y un bajo valor de CAI no son factores que, por sí solos, puedan -o deban- ser elementos discriminativos de secuencias identificadas como marcos de lectura abierta, independientemente de que muestren, o no, similitud con secuencias reportadas.

En relación con los métodos por búsqueda de contenido, basados en la autocorrelación de nucleótidos asociada a tripletes, el más conocido es el algoritmo de Shepherd (Shepherd, 1981; 1990); su utilización está basada en el reconocimiento -en una secuencia de nucleótidos -, de un patrón de posicionamiento de las bases en los codones, es decir, cada base muestra posiciones preferenciales en el codón, ajustándose, en términos estadísticos, al formato RNY, en donde R=purinas (adenina y guanina) Y=pirimidinas (citosina y timina) y N=purina o pirimidina. El análisis de múltiples genes a lo largo de la escala filogenética ha mostrado que las secuencias codificantes muestran, en sus codones, una periodicidad del tipo RNY; la explicación de esta periodicidad ha sido muy debatida, aunque existe cierto consenso de que el exceso de codones RNY es probablemente atribuible a la mayor abundancia de RNAs de transferencia correspondientes a codones del tipo RNY así como a una mayor presencia de aminoácidos codificados por codones de tipo RNY en proteínas (Staden, 1984;Jukes,1996). Este método es independiente del conocimiento que se tenga de la utilización preferencial de codones en la secuencia analizada; como consecuencia de las propiedades arriba mencionadas, el algoritmo de Shepherd indica la fase de lectura más probable de la secuencia, es decir, este programa reconoce el marco de lectura con mayor preponderancia de codones con el formato RNY. Este método ha sido empleado en el análisis de un gran número de

genes así como de regiones no codificantes, siendo aplicado al primer genoma del cual se tuvo la secuencia completa, el del virus ϕX174 (Shepherd, 1981); el programa fue valorado al analizar 2,000 secuencias tomadas al azar del banco de datos del EMBL (European Molecular Biology Laboratory), e incluyendo todo tipo de organismos, con el resultado de que en el 90% de los casos, el algoritmo fue un excelente elemento de predicción de las fases de lectura correctas, así como en la discriminación de secuencias codificantes y no codificantes (Staden, 1984; Stormo, 1987; Watson et al, 1987; Shepherd, 1990). Dado su alto nivel de confiabilidad, su versatilidad de aplicación a múltiples tipos de secuencias y su facilidad de manejo como programa informático, el algoritmo de Shepherd fue escogido por nosotros como algoritmo básico de análisis del genoma de *Saccharomyces cerevisiae*.

2- OBJETIVO.

En este contexto se ubica la presente investigación, la cual tuvo por objetivo la valoración de secuencias de tamaño reducido -menores de 150 codones- del genoma de la levadura, incluyendo los ORFs considerados como cuestionables, así como el análisis de secuencias de mayor tamaño -300 codones o mas -, todas ellas sin función conocida, discriminando los ORFs con potencial de codificación respecto a los generados en forma aleatoria, mediante la aplicación del algoritmo de Shepherd.

3- ESTRATEGIA DE ANALISIS.

- Breve descripción de la base de datos.

La secuencia completa del genoma de la levadura, es accesible a través de la Red Internet y presenta, en cualquiera de los dos bancos de datos en los cuales se localiza, la siguiente nomenclatura: Una letra inicial mayúscula Y igual para todos los ORFs (Y para Yeast = levadura), una segunda letra mayúscula indicando el número de cromosoma: A, representando el cromosoma I, B, por el cromosoma II, C, por el cromosoma III, y subsecuentes hasta la letra P, la cual representa al cromosoma XVI; una tercera letra mayúscula, L o R, indicando la ubicación del ORF en el brazo izquierdo -L = left- o en el brazo derecho -R = right- , de un determinado cromosoma; a continuación un número indica la posición del ORF a partir del

centrómero, 001, 023, 148, etc., y una letra minúscula w ó c, indicando si la secuencia se encuentra en la cadena Watson con polaridad 5'→3' o en la cadena Crick con polaridad 3'→5'; así, un ORF del genoma de la levadura es representado en los bancos de datos en la siguiente forma: (ejemplo) YGL173w. A cada ORF registrado en las bases de datos antes mencionadas se le asigna un par de coordenadas numéricas, las cuales señalan el número del nucleótido en el cual se inicia el ORF y el número del nucleótido en el cual finaliza (en la cadena Watson, un ORF siempre tiene coordenadas ascendentes, mientras que si se encuentra en la cadena Crick mostrará coordenadas descendentes).

La secuencia de cada ORF se puede presentar en diversos formatos, siendo el más adecuado para su conversión a programas de análisis el formato en el que la secuencia se presenta como una asociación de nucleótidos sin anotaciones marginales y sin encabezados numéricos; desde luego que el contexto físico de una secuencia, tal como el cromosoma en el que se encuentra y su posición en el mapa así como su nivel de similitud con otras secuencias, o su valor de CAI, son elementos muy importantes en la valoración de un ORF, pero esta información es anexada por separado en la recuperación de la secuencia. Esta recuperación puede ser llevada a cabo mediante diversas entradas: por registro del ORF, por el nombre del gen anotado o por sus correspondientes coordenadas, indicando el cromosoma. Desde luego que en la recuperación de una secuencia es posible anexar también cierta extensión de nucleótidos que flanqueen al ORF elegido en cualquier dirección; esto es importante para observar y analizar el contexto que rodea a una determinada secuencia; en la presente investigación, cada ORF seleccionado se recuperó junto con una extensión de 300 pares de bases contiguos al extremo 5' y 300 pares de bases contiguos al 3'.

- Selección de ORFs.

De los 16 cromosomas que integran el complemento genético haploide de *Saccharomyces cerevisiae*, se seleccionaron los cromosomas con número impar,

ya que entre estos se encuentran cromosomas representativos del genoma en su conjunto tales como cromosomas de tamaño pequeño -300 a 500 kilobases- (cromosomas I,III,V y IX), de tamaño intermedio -700 a 900 kilobases- (cromosomas XI y XIII), así como cromosomas de más de una megabase (cromosomas VII y XV). Como la densidad génica es similar en los 16 cromosomas, el muestreo de ORFs es un reflejo del número de secuencias no caracterizadas (ORFs cuestionables e hipotéticos) presentes en cada cromosoma en función de su longitud. De esta manera, los cromosomas grandes tendrán un número mayor de secuencias no caracterizadas, mientras que los cromosomas de tamaño intermedio o pequeño tendrán un número de ORFs proporcional a su extensión.

Como los ORFs más abundantes en el genoma de la levadura son aquellos de un tamaño comprendido entre 100 y 150 aminoácidos (Das et al, 1997; Andrade et al, 1997), se recuperaron las secuencias de todos los ORFs cuestionables presentes en los ocho cromosomas seleccionados, sumando un total de 126 secuencias, con la condición de que aparecieran registradas tanto en la base de datos de MIPS como en SGD; estas secuencias muestran una extensión aproximada de 300 pares de bases, correspondientes aproximadamente a 100 aminoácidos; también, de cada cromosoma seleccionado se muestrearon ORFs considerados como hipotéticos, es decir presumiblemente codificantes, de una longitud variable entre 300 y 900 pares de bases, con el fin de cubrir un rango amplio de longitudes en las secuencias sujetas a análisis. Así, el total de ORFs considerados como hipotéticos analizados fue de 157 secuencias, correspondiendo 103 a ORFs de 300 pares de bases y 54 a ORFs de 900 pares de bases o mayores; de esta manera, el total de secuencias analizadas respecto a su capacidad de codificación fue de 283. La lista completa de ORFs y genes analizados aparece en el Apéndice de esta tesis.

Como un control positivo del programa de Shepherd sometimos a análisis las secuencias completas de genes caracterizados funcionalmente, también denominados genes anotados, y de tamaño similar a las de los ORFs no caracterizados, analizando un total de 101 genes presentes en los ocho

cromosomas seleccionados. El programa de Shepherd reconoció el formato RNY en el 85% de estos genes, presentándose, en promedio, en el 74% ±12% de la longitud del ORF en su fase codificante, considerándose el valor de 62% como el umbral mínimo de reconocimiento del formato RNY (Figura 1). Como controles negativos del algoritmo utilizamos secuencias al azar con longitudes de 100, 300 y 1,600 pares de bases, generadas mediante el programa informático Shuffle, desarrollado por Genetics Computer Group (GCG) de la Universidad de Wisconsin; este programa, al serle dada una secuencia de bases, redistribuye al azar los nucleótidos de la secuencia, manteniendo constante su composición nucleotídica. Las secuencias al azar fueron analizadas por el programa de Shepherd, identificándose el patrón RNY entre el 30 y el 40% de la longitud total de la secuencia, en promedio, considerándose estas secuencias como controles negativos (Figura 1).

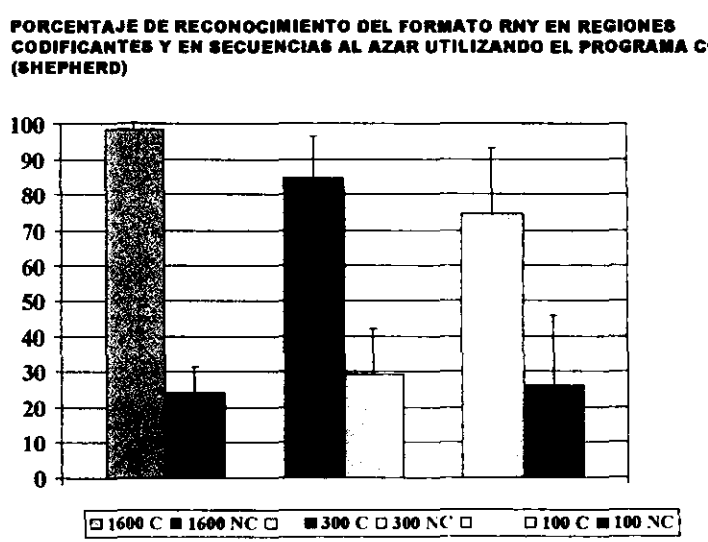
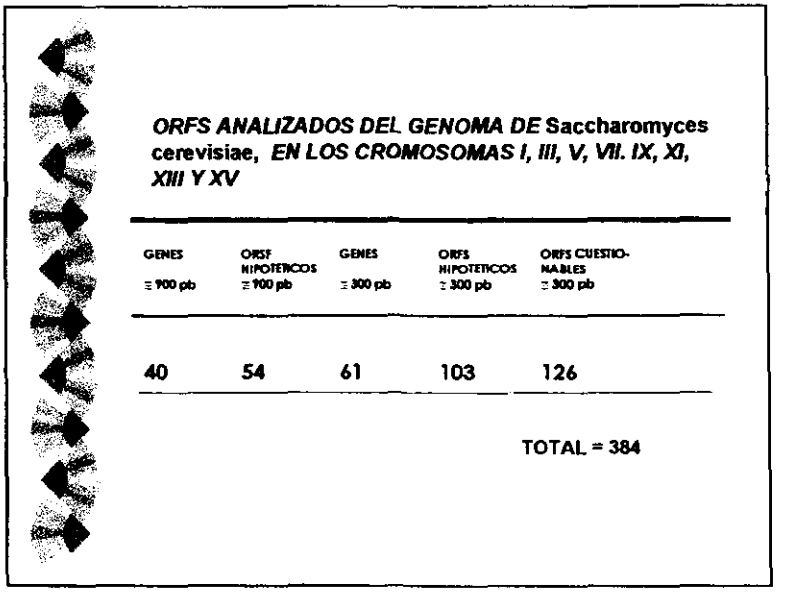


FIGURA 1.- La gráfica señala las diferencias en los porcentajes de reconocimiento del serial RNY por el programa de Shepherd (COD RNY) en secuencias codificantes (C) y no codificantes (NC) de 1600, 300 y 100 pares de bases, utilizadas como controles.

