PREDICCIONES DE FUNCIÓN BASADAS EN LA ESTRUCTURA DEL GEN. ANÁLISIS DE ORFS DE FUNCIÓN DESCONOCIDA

Esperanza Cerdán Villanueva y María Angeles Freire Picos
Departamento de Biología Celular y Molecular
Universidad de La Coruña

1. ¿DE LA FUNCIÓN A LA SECUENCIA O DE LA SECUENCIA A LA FUNCIÓN?

Los métodos tradicionales de la genética molecular, y más concretamente en levaduras, tenían como punto de partida la obtención de mutantes que eran reconocidos por una variación fenotípica respecto a la línea salvaje o no mutada. Este mutante era después transformado con una librería genómica y la recuperación del fenotipo salvaje servía para la selección de un clon que, potencialmente, contenía el gen no mutado. Una característica fenotípica, una expresión manifestada por una cualidad reconocible de una determinada función, se utilizaba por tanto para clonar un gen. Por ejemplo, si obtenemos un mutante de *Saccharomyces cerevisiae* que contiene un gen no funcional del citocromo *c*, este mutante será incapaz de crecer en una fuente de carbono no fermentable, como puede ser el lactato, ya que para ello necesita utilizar la cadena respiratoria de la que forma parte el citocromo *c*. Si este mutante es transformado con una librería genómica y seleccionamos entre los transformantes aquellos que sean capaces de crecer en lactato, es probable que en estos clones, respiratoriamente competentes, encontremos una copia intacta del gen que codifica para citocromo *c*. Una vez que obtenemos este clon podemos realizar un análisis de su estructura, primero a través de un mapa de restricción y posteriormente mediante su secuenciación. Es decir en este modelo clásico, mucho antes de conocer la secuencia del gen ya conocíamos su función, porque en realidad era un camino que comenzando de la función nos llevaba hacia la estructura.

En la actualidad el desarrollo de poderosas técnicas de secuenciación, su automatización, y el creciente interés que ha tomado en la comunidad científica internacional la secuenciación de genomas completos de organismos pertenecientes a toda la escala evolutiva, desde pequeños procariotas hasta el

genoma humano, han hecho que el camino desde la función a la secuencia tenga que ser recorrido en sentido inverso. Por ejemplo, la terminación del proyecto de secuenciación de levadura, cuya secuencia total fue hecha pública en Bruselas el 24 de abril de este año 1996, nos revela que son muy numerosas las secuencias con potencial capacidad codificadora (ORFs) pero que no encuentran homologías con secuencias de genes de función conocida, ya sea en la propia levadura o en otros organismos. Estas ORFs, cuya función sigue siendo un enigma no resuelto, se han bautizado como "huérfanas". La situación en que nos encontramos es que conocemos su secuencia pero tenemos que averiguar su función. El desarrollo de algunas de las posibles estrategias que pueden seguirse para resolver esta cuestión constituye el objetivo de este trabajo.

2. SECUENCIA, ORDENADOR Y PREDICCIÓN DE FUNCIÓN.

La correcta integración de todos los datos obtenidos en los proyectos de secuenciación sería imposible si no dispusiésemos además de un apoyo informático. Podemos afirmar que la bioinformática ha experimentado un notable avance en la última década y el *sofware* desarrollado con aplicaciones en Biología Molecular es cada vez más potente. Sin embargo, este rápido avance o desarrollo de herramientas informáticas tiene una contrapartida, y es que no toda la información y programas se encuentran físicamente localizados en un único punto y, según cuales sean nuestros intereses en el análisis, tendremos que utilizar una vía u otra. Además los *software* desarrollados son en muchas ocasiones especie-específicos, de tal forma que un algoritmo aplicable a levaduras puede no ser útil en *Drosophila* o en humanos. La correcta utilización de estos recursos requiere por tanto un adecuado entrenamiento.