La suma total de secuencias estudiadas del genoma fue de 384- incluyendo ORFs cuestionables e hipotéticos y genes anotados, lo cual equivale al 6.3% del genoma de *Saccharomyces cerevisiae* (Stanford Genome Data Base, 1998)(Figura 2).



ORFS ANALIZADOS DEL GENOMA DE* Saccharomyces cerevisiae, *EN LOS CROMOSOMAS I, III, V, VII. IX, XI, XIII Y XV

GENES ≥ 900 pb	ORSF HIPOTETICOS ≥ 900 pb	GENES ≥ 300 pb	ORFS HIPOTETICOS ≥ 300 pb	ORFS CUESTIO-NABLES ≥ 300 pb
40	54	61	103	126

TOTAL = 384

FIGURA 2.- La tabla muestra el número de ORFs analizados en el genoma de la levadura, correspondiendo a secuencias de genes anotados y a secuencias de ORFs hipotéticos y cuestionables.
El total de secuencias analizadas equivale aproximadamente al 6% del genoma de *Saccharomyces cerevisiae*.

-Descripción del Algoritmo de Análisis (Programa COD RNY de Shepherd).

Una secuencia de nucleótidos de una longitud determinada, L, es convertida a una serie de purinas R, y pirimidinas Y, y a partir del primer nucleótido, la secuencia es examinada en los 3 marcos de lectura posibles en una determinada polaridad

para definir cuál de las 3 fases muestra la menor desviación de una secuencia patrón formada únicamente de tripletes RNY, donde N = purinas ó pirimidinas. El programa divide la secuencia bajo análisis en grupos de 3 bases -tripletes- analizando los primeros tres nucleótidos para la fase de lectura 1, los nucleótidos 2, 3 y 4 para la fase de lectura 2 y los nucleótidos 3, 4 y 5 para las fase de lectura 3. Por iteración, el programa analiza toda la longitud dada registrando cuál marco de lectura se ajusta con mayor aproximación al serial RNY. El marco de lectura que presenta la menor desviación respecto a este codón de referencia es registrado y graficado en el punto medio de la longitud L, indicando que esta fase de lectura es la que muestra el menor número de desviaciones respecto al codón serial. El procedimiento se repite después de avanzar S número de bases (paso de avance) siempre en la dirección 5'→3' a lo largo de la secuencia seleccionada.

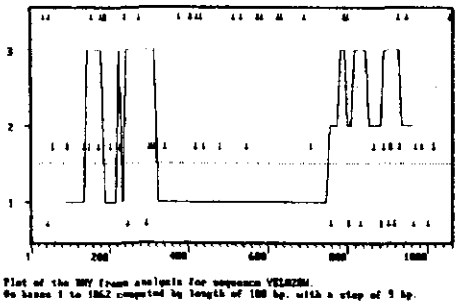
Un punto importante de señalar es el hecho de que en la selección de la extensión del paso de avance S, mientras más pequeño sea este paso mayor será la sensibilidad del análisis, ya que al mantener sin variación la longitud de análisis (ventana) y avanzar una unidad (tres pares de bases es la mínima unidad de paso de avance en el programa de Shepherd), el cálculo se vuelve a repetir, obteniéndose un mayor detalle de cada sección de la secuencia; esto es debido a que cada ventana se sobrelapa con la sección previa en toda su longitud, menos la unidad de avance (Shepherd, 1981, 1990; Staden, 1990). De esta manera, si escogemos una sección de longitud L de 60 nucleótidos y un paso de avance S de 3 nucleótidos, cada triplete sucesivo será inspeccionado desde la base 1 a la base 60 (fase 1), desde la base 2 hasta la base 61 (fase 2) y desde la base 3 a la base 63 (fase 3); una vez que el programa haya analizado los primeros 60, avanzará hasta el nucleótido 63 y realizará nuevamente la inspección de 60 nucleótidos, comenzando este nuevo análisis a partir del nucleótido 3 de la sección previa, y por iteración, hasta el final de la secuencia.

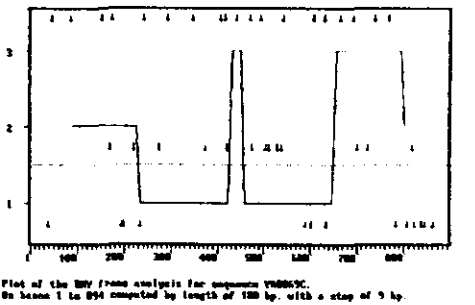
En los casos en donde se presentan iguales valores mínimos de desviación en 2 ó 3 fases de lectura respecto al serial RNY, el cómputo se lleva a cabo sobre una

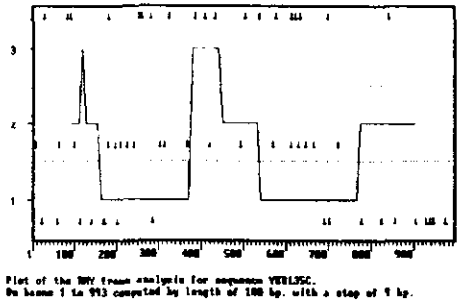
extensión mayor pero aún centrado sobre el mismo punto medio de la sección previa.

Los resultados son presentados en forma gráfica: la secuencia es representada en el eje de las X y los resultados analíticos son graficados en el eje de las Y, en el cual se registran los 3 marcos de lectura, con sus respectivas señales de terminación si es que están presentes. La gráfica indica cuál de los 3 marcos de lectura presenta la menor desviación del serial RNY en cada triplete a lo largo de la secuencia, señalando con un punto dicha posición; si una fase de lectura muestra la menor desviación en muchas posiciones consecutivas (tripletes contiguos), los puntos producirán una línea continua en el marco de lectura correspondiente, indicando la fase con mayor probabilidad de codificación potencial. Por el contrario, si los puntos muestran alternancia entre 2 ó 3 fases, la gráfica mostrará una serie de picos y mesetas, indicando desviaciones respecto al formato RNY entre 2 ó 3 marcos. En otros casos una fase de lectura puede mostrar una preponderancia de la seriación RNY -línea continua- punteada por desviaciones hacia los otros marcos de lectura (estos picos parásitos se han interpretado como eventos mutacionales).

Con base en observaciones de los patrones de gráficas obtenidas de secuencias codificantes, se ha propuesto que las regiones codificantes muestran en general pocos cambios de marcos de lectura, y cuando esto sucede, se observa principalmente en genes con intrones, en donde los exones pueden ser leídos en distintas fases, mientras que en regiones no codificantes no existe una preferencia hacia una fase de lectura determinada en la cual se conserve el formato RNY (Stormo, 1987; Shepherd , 1990). La cuantificación del porcentaje del serial RNY se llevó a cabo midiendo la extensión del ORF en la cual se presenta el formato RNY, considerándose el total de bases como el 100%; de esta manera, los ORFs analizados presentaron porcentajes variables del serial RNY, lo cual, según el criterio de Shepherd, está en relación con su potencial de codificación (Figura 3).







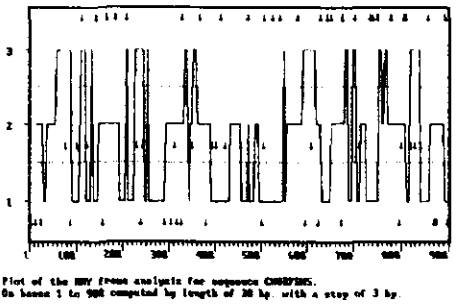


FIGURA 3.- Estas cuatro gráficas de Shepherd muestran los distintos porcentajes de reconocimiento del formato RNY en diferentes secuencias : 100% en la gráfica del extremo superior izquierdo, 75% en la gráfica del extremo superior derecho y 50% en la gráfica inferior izquierda; en estas tres gráficas la fase de lectura abierta se encuentra en el marco 1. La gráfica inferior derecha, por su parte, muestra el comportamiento de una secuencia generada al azar en la cual no se aprecia un marco de lectura preponderante, correspondiendo a cada fase aproximadamente un 30% del serial RNY.

El programa de Shepherd, incluido en el paquete informático PC Gene de Intellegenetics Inc, and Genofit SA, Ver. 6.01 (1989), con denominación COD RNY, se corrió en una computadora Acer 386 con 640 kb de memoria RAM; las

secuencias fueron copiadas en formato GCG y recuperadas de la base de datos Stanford Genome Data Base (SGD), utilizando el listado de ORFs del genoma de *Saccharomyces cerevisiae* elaborado por el Martinsried Institute For Protein Secuences (MIPS) en Munich, Alemania.

Programa BLAST.- Este programa de búsqueda de similitudes entre secuencias de proteínas fue utilizado para tratar de identificar secuencias de aminoácidos correspondientes a ORFs cuestionables e hipotéticos; el programa BLAST (Basic Alignment Search Tool) es un algoritmo de búsqueda de similitudes entre secuencias, comparando segmentos similares entre una secuencia dada y las secuencias registradas en las bases de datos; el programa evalúa la significancia estadística de las similitudes encontradas y reporta sólo los apareamientos que satisfagan un umbral de significancia establecido previamente.

4- RESULTADOS.

Los resultados de los análisis realizados son descritos para cada cromosoma por separado.

CROMOSOMA I -. Para el cromosoma 1, MIPS señala 2 ORFs cuestionables, YAL 056c-A y YAL 034c-B, de 116 y 117 codones respectivamente. Stanford no tiene registrados estos ORFs, por lo que no se recuperaron dichas secuencias. Este cromosoma tiene 15 ORFs registrados como proteínas hipotéticas sin similitud con alguna secuencia reportada; de estos 15 ORFs, *ocho secuencias mostraron* porcentajes de reconocimiento del serial RNY mayores de 62%, considerado como cota mínima de reconocimiento, de las cuales seis tienen una longitud promedio de 300 pares de bases y las dos restantes una longitud promedio de 900 pares de bases; estos ocho ORFs, considerados como genes putativos según el algoritmo de análisis, muestran un valor de CAI de 0.10 en promedio, el cual es considerado un valor bajo, indicando que en estas secuencias no es manifiesta la utilización preferencial de codones consignada para los genes de la levadura. El único ORF que mostró un valor de CAI elevado, 0.48, y con un porcentaje de concordancia con el formato RNY de 100% fue YAR 020c el cual muestra similitud con una proteína registrada en la base de datos de MIPS (Figura 4). Los ocho ORFs reconocidos con potencial codificante por el programa de Shepherd fueron: YAL 045c (102 codones), YAR 020c (55 codones), YAR 040c (119 codones), YAR 064w (99 codones), YAR 069c (97 codones), YAR 070c (99 codones),YAL 056w(847 codones) y YAL 027w (261 codones).

Respecto a los genes anotados y utilizados como control, el algoritmo de Shepherd reconoció el formato RNY en tres de los cinco genes analizados; los dos genes no reconocidos por Shepherd corresponden a una proteína de membrana y a una proteína del sistema del holo-citocromo C, con una CAI promedio para ambos genes de 0.12.

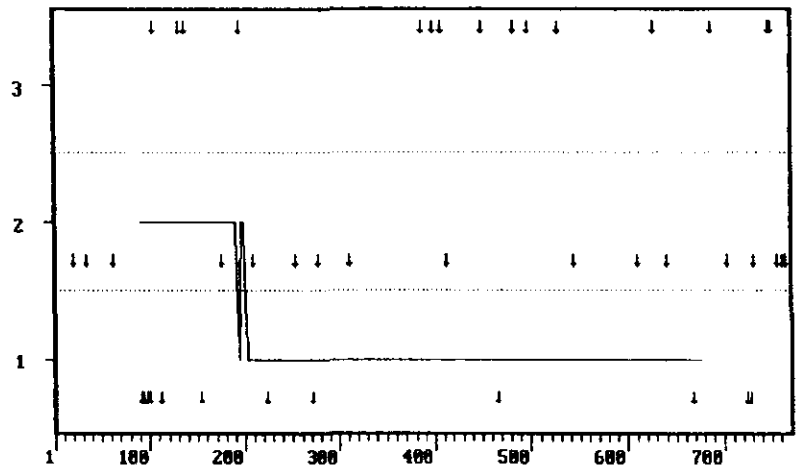


FIGURA 4.- En esta secuencia se observa que el ORF correspondiente a la secuencia YAR020c comienza aproximadamente en la base 270 y termina en la base 465 (eje de abcisas), en el marco de lectura 1 (eje de ordenadas); el algoritmo de Shepherd identifica el patrón RNY a todo lo largo de la secuencia (línea horizontal), extendiéndose, inclusive, mas allá del codón de término. Las flechas verticales, presentes en los tres marcos de lectura, señalan la posición de los codones de término.

CROMOSOMA III-. En este cromosoma se tienen registrados nueve ORFs cuestionables en los bancos de datos MIPS y Stanford, de los cuales tres, al cuantificar su porcentaje de seguimiento del formato RNY, fueron considerados como secuencias con potencial codificante: YCL 006c (109 codones), YCL 023c (115 codones) y YCR 041w (110 codones). El valor promedio de CAI para estas secuencias es de 0.1. De estos tres ORFs, consideradas como genes potenciales, YCL 006c e YCL 023c muestran una concordancia absoluta con el formato RNY, al no presentar desviaciones respecto a este codón de referencia.

En relación con los ORFs considerados como hipotéticos por MIPS y Stanford, se analizaron catorce secuencias, de las cuales cinco mostraron porcentajes de concordancia con el serial RNY en el 80% de su longitud; estas secuencias son: YCL 058c (152 codones), YCR 006c (157 codones), YCR 043c (127 codones), YCR 085w (117 codones), YCL 061c (329 codones) e YCR 017c (953 codones); las dos últimas secuencias, de considerable extensión, señalan hacia una alta probabilidad de codificación, ya que su longitud, aunada a un valor promedio de CAI de 0.16, refuerza la probabilidad de ser genes potenciales.

Es de señalar que ninguna de estas cinco secuencias mostraron similitud con alguna otra reportada en las bases de datos.

Respecto a los genes anotados, de diez secuencias seleccionadas, siete fueron cuantificados como codificantes; entre estas secuencias figura el gen ribosomal YCR 031c, el cual contiene un intrón que abarca desde el codón 3 hasta el codón 103, seguido por un exón terminal de 135 codones. Este exón es reconocido por el programa de Shepherd en fase 2, de manera lineal, hasta el codón de término (Figura 5). El primer exón abarca los dos primeros codones, presentándose en el marco de lectura 1. La gráfica ilustra de manera clara la capacidad de detección del algoritmo de Shepherd de regiones no codificantes (intrones) y codificantes (exones) así como sus transiciones de fase .

CROMOSOMA V-. Para este cromosoma, en el catálogo de MIPS aparecen señalados como ORFs cuestionables trece secuencias, de las cuales solamente figuran tres en la base de datos del SGD; como los criterios para considerar como ORFs a secuencias no anotadas varían entre ambas bases de datos, se consideró adecuado incluir solamente aquellos ORFs que estuvieran registrados en las dos bases de datos y que no mostraran discordancia respecto a la asignación de coordenadas; por esta razón, se recuperaron tres ORFs considerados como cuestionables, de los cuales solamente uno, YER 119c-A, generó una gráfica lineal aunque en fase 3, la cual está puntuada por codones de término, mientras que el

ORF está registrado en fase 1. La interpretación de la gráfica señala que este ORF no tiene potencial codificante, aunque por el corrimiento en la distribución del formato RNY, el programa registra la serie codificante en la fase 3.

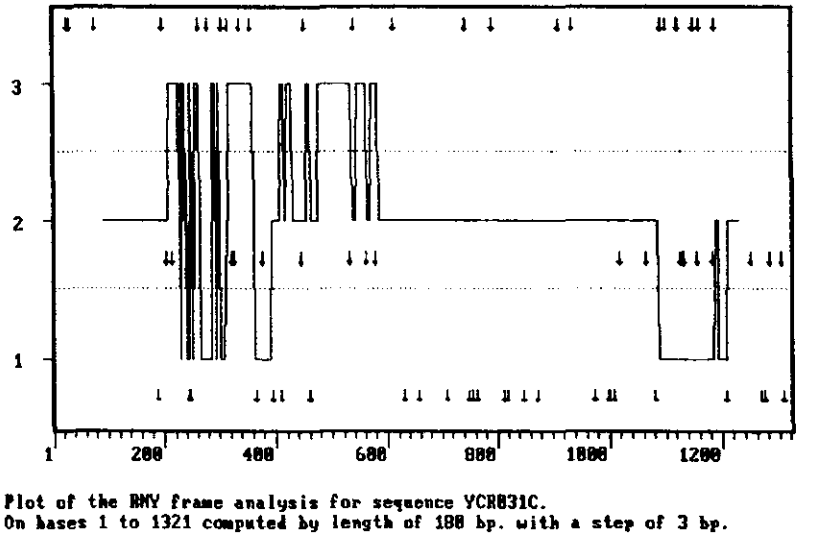


FIGURA 5.– En este gráfica de Shepherd del gen YCR031c, se muestra al intrón abarcando desde la base 307 hasta la base 612, observándose fluctuaciones del patrón RNY hacia los tres marcos de lectura; a partir de la base 612, se abre un ORF en fase 2, extendiéndose hasta la base 1018, en la que aparece un codón de término señalando el fin del exón. El formato RNY es reconocido por el programa y señalado como una línea horizontal en toda la extensión de la secuencia del exón en fase 2.

Respecto a los ORFs hipotéticos, de las veinte secuencias seleccionadas en este cromosoma, quince mostraron porcentajes del formato RNY mayores de 62% - 82% en promedio -, de las cuales diez corresponden a ORFs de alrededor de 150 codones y las otras cinco a ORFs de 500 codones promedio. Estas secuencias con potencial codificante fueron: YEL 008w (126 codones), YEL 010w (116 codones), YEL 014c (101 codones), YEL 028w (153 codones), YEL 068c (110 codones), YEL 075c (122 codones), YER 035w (145 codones), YER 067w (161 codones), YER 071c (126

codones), YER 121w (114 codones), YEL 017w (337 codones), YEL 023c (682 codones), YEL 057c (234 codones), YER 038c (464 codones) e YER 080w (627 codones). Estos ORFs muestran un valor de CAI promedio de 0.12.

Las gráficas de la mayoría de las secuencias analizadas mostraron algunas desviaciones respecto a la fase de lectura en la cual se encuentra el ORF principal, aunque en todos los casos el algoritmo de Shepherd señala el marco de lectura coincidente con el anotado por las bases de datos.

Respecto a los genes anotados y utilizados como control, de catorce genes seleccionados, el programa de Shepherd registró nueve como concordantes con el formato RNY. Entre estos genes figura YER 117w (137 codones), el cual codifica para una proteína ribosomal; como en el caso de YCR031c, este gen también contiene un intrón, el cual fue registrado por el algoritmo, identificando la fase de codificación del exón. Nuevamente, en este caso se demuestra el valor predictivo del programa en relación con las regiones con potencial codificante.

CROMOSOMA VII-. En este cromosoma fueron identificados treinta y siete ORF's considerados como cuestionables; de este total de secuencias, diez fueron cuantificadas como posibles secuencias codificantes, aunque cinco de ellas presentaron, en algunos casos, marcadas fluctuaciones respecto a la fase codificante putativa. Por otra parte, los cinco ORFs restantes generaron registros completamente lineales en toda su extensión en concordancia con el formato RNY en fase 1. Estos cinco ORFs fueron los siguientes: YGL 072c (119 codones), YGR 064w (122 codones), YGR 114c (129 codones), YGR 219w (113 codones) e YGR 242w (102 codones). Los valores de CAI para estas secuencias oscilaron entre 0.07 y 0.24, correspondiendo este último valor al ORF YGR 219w, lo cual lo señala como un ORF con una probabilidad de codificación.

Respecto a los cinco ORFs cuestionables que mostraron desviaciones respecto al serial RNY, su valor promedio de CAI fue de 0.10y su porcentaje de seguimiento del

serial RNY fue de 74% en promedio. Estos ORFs fueron los siguientes: YGL 007w (125 codones), YGL 034c (121 codones), YGR 045c (120 codones), YGR 164w (111 codones) e YGR 269w (108 codones).

En relación con los ORFs hipotéticos, se analizaron veintiún secuencias, de las cuales nueve ORFs mostraron un promedio de 85% de concordancia con el formato RNY, considerándose como secuencias potencialmente codificantes; los tamaños de estos ORFs varían entre 57 y 713 codones. Las secuencias que mostraron gráficas lineales correspondiendo con el serial RNY, es decir, concordantes, en toda su extensión, con la característica de codificación, fueron YGL 188c (57 codones) - este ORF es uno de los de menor longitud y con potencial codificante detectados en esta investigación (Figura 6)-, YGR 030c (158 codones) e YGR 294w (120 codones). Esta última secuencia mostró un valor de CAI de 0.65, con una alta probabilidad de ser un gen potencial. Otros dos ORFs, YGL 230c (147 codones) e YGR 035c (116 codones), también mostraron una pauta de lectura asociada a la fase 1, aunque presentaron fluctuaciones hacia la fase 3 a lo largo de sus secuencias. Los cuatro ORFs restantes de mayor tamaño que mostraron en sus gráficas de Shepherd mayor concordancia con el formato RNY fueron: YGL 138c (345 codones), YGL 183c (174 codones), YGR 103w (605 codones), YGR 113w (335 codones) e YGR 128c (703 codones); el valor medio de CAI fue de 0.16 con un porcentaje de seguimiento del serial RNY de 75%, en promedio.