3. IDENTIFICACIÓN DE GENES

Recientemente J. W. Fickett ha realizado una interesante y amena revisión sobre el estado actual del reconocimiento de regiones codificadoras mediante programas de ordenador (Fickett, 1996) del que podemos tomar algunos datos como punto de partida (Tabla I). Los primitivos programas de identificación de regiones codificadoras estaban basados en la determinación de secuencias ATG, o alternativas a este codón de iniciación según especies, que permitiesen pautas de lectura hasta un determinado codón de terminación TAA, TGA, TAG. Sin embargo, la localización por este sistema representa tener una secuencia convenientemente verificada, sin fallos en la lectura, por lo que en la actualidad se utilizan programas que están basados en otros criterios como puede ser el contenido G+C, el uso de codones y otros. Cuando se trabaja con secuencias extensas procedentes de organismos eucariotas antes de realizar la búsqueda de ORFs conviene hacer una eliminación de regiones de DNA repetitivo. Aunque estas repeticiones pueden solapar con regiones transcritas por la RNA polimerasa II, no suelen solapar con regiones que pertenecen a exones o al promotor, por tanto su localización proporciona información sobre dónde no

están los genes. Muchos programas de detección de regiones codificantes combinan varios criterios mediante la aplicación de un algoritmo que resulta en un valor numérico al que se denomina discriminante. La principal ventaja de estos programas es que pueden ser aplicados a secuencias no verificadas y no demasiado extensas; el límite puede establecerse en torno a 100 pb para que el valor discriminante resulte significativo.

Tabla I.- Búsqueda de genes a través de Internet. (Traducida de Fickett, 1996)

Tipo de búsqueda	Nombre	Organismo	Acceso
Repeticiones	Pythia	Humanos	pythia@anl.gov
	Repbase	Humanos	ftp://ncbi.nlm.nih.gov
Homología	BLAST	Todos	blast@ncbi.nlm.nih.gov
	FASTA	Todos	fasta@ebi.ac.uk
Dominios	BLOCKS	Todos	blocks@howard.fhcrc.org
	ProfileScan	Todos	http://ulrec3.unil.ch/software/ PFSCAN_form.html
	MotifFinder	Todos	motif@genome.ad.jp
Identificación de genes (Integrados)	FGENEH	Humanos	service @theory.bchs.uh.edu
	GeneID	Vertebrados	geneid@bir.cedb.uwf.edu
	GeneMark	Varias sp.	genemark@ford.gatech.edu
	GeneParser	Humanos	http://beagle.colorado.edu/ eesnyder/GeneParser.html
	GenLang	*Drosophila*	*a*genlang@cbil.humgen.upenn.edu
	GRAIL	Humanos	grail@ornl.gov
	EcoParse	*E. coli*	ecoparse@cse.ucsc.edu
Reconocimiento de señales	PromoterScan	Eucariotas	danp@biosci.cbs.umn.edu
	NetGene	Humanos	netgene@virus.fki.dth.dk

Otros criterios para el reconocimiento de genes pueden estar basados en patrones comunes que se espera encontrar en el gen, por ejemplo en su promotor o en los sitios de corte y empalme de exones, en las inmediaciones del codón de iniciación ATG etc. En este caso se trata de buscar homologías de la secuencia respecto a consensos, algo similar a lo que se aplica en los progamas de búsqueda de sitios reconocidos por las enzimas de restricción. Sin embargo, hay una notable diferencia, y es que, si bien los sitios de restricción están constituidos por secuencias únicas en la mayoría de los casos, las regiones consenso pueden variar en determinadas posiciones. Para evitar este inconveniente se utilizan algoritmos basados en matrices ponderadas. Con esta estrategia se intenta incluir la información disponible sobre la importancia de cada base en cada posición de la señal.

Los programas más avanzados de que disponemos en la actualidad para la identificación de genes combinan varios de los criterios anteriormente expuestos y además la correcta interrrelación entre ellos, por eso se denominan programas integrados. Sin embargo, siempre hay que tener en cuenta que estos algoritmos pueden tener limitaciones, y además que sólo utilizan una fracción pequeña de todo el conocimiento científico. A modo de ejemplo podemos comentar que ninguno de los algoritmos existentes tiene en cuenta hechos científicos tan bien comprobados como la existencia de promotores carentes de TATAs o el *splicing* alternativo. En cualquier caso, dada la magnitud de los proyectos de secuenciación y la velocidad de obtención de datos, todos estos programas resultan imprescindibles para poder empezar a delimitar qué regiones de las secuencias leídas pueden corresponder a genes.