Entre los genes anotados y analizados como controles, de los trece seleccionados, el algoritmo identificó a doce como secuencias codificantes. Dos de estos genes contienen intrones, YGL 103w (149 codones) (proteína ribosomal) e YGL 087c (137 codones) (ubiquitina). En ambos genes el programa registra las regiones de los exones en forma precisa.

CROMOSOMA IX-. De este cromosoma, se recuperaron de la base de datos tres ORFs clasificados como cuestionables, de los cuales dos fueron registrados por el programa y posteriormente cuantificados como secuencias con potencial

codificante: YIL 141w (129 codones) e YIL 163c (117 codones). El valor de CAI para ambas secuencias es de 0.11, presentando ambos ORFs desviaciones mínimas respecto al codón RNY. El tercer ORF recuperado, YIL 060w (144 codones), es reconocido por el algoritmo, con su pauta de lectura mostrando absoluta linearidad con la fase 3, mientras que el ORF registrado en la base de datos se despliega en la fase 1; este efecto puede ser debido a la persistencia del formato RNY en la fase 3 a pesar de la presencia de codones de término. Desde luego, los codones de terminación presentes en la fase 3 eliminan la posibilidad de que la secuencia codificante se encuentre en esta fase.

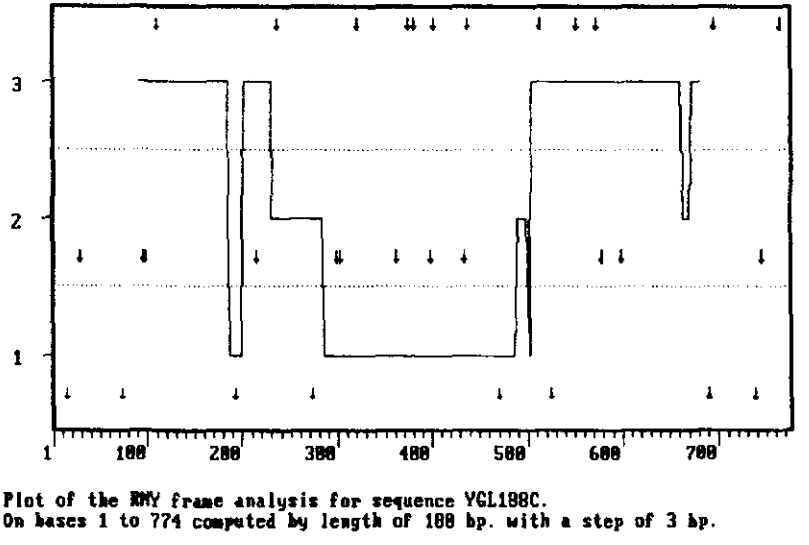


FIGURA 6.- YGL188c es uno de los ORFs hipotéticos más pequeños identificados en el genoma de la levadura. Este ORF se extiende desde la base 301 hasta la base 472 (57 codones) en el marco de lectura 1; el algoritmo de Shepherd identifica el formato RNY asociado a esta secuencia en toda su longitud. Se pueden apreciar las fluctuaciones o cambios de fase en el patrón RNY que flanquean la región del ORF.

De nueve ORFs hipotéticos seleccionados de este cromosoma, siete secuencias mostraron una fuerte correspondencia - 89% - entre el serial RNY y el registro generado por el programa. El valor de CAI para estas secuencias es, en promedio, de 0.13, con excepción de los ORFs YIL 176c(120 codones) y de YIR 041w (124 codones) los cuales mostraron valores de CAI de 0.65. Cabe señalar que la región que flanquea a las secuencias de -300 a +1 es idéntica en ambos ORFs (Figura 7); probablemente ambas secuencias son un ejemplo de duplicación génica, aunque en el análisis de las regiones duplicadas del genoma de la levadura realizado por Wolfe y Shields (1977),no señalan a ambas secuencias como secciones duplicadas. En el laboratorio comparamos estas secuencias mediante un alineamiento apareado, utilizando el programa NALIGN (Myers y Miller, 1988), mostrando ambas una similitud de 87.3% en la región del marco de lectura abierta y un 91.3% de similitud al adicionar la sección comprendida entre -600 y el sitio de inicio; estos resultados indican que ambos ORFs, residiendo en posiciones extremas del cromosoma IX y muy cercanas a las regiones teloméricas - YIL 176c en el brazo izquierdo y YIR041w en el brazo derecho - pueden representar un evento de duplicación génica no sólo de las regiones codificantes sino también de las secuencias precedentes, "upstream", probablemente de carácter regulatorio.

Las secuencias correspondientes a ORFs hipotéticos que mostraron porcentajes elevados de seguimiento del serial RNY -92% en promedio- y consideradas con potencial codificante fueron las siguientes: YIL 008w (99 codones), YIL 032c (118 codones), YIL 058w (94 codones), YIL 059c (121 codones), YIL 086c (102 codones), YIR 007w (764 codones) y YIR020c (100 codones).

CROMOSOMA XI-. En este cromosoma se seleccionaron diecinueve ORFs cuestionables, de los cuales únicamente cinco fueron considerados como secuencias con potencial de codificación; estas cinco secuencias muestran valores de CAI situados entre 0.09 y 0.20, lo cual, par a ORFs considerados como cuestionables, representan valores intermedios. Las secuencias fueron las siguientes: YKL 053w (124 codones), YKL118w (103 codones), YKL 153w (169 codones), YKL169c

(127 codones) y YKR 040c (167 codones). En el caso de YKL 153w, toda la secuencia registra una concordancia absoluta respecto al formato RNY, mientras que las otras dos secuencias presentan picos parásitos al inicio y al término de su longitud. El porcentaje de concordancia con el serial RNY en estos cinco ORFs es de 91%, lo cual los sitúa como probables genes putativos.

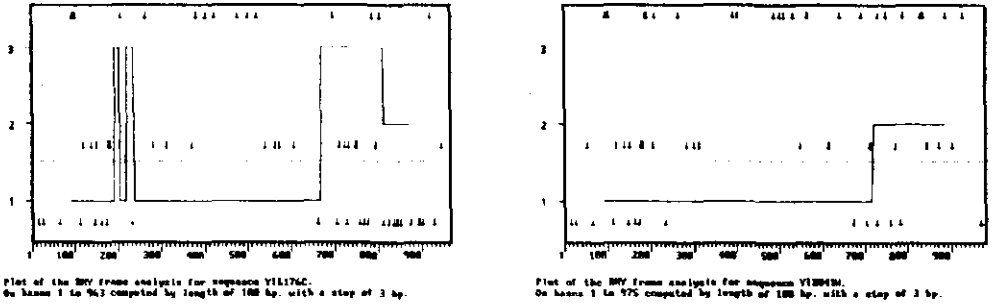


FIGURA 7.- Aquí se observa que las secuencias YIL176c y YIR041w, de similar longitud, comparten las mismas posiciones de sus codones de término en los 230 pares de bases previos al inicio de sus respectivos ORFs, ambos en el marco de lectura 1. El análisis de sus secuencias reveló un porcentaje de similitud de 87% en sus regiones codificantes putativas, las cuales se extienden desde la base 301 hasta la base 660, aproximadamente y un 91% de similitud en los 600 pares de bases previos al codón de inicio. Estos porcentajes sugieren que ambas secuencias son producto de un evento de duplicación.

Respecto a los ORFs considerados como proteínas hipotéticas, de veinticuatro secuencias seleccionadas , catorce fueron consideradas como potencialmente codificantes, debido al grado de concordancia con el serial RNY; entre los ORFs de mayor tamaño se encuentra uno de 1,764 codones, así como otros de tamaño intermedio, lo cual sugiere un origen no aleatorio para estas secuencias. Por otra parte, la mayoría de estos ORFs no muestran similitud con secuencias reportadas previamente. Las secuencias con probabilidad de codificación - 78% de seguimiento del formato RNY - fueron las siguientes: YKL 031w (137 codones) YKL 102c (101 codones), YKL 137w (103 codones), YKL 225w (115 codones), YKL 061w

(113 codones), YKL 014c (1,764 codones), YKL 183w (306 codones), YKL 189w (399 codones), YKL 221w (473 codones), YKL 222c (705 codones), YKR 044w (443 codones), YKL 107w (309 codones), YKL 151c (337 codones) e YKL 195w (427 codones).

Respecto a los genes anotados, en el 90 % de los casos el algoritmo registra concordancia entre el formato RNY y el marco de lectura abierta propio del gen; al igual que en otros cromosomas, en el caso de genes de proteínas ribosomales con intrones, el algoritmo asigna las pautas de lectura del serial RNY a las fases en las cuales están expresados los exones.

CROMOSOMA XIII-. En este cromosoma, uno de los de mayor tamaño, con 924,430 pares de bases, se tienen registradas veintidós secuencias consideradas como cuestionables; cinco de estas secuencias mostraron concordancia con el codón RNY en la fase 1. De estos cinco ORFs, cuatro mostraron ligeras desviaciones a lo largo de sus secuencias respecto al formato RNY; también se observa una alta correspondencia - 75% - entre este serial y el marco de lectura abierta registrado por el programa; la nomenclatura de estos ORFs es la siguiente: YML 089c (122 codones), YMR 031w-A (108 codones), YMR 244c-A (104 codones) e YMR 316c-A (103 codones); el quinto ORF considerado con potencial codificante, YMR 173w-A (394 codones), mostró una serie de características particulares, las cuales se detallan a continuación:

a) Su secuencia de bases muestra 2 marcos de lectura abierta, correspondientes a las fases 1 y 3 : en la fase 1, el ORF tiene una extensión de 1,182 pares de bases, correspondiendo a 394 codones -este ORF, con un valor de CAI de 0.09, corresponde al registrado en las bases de datos. El segundo ORF, en fase 3, tiene una extensión de 1,290 pares de bases (430 codones) y un valor de CAI de 0.35, mostrando una región de sobrelape de 1,081 pares de bases con el ORF anterior. El

programa de Shepherd visualiza la fase de lectura de esta secuencia en el marco 3, sin presentar desviaciones a otras fases (Figura 8).

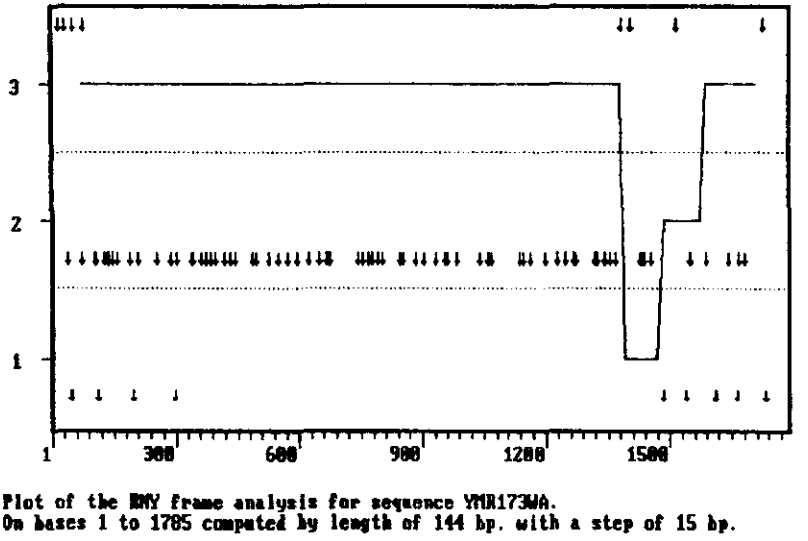


FIGURA 8.- Dos ORFs sobrelapados. En este gráfico de Shepherd se observan dos marcos de lectura abierta, situados en las fases 1 y 3, mostrando un sobrelape de 1380 pares de bases (460 codones); el ORF que aparece en fase 3 corresponde al gen DDR48, codificante para una proteína de choque térmico y el cual es identificado por el algoritmo de análisis. El ORF en fase 1 posiblemente corresponde a una secuencia codificante para una proteína de membrana; este ORF está considerado como cuestionable hasta el momento.

b) Las bases de datos registran un ORF contiguo a YMR 173w-A designándolo como YMR 173w y codificante para una proteína de 430 aminoácidos, anotada como proteína de choque térmico (DDR 48). Este es el ORF que visualiza el programa de Shepherd; sin embargo, el ORF consignado por las bases de datos, con una longitud de 394 codones (YMR173w-a)y no visualizado por el programa de Shepherd, muestra, al analizar su secuencia de aminoácidos, ocho dominios de

transmembrana, sugiriendo un posible papel de proteína integral. Esta proteína putativa tiene un peso molecular estimado de 43.9KD y un valor de CAI de 0.09.

c) Al realizar una búsqueda de similitud del ORF YMR173w-a de 394 pb con otras secuencias, utilizando el programa BLASTP 2.0.4., se encontró una similitud parcial de 59% con una proteína de 63.5 KD de la levadura de fisión *S. pombe*; esta proteína pertenece a la familia de traslocasas para resistencia a fármacos; asimismo, el programa BLASTP reportó una similitud parcial de 55% con otra proteína de transmembrana (hipotética) de *Haemophilus influenzae*, designada como proteína integral de membrana. Estos resultados apuntan a que la región YMR 173w contiene, probablemente, dos genes sobrelapados, uno de ellos anotado -DDR 48- y otro de carácter putativo – YMR173w-a con potencial codificante para una proteína de membrana. Este hecho de sobrelapamiento de genes en la levadura no ha sido reportado anteriormente, representando una vía de investigación prometedora respecto al estudio de la organización del genoma de la levadura.

En relación a los ORFs hipotéticos, se seleccionaron veinticuatro secuencias, de las cuales trece mostraron patrones compatibles con el formato RNY; entre estas secuencias, cuatro mostraron relaciones casi lineales entre la fase de lectura que cumple con el serial RNY y el marco de lectura abierta registrado para estas secuencias, por lo que fueron consideradas como ORFs potencialmente codificantes. Las secuencias que no mostraron relaciones lineales absolutas presentan picos parásitos intemedios o periféricos a lo largo del ORF, aunque la mayor parte de su extensión se adapta al formato RNY, con una concordancia promedio de 88%. Dichas secuencias son las siguientes: YML 108w (105 codones), YML 090w (128 codones), YMR 324c (80 codones), YMR 325w (124 codones), YMR 107w (115 codones), YMR 122c (124 codones), YMR 195w (127 codones), YMR 252c (134 codones), YMR 326c (102 codones), YML 071c (607 codones), YMR 114c (368 codones), YMR 130w (302 codones) e YMR 204c (420 codones).

Respecto a los genes anotados de control, en el 90% de los casos el programa identificó en forma correcta la fase de lectura; inclusive, en el caso del gen YML 085c, correspondiente al gen de la tubulina 1, el algoritmo registró en forma correcta el marco de lectura abierta de este gen, que corresponde, como caso particular, a la fase 3 del ORF.

CROMOSOMA XV-. De este cromosoma se recuperaron las treinta y cinco secuencias consideradas como ORFs cuestionables, identificándose ocho secuencias, las cuales mostraron una alta correspondencia - 83% - entre la fase de lectura y el serial RNY. Esta alta correspondencia nos llevó a conferirle a estas secuencias capacidad potencial codificante. Estas secuencias fueron : YOL 106w (117 codones), YOL 050c (106 codones), YOL 046c (224 codones), YOR 121c (101 codones), YOR 169c (154 codones), YOR 225w (109 codones), YOR 379c (112 codones) e YOR 218c (139 codones); este último ORF mostró una concordancia absoluta entre el formato RNY y la pauta de lectura en la fase 1.

Respecto a los ORFs considerados como hipotéticos, se seleccionaron treinta secuencias, de las cuales diecisiete fueron cuantificadas y consideradas como potencialmente codificantes debido a la elevada concordancia - 84% - entre la serie RNY y la fase de lectura abierta, la cual corresponde, en todos los casos, a la fase 1. Las secuencias de entre 100 y 150 codones, con potencial codificante, fueron las siguientes: YOL 166c (112 codones), YOL 160w (113 codones), YOL 118c (102 codones), YOL 085c (113 codones), YOR 252w (141 codones), YOL 026c (113 codones), YOR 015w (119 codones). Estas dos últimas secuencias mostraron una alta correspondencia entre el codón de referencia RNY y el marco de lectura abierta. Como caso especial se puede mencionar al ORF YOR 053w - codificante para una acetiltransferasa - en cuya secuencia se presentan 3 ORFs simultáneos, ya que no se observan codones de término al interior de los 3 marcos de lectura abierta; en este caso, el algoritmo de Shepherd identifica el formato RNY en forma absoluta en fase 2, mientras que las bases de datos registran el ORF principal en la fase 1. Por otra parte, el tercer ORF se presenta en la fase 3 con una extensión casi

idéntica a la secuencia de la fase 2 (Figura 9). Es posible que en este caso también se esté presentando un sobrelapamiento de las secuencias codificantes, pues la ausencia de codones de término abre la posibilidad de la existencia de genes sobrelapados, aún sin caracterizar: este gen muestra un valor de CAI de 0.10, desconociéndose los valores para las otras 2 fases. Por otra parte, las secuencias consideradas como potencialmente codificantes -con una ocurrencia del formato RNY en el 75 % de la longitud de los ORFs, en promedio, así como una extensión de entre 300 y 1200 codones- fueron las siguientes: YOL 091w (609 codones), YOL 078w (1,176 codones), YOL 070c (501 codones), YOL 063c (957 codones), YOR 129c (893 codones), YOR 175c (619 codones),YOR 292c (309 codones), YOR 342c (319 codones) e YOR 352w (343 codones).

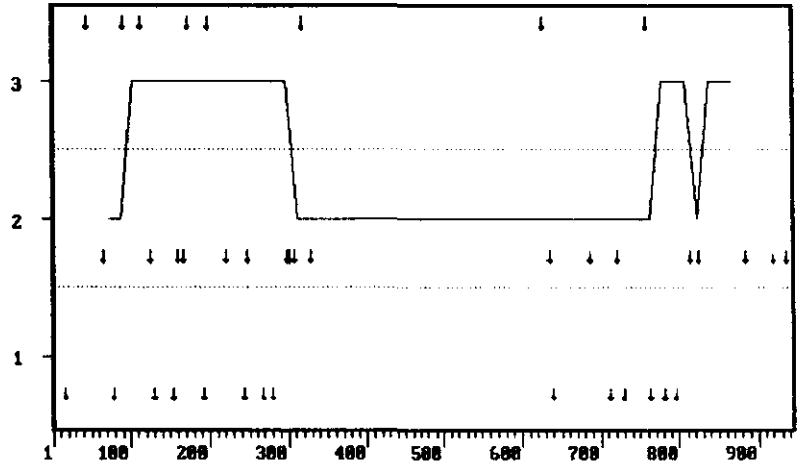


FIGURA 9.- Tres ORFs sobrelapados. Aquí se pueden apreciar tres regiones de marco de lectura abierta sobrelapadas y compartiendo una sección de 310 pares de bases (aproximadamente 103 codones); el programa de Shepherd registra el formato RNY en la fase 2, mientras que el ORF en fase 1 corresponde al gen de la acetiltransferasa y es el que consignan las bases de datos. Se desconoce si alguno de los ORFs en las fases 2 y 3 es codificante.

Respecto a genes anotados utilizados como control del programa, el 90% de las secuencias seleccionadas en este cromosoma fueron reconocidas por el algoritmo en correspondencia con el formato RNY. Esta calibración del programa informático es la base para asignarle una alta confiabilidad en su capacidad predictora respecto al reconocimiento de regiones genómicas con potencial codificante.

5-. DISCUSION Y CONCLUSIONES.

Al integrar los resultados del análisis llevado a cabo en ocho cromosomas de *Saccharomyces cerevisiae*, llama la atención la diferencia entre el porcentaje encontrado de ORFs cuestionables pero con potencial codificante, equivalente a 27.7%, respecto a los ORFs hipotéticos que también mostraron este potencial, 62% (Figura 10); aunque estos resultados son concordantes con las categorías bajo las cuales están clasificados estos ORFs en las bases de datos, esto es, como cuestionables y como hipotéticos, esta diferencia señala una característica del genoma de la levadura: la mayoría de los genes potenciales analizados pertenecen a la fracción de ORFs con marcos de lectura abierta de una longitud mayor de 100 codones, sin que los valores bajos de CAI representen un criterio definitivo para discriminar entre regiones con ó sin potencial de codificación; la longitud promedio de los ORFs cuestionables potencialmente codificantes analizados en este estudio fue de 135 codones, mientras que la de los ORFs hipotéticos con potencial codificante fue de 286 codones; esta diferencia refleja la gran variabilidad en la longitud de los ORFs en los que se manifiesta el formato RNY. Ahora bien, el análisis del transcriptoma de la levadura ha indicado la presencia de al menos 160 genes no caracterizados ni anotados como ORFs en las bases de datos, pero cuyos niveles de expresión son detectables (Velculescu et al,1997); una característica común en estos ORFs es su tamaño reducido, del orden de 66 codones en promedio, lo cual señala la necesidad de prestar más atención a secuencias de longitud reducida. Por otra parte, estos 160 ORFs apenas representan el 3.4 % del total de genes expresados detectados y tan sólo el 2.6 % del total de ORFs anotados en las bases de datos; una limitación importante respecto al conocimiento de estas secuencias es que aún no están disponibles sus coordenadas, impidiendo, por ahora, su análisis por medios computacionales. Por otra parte, los dos tipos de ORFs analizados en el presente estudio y que mostraron potencial codificante, no presentaron diferencias apreciables en sus valores de CAI

: 0.12 para los ORFs cuestionables y 0.14 para los ORFs considerados como hipotéticos. Esta falta de correlación entre el tamaño de un ORF y su valor de CAI fue observada por Dujon et al (1994) al analizar la secuencia del cromosoma XI, aunque en el mismo estudio, los autores señalan que la búsqueda de ORFs como genes funcionales se ha basado, generalmente, en la utilización del CAI como criterio de discriminación, junto con su tamaño; la presente investigación sugiere que, para ORFs de extensión moderada -menores de 300 codones- el valor de CAI no es un elemento crítico para la definición de un gene potencial, mientras que para ORFs de mayor longitud, su misma extensión minimiza su probabilidad de ocurrencia al azar a valores por debajo del 0.2 % (Sharp y Cowe, 1991; Oliver et al, 1992).

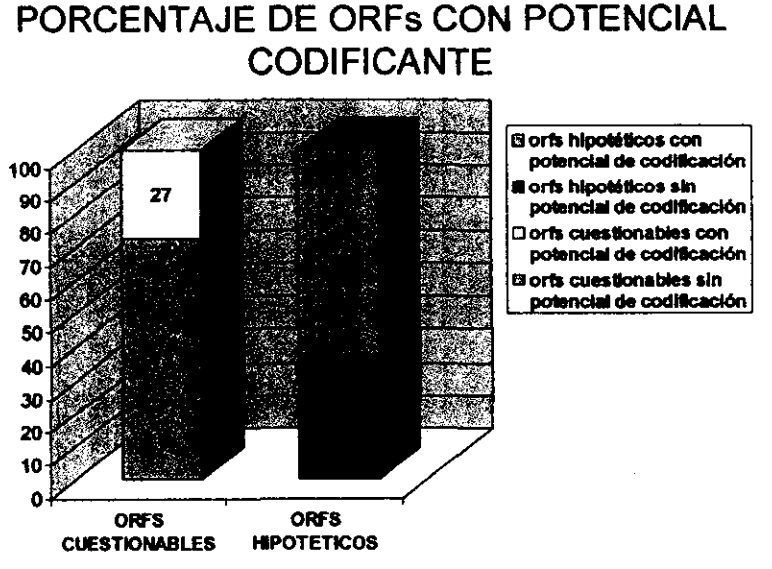


FIGURA 10.- La gráfica señala las diferencias en el potencial de codificación en ORFs cuestionables e hipotéticos; los ORFs cuestionables mostraron un 27% de potencial codificante mientras que en los ORFs hipotéticos este porcentaje es de 62%; estos resultados concuerdan con las estimaciones de los autores del Proyecto de Secuenciación del Genoma de *Saccharomyces cerevisiae* para estas secuencias.

Otro resultado importante que se observó respecto al tamaño de los ORFs con potencial codificante consistió en que las ocho secuencias menores a 100 codones que se identificaron, corresponden a la categoría de ORFs hipotéticos y no a ORFs cuestionables, mostrando un valor promedio de CAI de 0.14, propio de ORFs hipotéticos. Entre estas secuencias, la de menor tamaño que fue reconocida por el programa de Shepherd con potencial de codificación se encontró en el cromosoma VII, con una longitud de 57 codones y sin similitud con otra secuencia reportada; su valor de CAI se ubicó en 0.20. Por otra parte, una secuencia con extensión menor a 100 codones fue reconocida en el cromosoma I, constando de 99 codones, sin similitud conocida y con un valor de CAI de 0.09; de esta manera, el escaso número de secuencias con potencial codificante de una longitud menor de 100 codones, así como el bajo número total de ORFs pequeños -291 secuencias menores de 100 codones con ó sin función conocida -, presentes en el genoma de *Saccharomyces cerevisiae*, refuerza la idea de que la búsqueda computacional de ORFs con capacidad codificante (entre las más de 2,800 secuencias sin caracterización genética), debe poner especial énfasis en la respuesta que generen las secuencias a los algoritmos de búsqueda por contenido, mientras que el criterio del valor de CAI -de indudable utilidad-, puede ser aplicado para estimar el nivel de expresión de los ORFs identificados como potencialmente codificantes. Estas dos condiciones, junto con los resultados obtenidos de la caracterización del transcriptoma de la levadura, pueden servir para definir prioridades en los programas experimentales de investigación genética, bioquímica y biológica de las regiones codificantes del genoma de *Saccharomyces cerevisiae*.

Respecto al algoritmo de búsqueda por contenido utilizado en esta investigación, el programa de Shepherd demostró una gran capacidad de discriminación de las regiones con potencial codificante respecto a las no codificantes en las dos clases de secuencias analizadas correspondientes a genes anotados y ORFs hipotéticos y cuestionables (Figura 11), así como con genes conteniendo intrones, discriminando,

en estos últimos, las secuencias correspondientes a los exones, y señalando, con gran precisión, los límites exón -intrón- exón.

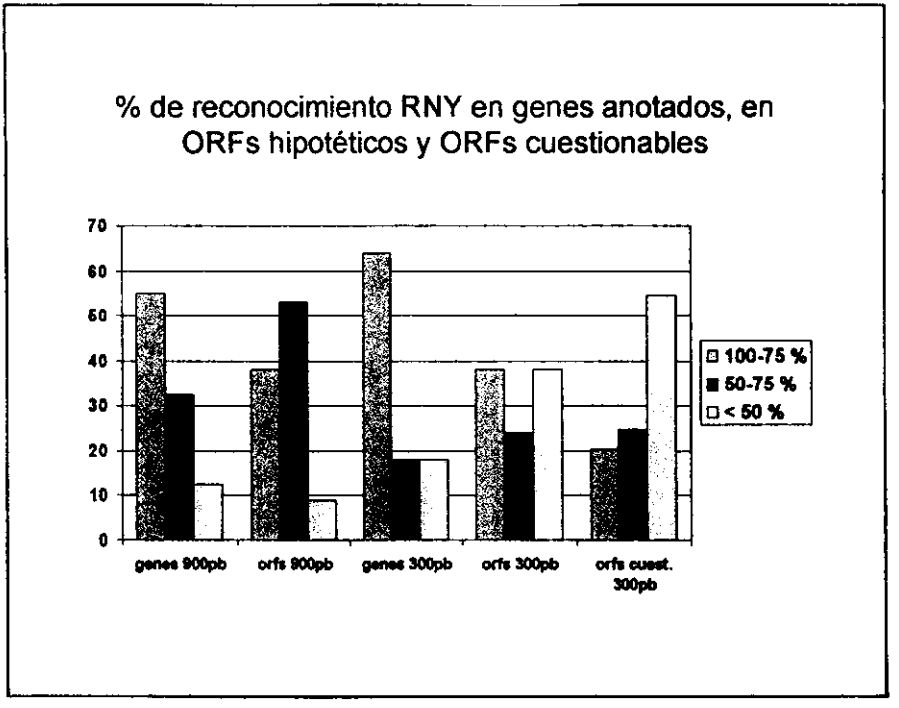


FIGURA 11.- La gráfica muestra los porcentajes de reconocimiento del formato RNY en genes y ORFs de diferente longitud. Los rangos a la derecha de la gráfica indican, en términos porcentuales, la extensión de la secuencia en la que está presente el formato RNY. Los mayores porcentajes de presencia del serial RNY se encontraron en genes de 900 y 300 pares de bases, mientras que los menores porcentajes se observaron en ORFs considerados como cuestionables.

El programa informático se aplicó escogiendo los parámetros que ofrecieran la mayor sensibilidad de análisis, aunque, en algunos casos, se amplificaron las desviaciones respecto al serial RNY, ocasionando "ruido" (presencia de picos parásitos en la gráfica), sin que este efecto obstruyera la interpretación de los datos. Tal como los resultados señalan, el programa generó dos tipos principales de gráficas: las de tipo lineal, en la cual el ORF muestra una correspondencia absoluta con el formato RNY, y la de tipo puntuado, en la que aparecen desviaciones, en la

secuencia del ORF, respecto a este mismo formato. En términos generales, las configuraciones lineales se observaron en ORFs de una extensión de 100 a 150 codones, mientras que lo registros gráficos con desviaciones del serial RNY se presentaron en aquellas secuencias con longitudes comprendidas entre 200 y 700 codones; desde luego, pueden presentarse casos de ORFs pequeños con desviaciones del serial RNY así como secuencias de longitud considerable -mayores de 300 codones- que muestren comportamientos lineales en concordancia con este codón de referencia (Shepherd, 1990; Staden, 1984,1990).

En relación con la probable presencia de genes sobrelapados en el genoma de la levadura, el caso observado en el cromosoma XIII en la región YMR 173 apunta a las siguientes consideraciones: la presencia de dos marcos abiertos sobrepuestos en la misma secuencia es, evidentemente, la condición necesaria, aunque no suficiente, para la existencia de genes sobrelapados; se conoce que la secuencia correspondiente al gen anotado, DDR 48, tiene una longitud de 430 codones presentes en la fase de lectura 3, mientras que el segundo ORF muestra una longitud de 1,182 pares de bases -394 codones- presentes en fase 1, generando una región de sobrelapamiento de 1,081 pares de bases -360 codones-; una característica importante de estos 2 ORFs es que ninguno de ellos está contenido en el otro, y que ambos se presentan con la misma polaridad 5'→3'. El gen DDR 48 se extiende del nucleótido 93 al 1,382, mientras que el ORF considerado cuestionable se extiende del nucleótido 301 al 1,482, es decir, el gen DDR 48 principia aproximadamente 200 pares de bases antes que el ORF cuestionable y termina 100 pares de bases antes; es importante señalar que el valor de CAI asignado para DDR 48 es de 0.35 mientras que este valor para el ORF YMR 173w-A es de 0.09. El hecho de que el ORF considerado como cuestionable tenga una extensión de 394 codones hace muy improbable su presencia por procesos aleatorios aunque no se puede descartar esta posibilidad. Respecto a la búsqueda de similitud de la secuencia del ORF cuestionable con alguna otra reportada, se encontró correspondencia con dos proteínas de membrana, lo cual,

aunado al análisis de su composición de aminoácidos, señala que es muy verosímil que este ORF corresponda a un gen codificante para un polipéptido presente en membrana. La predicción sobre la capacidad codificante de este ORF, YMR173w-a, puede ser verificada experimentalmente mediante la técnica del análisis serial de la expresión del gene, SAGE, desconociéndose hasta el momento, si el análisis del proteoma de *Saccharomyces* ha identificado esta región respecto a sus productos de expresión génica. Por otra parte, la observación de secuencias pertenecientes a ORFs cuestionables y en las cuales se presentan ORFs sobrelapados entre las fases 1 y 3, indica que la compactación de genes en el genoma de la levadura puede favorecer el sobrelape de regiones potencialmente codificantes; como en el caso del cromosoma XIII, en el XV las secuencias sobrelapadas se presentan en las fases 1 y 3, aunque el programa de Shepherd registra únicamente una fase como codificante, la cual muestra, en forma lineal, una total concordancia con el patrón RNY. De esta manera, resulta de sumo interés el caracterizar, de forma experimental, la probable capacidad de codificación de los ORFs que muestran sobrelapamiento en sus secuencias, independientemente de que estén considerados como ORFs cuestionables, condición ésta que no representa una designación excluyente respecto al potencial de codificación de estas secuencias.

En resumen, el panorama resultante de esta investigación puede ser visualizado de la siguiente manera:

1.- Mediante este estudio se descubrió que el 27% de los ORFs cuestionables y el 62% de los ORFs hipotéticos analizados mostraron potencial codificante, lo cual sugiere un esquema de prioridades en el análisis funcional de las secuencias que pudieran corresponder a nuevos genes.

2-. Los ORFs pequeños considerados como cuestionables pueden ser estudiados por algoritmos de búsqueda por contenido, independientemente de sus valores de CAI, los cuales, pueden vincularse con los niveles de expresión y no tomarse como

criterios restrictivos respecto a su capacidad codificante; esto es especialmente señalado para ORFs que muestran valores de CAI ubicados entre 0.05 y 0.12.

3.- La distribución de ORFs cuestionables e hipotéticos en los ocho cromosomas de la levadura estudiados en esta investigación, está en función de la longitud y de la densidad génica en los cromosomas, sin que se observe un sesgo respecto a la frecuencia de aparición de esta clase de ORFs en algún cromosoma en particular; la anomalía respecto a la carencia de algunos ORFs cuestionables en la base de datos de Stanford puede deberse a problemas en la asignación de las coordenadas correspondientes, lo cual revela las discordancias en la anotación de secuencias entre las dos principales bases de datos que manejan la secuencia completa del genoma de *Saccharomyces cerevisiae*.

4-. El descubrimiento de una región de sobrelaparniento en las secuencias de ORFs en el cromosoma XIII y otras regiones en el cromosoma XV, apunta hacia un análisis en profundidad de estas secuencias, en dos direcciones complementarias: por una parte, la búsqueda de similitudes con otras secuencias reportadas, y por otra, su correlación con los resultados derivados del análisis del transcriptoma, el cual a la fecha, está siendo publicado en forma parcial por el grupo de investigación de la Escuela de Medicina de la Universidad Johns Hopkins en Baltimore (Velculescu et al, 1997).

5-. El algoritmo de Shepherd demostró su capacidad de reconocimiento de secuencias en relación con su potencial de codificación, registrando con gran precisión secuencias que por sus características de cambios de fase, presentan dificultades para su detección, incluso utilizando algoritmos de búsqueda por señal - como es el caso de genes conteniendo intrones- (Claverie, et al, 1990). Este programa informático presenta una gran versatilidad en sus aplicaciones, pues al basar su búsqueda de regiones codificantes en torno al formato RNY, prescinde de criterios más limitantes tales como la utilización preferencial de codones - particularmente en genes putativos con bajo nivel de expresión -, la ubicación de

codones de inicio o de término o la presencia de secuencias consenso en el ORF analizado y en regiones contiguas. Es, por lo tanto, un algoritmo que debería ser más utilizado en los programas de búsqueda por contenido.

CONCLUSIONES FINALES

1) Se logró la identificación diferencial de ORFs pequeños con potencial codificante y sin potencial codificante dentro de las categorías de ORFs cuestionables e hipotéticos, los cuales no habían sido previamente definidos en el genoma de la levadura.

2) Los resultados señalan la factibilidad de un análisis experimental detallado en los ORFs con potencial codificante, señalando prioridades y posibilitando la identificación funcional de nuevos genes.

3) La aplicación del algoritmo de Shepherd -basado en la identificación del formato RNY en ORFs aún no caracterizados en el genoma de la levadura- es una herramienta valiosa en el reconocimiento de regiones potencialmente codificantes; esto es especialmente apropiado en ORFs considerados como cuestionables por criterios tales como su limitada extensión y por presentar bajos valores de CAI.

REFERENCIAS.

1. **Andrade, M., A. Daruvart, G. Casari, R. Schneider, M. Termier and C. Sander.** (1997). Characterization of new proteins found by analysis of short open reading frames from the full yeast genome. Yeast, Vol. 13: 1363-1374.

2. **Barry, C., G. Fichant, A. Kalogeropoulos and Y. Quentin.** (1996). A computer filtering method to drive out tiny gene from the yeast genome. Yeast, Vol. 12: 1163-1178.

3. **Baudin, A., O. Ozier-Kalogeropoulos, A. Denouel, F. Lacroute and C. Cullin.** (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. Nucleic Acids Research, Vol. 21: 3329-3330.

4. **Beggs, J.** (1978). *Transformation of yeast by a replicating hybrid plasmid*. Nature, Vol. 275: 104-109.

5. **Boucherie, H., F. Sagliocco, R. Joubert, 1. Maillet, J. Labarre and M. Perrot.** (1996). Two-dimensional gel protein database of *Saccharomyces cerevisiae*. Electrophoresis, Vol. 17: 1683-1699.

6. **Carbon, J.** (1993). Genes, Replicators, and Centromeres: The First Artificial Chromosomes. En: M. Hall and P. Linder (eds) The Early Days of Yeast Genetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. pp. 375-390.

7. **Carle, G., and M. Olson.** (1985). Separation of Chromosomal DNA Molecules from Yeast by Orthogonal-Field-Alternation Gel Electrophoresis. Nucleic Acids Research, Vol. 12: 5647-5664.

8. **Chinault, A. and J. Carbon.** (1979). Overlap hybridization screening: Isolation and characterization of overlapping DNA fragments surrounding the leu2 gene on yeast chromosome III. Gene, Vol. 5: 111-126.

9. **Clarke, L. and J. Carbon.** (1980). Isolation of a yeast centromere and construction of functional small circular chromosomes. Nature, Vol. 287: 504-509.

10. **Claverie, J., I. Sauvaget and L. Bougueleret.** (1990). k-Tuple Frequence Analysis: From Intron/Exon Discrimination to T-Cell Epitope Mapping. En: R. Doolittle (ed) Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzymology, Vol. 183 Academic Press, San Diego. pp. 237-252.

11. **Clayton, R., O. White, K. Ketchum and J. C. Venter.** (1997). The first genome from the third domain of life. Nature, Vol. 387: 459-462.

12. **Das, S., L.Yu, C. Gaffatzes, R. Rogers, J. Freeman, J. Bienkowska, R. Adams and T. Smith.** (1997). Biology's new Rosetta stone. Nature, Vol. 235: 29-30.

13. **Doolittle, R.** (1987). OF URFS AND ORFS: A Primer on How to Analyze Derived Amino Acid Sequences. *University Science Books, Mill Valley, California.* 102 p.

14. **Doolittle, R.** (1990). Searching through Sequence Databases. En: R. Doolittle (ed) Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzimology. Vol. 183, Academic Press, San Diego. pp. 99-110.

15. **Dujon, B. D. Alexandraki, B. Andre, W. Ansorge, V. Baladron, J. Ballesta, A. Banrevi et al.** (1994). Complete DNA sequence of yeast chromosome XI. Nature, Vol. 369: 371-378.

16. **Fey, S., A. Nawrocki, M. Larsen, A. Gorg, P. Roepstorff, G. Skews, R. Williams and P. Larsen.** (1997). Proteome analysis of *Saccharomyces cerevisiae*: a methodological outline. Electrophoresis, Vol. 18: 1361-1372.

17. **Fickett, J. and C. Tung.** (1992). Assessment of protein coding measures. Nucleic Acids Research, Vol. 20: 6441-6450.

18. **Frommont-Racine, M. J. Rain and P. Legrain.** (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat. Gent., Vol. 16: 277-282.

19. **Garrels, J., C. McLaughlin, J. Warner, B. Futcher, G. Latter et al.** (1997). Proteome studies of *Saccharomyces cerevisiae*: Identification and characterization of abundant proteins. Electrophoresis, Vol. 18: 1347-1360.

20. **Gelfand, M.** (1990). Global methods for the computer prediction of protein-coding regions in nucleotide sequence. Biotechnology Software, July-Aug.: 3-11.

21. **Goffeau, A. B. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert et al.** (1996). Life whit 6000 Genes. Science, Vol. 274: 546-567.

22. **González-Pastor, J., J. San Millán and F. Moreno.** (1994). The smallest known gene. Nature, Vol. 369: 281.

23. **Gribskov, M., J. Devereux, and R. Burgess.** (1984). The codon preference plot: graphic analysis of protein coding sequence and prediction of gene expression. Nucleic Acids Research, Vol. 12: 539-549.

24. **Güldener, U., S. Heck, T. Fiedler, J. Beinhauer and J. Hegemann.** (1996). A new efficient gene disruption cassette for repeated use in budding yeast. Nucleic Acids Research, Vol. 24: 2519-2524.

25. **Hsiao, C. and J. Carbon.** (1979). High frecuency transformation of yeast by plasmids containing the cloned yeast ARG4 gene. Proc. Natl. Acad. Sci., Vol. 76: 3829-3833.

26. **Ikemura, T.** (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol., Vol. 2: 13-34.

27. **Jukes, T.** (1996). On the Prevalence of Certain Codons ("RNY") in Genes for Proteins. J. Mol. Evol., Vol. 42: 377-381.

28. **Koonin, E., A. Mushegian and K. Rudd.** (1996). Sequencing and analysis of bacterial genomes. Current Biology, Vol. 6: 404-416.

29. **Lashkari, D., J. De Risi, J. McCusker, A. Namath, C. Gentile, S. Hwang, P. Brown and R. Davis.** (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci., Vol. 94: 13057-13062.

30. **Li, H. and L. Luo.** (1996). The Relation between Codon Usage, Base Correlation and Gene Expression Level in Escherichia coli and Yeast. J. Theor. Biol., Vol. 181: 111-124.

31. **MIPS.** (1988). http://speedy.mips.biochem.mpg.de/mips/yeast-genome.htmlx.

32. **Mortimer, R.** (1993). Ojvind Winge: Founder of Yeast Genetics. En: M. Hall and P. Linder (eds) The Early Days of Yeast Genetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

33. **Murray, A. and J. Szostak.** (1983). Construction of artificial chromosomes in yeast. Nature, Vol. 305: 189-193.

34. **Nowak, R.** (1995). Entering the Postgenome Era. Science, Vol. 270: 368 -371.

35. **Olivas, W., D. Muhrad and R. Parker.** (1997). Analysis of the yeast: Identification of new non-coding and small ORF-containing RNAs. Nucleic Acids Research, Vol. 25: 4619-4625.

36. **Oliver, S. Q. van der Aart, M. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki et al.** (1992). The complete DNA sequence of yeast chromosome III. Nature, Vol. 357: 38-46.

37. **Oliver, S.** (1996). From DNA sequence to biological function. Nature, Vol. 379: 597-600.

38. **Oliver, S.** (1997). From gene to screen with yeast. Curr. Opin. Genet. Dev., Vol 7: 405-409.

39. **Olson, M.** (1991). The Genome of *Saccharomyces cerevisiae*. En: Broach, J., J. Pringle and E. Jones (eds). The Molecular and Cellular Biology of the Yeast *Saccharomyces cerevisiae*. Cold Spring Harbor Laboratory, New York. pp. 1-39.

40. **Oshima, Y.** (1993). Homothallism, Mating-type Switching, and the Controlling Element Model in *Saccharomyces cerevisiae*. En: M. Hall and P. Linder (eds) The Early Days of Yeast Genetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. pp. 291-304.

41. **Petes, T. and D. Botstein.** (1977). Simple Mendelian inheritance of the Reiterated ribosomal DNA of yeast. Proc. Natl. Acad. Sci., Vol. 74: 5091-5095.

42. **Ratzkin, B. and J. Carbon.** (1977). Functional expression of cloned yeast DNA in *E. coli*. Proc. Natl. Acad. Sci., Vol. 74: 487-491.

43. **Roman, H.** (1986). The Early Days of Yeast Genetics: A personal narrative. Annual Review of Genetics, Vol. 20 : 1-12.

44. **Rodriguez-Medina, J. and B. Rymond.** (1994). Prevalence and distribution of introns in non-ribosomal proteins genes of yeast. Mol. Gen. Genet., Vol. 243: 532-539.

45. **Rothstein, R.** (1983). One-step Gene Disruption in Yeast. Methods in Enzymology, Vol. 101: 202-211.

46. **Rudd, K., 1. Humphery-Smith, V. Wasinger and A. Bairoch.** (1998). Low molecular weight proteins: a challenge for post-genomic research. Electrophoresis, Vol. 19: 536-544.

47. **SGD.** (1998). http//genome-www.stanford.edu/*Saccharomyces*.

48. **Sharp, E. and W. Li.** (1987). The codon adaptation index-: a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research, Vol. 15: 1281-1295.

49. **Shepherd, J.** (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci., Vol. 78: 1596-1600.

50. **Shepherd, J.** (1990). Ancient Patterns in Nucleic Acid Sequences. En: R. Doolittle (ed) Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzymology, Vol. 183, Academic Press, San Diego. pp. 180-192.

51. **Shevchenko, A., O. Jensen, A. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H Boucherie and M. Mann.** (1996). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc. Natl. Acad. Sci., Vol. 93: 14440-14445.

52. **Shoemaker, D., D. Lashkari, D. Morris, M. Mittmann and R. Davis.** (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nature Genetics, Vol. 14: 450-456.

53. **Snyder, E. and G. Stormo.** (1993). Identification of coding regions in genomic DNA sequence: an application of dynamic programming and neural networks. Nucleic Acids Research, Vol. 21: 607-613.

54. **Snyder, E. and G. Stormo.**(1995). Identification of Coding Regions in Genomic DNA. J. Mol. Biol., Vol. 248: 1-18.

55. **Staden, R.** (1984). Measurement of the effects that coding for a protein has on a DNA sequence and their use for finding genes. Nucleic Acids Research, Vol. 12: 551-567.

56. **Staden, R.** (1990). Searching for Patterns in Protein and Nucleic Acid Sequences. En: R. Doolittle (ed) Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzymology, Vol. 183, Academic Press, San Diego. pp. 193-211.

57. **Staden, R.** (1990). Finding Protein Coding Regions in Genomic Sequences. En: R. Doolittle (ed) Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzymology. Vol. 183, Academic Press, San Diego. pp. 163-180.

58. **Stormo, G.** (1987). Identifying coding sequences. En: M.J. Bishop and C. J. Rawlings (eds) Nucleic acid and protein sequence analysis -A practical approach. IRL Press, Oxford-Washington D. C., pp. 231-258.

59. **Stormo, G.** (1990). Consensus Patterns in DNA. En: R. Doolittle (ed) Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzymology, Vol. 183, Academic Press, San Diego. pp. 211-221.

60. **Struhl, K.** (1983). The New Yeast Genetics. Nature, Vol. 305: 391.

61. **Valadez, J., C. Ontiveros, J. Hernández, J. Vega, B. Aguilar, M. A. Magos y G. Guarneros.** (1997). MINIGENES TOXICOS: Elementos Novedosos en la Regulación de la Síntesis de Proteínas. Boletín de la Sociedad Mexicana de Bioquímica, Vol. 8, 2: 6-15.

62. **Velculescu, V. L. Zhang, B. Vogelstein and K. Kinzler.** (1995). Serial Analysis of Gene Expression. Science, Vol. 270: 484-487.

63. **Velculescu, V. L. Zhang, W. Zhou, J. Vogelstein, M. Balsrai, D. Bassett, P. Hieter, B. Vogelstein and K. Kinzler.** (1997). Characterization of the Yeast Transcriptome. Cell, Vol. 88: 243-251.

64. **Wach, A., A. Brachat, R. Pohlmann and P. Philippsen.** (1994). New heterologous modules for classical or PCR-based gene disruptions in *Saccharomyces c erevisiae*. Yeast, Vol. 10: 1793-1808.

65. **Watson, J., N. Hopkins, J. Roberts, J. Steitz and A. Weiner.** (1987). Molecular Biology of the Gene, 4th ed. Benjamin/Cummings, Menlo Park, California.

66. **Wolfe, K. and D. Shields.** (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature, Vol. 387: 708-713.

APENDICE

LISTADO DE ORFs ANALIZADOS EN EL GENOMA DE *Saccharomyces cerevisiae*				
GENES ANOTADOS		ORFs HIPOTETICOS		ORFs CUESTIONABLES
≈ 900 pb	≈ 300 pb	≈ 900 pb	≈ 300 pb	≈ 300 pb
YAL005C	YAL030W	YAL027W	YAL030C	YCL006C
YAL020C	YAR029W	YAL034WA	YAL045C	YCL023C
YAL039C	YCL007C	YAL056W	YAL046C	YCL041C
YCL029C	YCL022C	YCL061C	YAL046C	YCL042W
YCR005C	YCL066W	YCR017C	YAL064W	YCL046W
YCR021C	YCR014C	YEL017W	YAL066W	YCR041W
YEL030W	YCR020CA	YEL023C	YAR020C	YCR049C
YEL063C	YCR024CA	YEL057C	YAR040C	YCR064C
YER011W	YCR096C	YER038C	YAR047C	YCR087W
YGL073W	YEL003W	YER080W	YAR064W	YER084W
YGL175C	YEL030W	YGL075C	YAR069C	YER119CA
YGL212W	YEL039C	YGL138C	YAR070C	YER181C
YGL225W	YEL049W	YGL183C	YCL002C	YGL007W
YIL035C	YEL054C	YGL196W	YCL016C	YGL024W
YIL134W	YEL063C	YGL219C	YCL056C	YGL034C
YIL142W	YER011W	YGR002C	YCL058C	YGL039C
YIR008C	YER058W	YGR103W	YCR001W	YGL072C
YIR034C	YER109CA	YGR113W	YCR006C	YGL074C
YKL011	YER117W	YGR128C	YCR022C	YGL088W
YKL019W	YER131W	YIL101C	YCR025C	YGL109W
YKL073W	YER159C	YIR007W	YCR043C	YGR011W
YKL085W	YGL058W	YKL014C	YCR068W	YGL118C
YKL094W	YGL070C	YKL107W	YCR085W	YGR018C
YKL201C	YGL087C	YKL108W	YEL008W	YGR039W
YKR048C	YGL089C	YKL151C	YEL010W	YGR045C
YKR052C	YGL103W	YKL183W	YEL014C	YGR050C
YML032C	YGR008C	YKL189W	YEL028W	YGR051C
YML085C	YGR020C	YKL195W	YEL067W	YGR052W
YML110C	YGR037C	YKL206C	YEL068C	YGR064W
YOL006C	YGR063C	YKL221W	YEL075C	YGR069W
YOR006C	YIL062C	YKL222C	YER035W	YGR104W
YOR027W	YIL069C	YKR044W	YER044C	YGR115C
YOR061W	YIL171W	YML071C	YER046W	YGR137W
YOR136W	YIR009W	YML114C	YER071C	YGR139W
YOR143C	YKL049C	YMR119W	YER092W	YGR151C

LISTADO DE ORFs ANALIZADOS EN EL GENOMA DE *Saccharomyces cerevisiae*				
GENES ANOTADOS		ORFs HIPOTETICOS		ORFs CUESTIONABLES
≈ 900 pb	≈ 300 pb	≈ 900 pb	≈ 300 pb	≈ 300 pb
YOR237W	YKL058W	YMR130W	YER121W	YGR160W
YOR288C	YKL097WA	YMR135C	YER135C	YGR164W
YOR299W	YKL156W	YMR204C	YER137C	YGR176W
YOR313C	YKL192C	YMR163C	YGL010W	YGR182C
	YKR057W	YOL063C	YGL015C	YGR190C
	YML009C	YOL070C	YGL121C	YGR228W
	YML129C	YOL072W	YGL188C	YGR230W
	YMR022W	YOL078W	YGL230C	YGR236C
	YMR042W	YOL091W	YGR030C	YGR259C
	YMR123W	YOR129C	YGR035C	YGR265W
	YMR175W	YOR138C	YGR215C	YGR269W
	YMR194W	YOR175C	YGR290W	YIL060W
	YMR251WA	YOR205C	YGR291C	YIL141W
	YMR256C	YOR245C	YGR293C	YIL163C
	YOL020C	YOR292C	YGR294W	YKL030W
	YOL053CA	YOR301W	YIL008W	YKL036C
	YOL109W	YOR342C	YIL032C	YKL076C
	YOR167C	YOR352W	YIL058W	YKL083W
	YOR210W		YIL059C	YKL11C
	YOR224C		YIL086C	YKL115C
	YOR265W		YIR020C	YKL118W
	YOR304CA		YIR040C	YKL123W
			YKL031W	YKL131W
			YKL044W	YKL136W
			YKL061W	YKL147C
			YKL084W	YKL177W
			YKL102C	YKL202W
			YKL137W	YKR012C
			YKL158W	YKR033C
			YKL223W	YKR047W
			YKL225W	YML013CA
			YKR032W	YML031WA
			YKR049C	YML058CA
			YKR073C	YML089C
			YKR083C	YML102CA
			YML058W	YMR052CA
			YML090W	YMR075CA
			YML108W	YMR086CA

LISTADO DE ORFs ANALIZADOS EN EL GENOMA DE *Saccharomyces cerevisiae*				
		ORFs HIPOTETICOS		ORFs CUESTIONABLES
			≈ 300 pb	≈ 300 pb
			YML122C	YMR119WA
			YMR057C	YMR135WA
			YMR082C	YMR153CA
			YMR103C	YMR158WA
			YMR107W	YMR172CA
			YMR122C	YMR173WA
			YMR141C	YMR193CA
			YMR195W	YMR290WA
			YMR230W	YMR294WA
			YMR252C	YMR304CA
			YMR254C	YMR306CA
			YMR324C	YMR316CB
			YMR325W	YOL035C
			YMR326C	YOL037C
			YOL026C	YOL046C
			YOL048C	YOL050C
			YOL085C	YOL099C
			YOL118C	YOL106W
			YOL131W	YOL134C
			YOL160W	YOL150C
			YOL166C	YOR041C
			YOR015W	YOR055W
			YOR024W	YOR082C
			YOR029W	YOR102W
			YOR053W	YOR105W
			YOR183W	YOR121C
			YOR252W	YOR135C
			YOR268C	YOR139C
			YOR343C	YOR146W
			YOR376W	YOR169C
				YOR170W
				YOR199W
				YOR200W
				YOR218C
				YOR225W
				YOR235W
				YOR248W
				YOR263C

LISTADO DE ORFs ANALIZADOS EN EL GENOMA DE *Saccharomyces cerevisiae*				
				ORFs CUESTIONABLES
				≈ 300 pb
				YOR277C
				YOR282W
				YOR300W
				YOR309C
				YOR331C
				YOR333C
				YOR345C
				YOR366W
				YOR379C