En relación con el proyecto de sequenciación del genoma de levadura se ha planteado una interesante pregunta a la hora de determinar qué ORFs de función desconocida serán analizadas en una segunda fase de análisis funcional. La pregunta proviene de cuál es el número de codones que debe contener una determinada ORF para ser considerada como tal. A efectos prácticos la selección se basó en que sólo se considerarían como potenciales ORFs las que fuesen mayores de 100 codones. La probabilidad de que una ORF de 100 codones aparezca por casualidad es muy pequeña, y por eso se decidió eliminar las que fueran menores a este límite ya que de este modo las ORFs elegidas tenían buenas perspectivas de ser funcionales; éste fue un criterio eminentemente práctico pero no con base funcional ya que existen ORFs con funciones conocidas y tamaños menores; por ejemplo, la menor descrita hasta el momento, PMP1, tiene 40 codones y codifica para una proteína de membrana. Tenemos pues certeza de que algunas de las ORFs que han sido relegadas del análisis funcional tienen sin embargo una función. El problema es que su número total es muy elevado y resultaría muy costoso hacer un análisis funcional para todas ellas a sabiendas de que muchas no serán ORFs reales. Simplemente analizando la mitad del número de cromosomas (I, II, III, V, VI, VIII, IX, XI) ya se encuentran más de 2.485 ORFs pequeñas. Es necesario buscar algún criterio para seleccionar entre estas las que sean potencialmente funcionales. El problema es que la mayoría de los programas aplicados para la identificación de genes se basan de alguna manera en la frecuencia de codones, y ésta es más difícil de evaluar correctamente a medida que disminuye el número de codones. Además, en *S. cerevisiae* existe una gran heterogeneidad en el uso de codones; la mayor parte de las ORFs presentan unos índices de desviación (*bias*) en el uso de codones bajos, y sólo una pequeña parte, que se correlaciona con los genes de alta expresión, presenta altos índices. Guiados por esta idea un grupo de investigadores (Barry *et al.*, 1996) han desarrollado un programa que permite actuar de filtro sobre estos genes pequeños. El método está basado en un análisis de correspondencia aplicado a las frecuencias de di-codones (hexámeros) en fase. Estas frecuencias se analizan en tres grupos testigo o de aprendizaje. Uno de estos grupos está constituído por ORFs que se sabe que codifican para proteínas y que presentan una gran desviación en el uso de codones (CBI > 0.3), otro por las ORFs que codifican para proteínas con escaso *bias* (CBI < 0.3) el otro grupo se construye por simulación al azar. De la comparación de estos grupos se obtiene una función discriminante que puede ser después aplicada a

una ORF problema. El resultado de aplicar el programa a las 2.485 pequeñas ORFs que mencionábamos resulta en una selección en la que 140 ORFs son posibles candidatas a ORFs reales, con una probabilidad menor de 0,01 de ser no codificantes. Con este procedimiento se han detectado también algunas repeticiones en estas ORFs pequeñas y pseudogenes. Estas 140 potenciales ORFs se sometieron a análisis de homología y búsqueda de dominios que se describen más adelante en este mismo apartado y también se compararon con las bases de datos de EST (Expressed Sequence Tags).

4. BÚSQUEDA DE HOMOLOGÍAS

Una vez identificado un gen, una posible manera de predecir su función consiste en comparar la homología de su secuencia, o la de la proteína codificada, con otras secuencias disponibles en las bases de datos. La lista de bases de datos que pueden ser consultadas es muy extensa y algunas de las más frecuentes se resumen el la Tabla II. Lo mismo sucede con los programas de búsqueda de homologías. Entre estos los más populares son los denominados FASTA y BLAST. En ambos casos se trata de programas de alineamiento dual que establecen comparaciones entre la secuencia problema y cada una de las secuencias de la base de datos, sus algoritmos tienen en cuenta el número de identidades (coincidencias exactas), el número de homologías (sustituciones equivalentes) y penalizan los *gaps* o huecos sin homología. Steven Brenner ha realizado una reciente y concisa revisión sobre estos temas (Brenner, 1996).

Tabla II.- Bases de datos distribuídas por EMBL*

Nombre	Descripción
3D ali	Alineamientos basados en estructuras
Alu	Secuencias Alu y alineamientos
Berlin RNA	Secuencias de RNA ribosómico de 5S
Bio-Catalog	Directorio de software de genética y biología molecular
Blocks	Base de datos de proteínas, Bloks
CpGisle	Base de datos de regiones CpG
Cutg	Uso de codones tabulado por GenBank
dbEST	Secuencias tags expresadas
dbSTS	Secuencias de sitios tags
DSSP	Estructura secundaria asignada a las proteínas recopiladas en PDB
ECDR	Base de datos de secuencias de *Escherichia coli*
EMBL	Base de datos de nucleótidos
Enzime	Base de datos de nomenclatura de enzimas
EPD	Base de datos de promotores eucariotas
FlyBase	Mapas genéticos de *Drosophila*
FSSP	Familias de proteínas con similitud estructural
HaemA	Hemofilias del grupo A
HaemB	Hemofilias del grupo B
HLA	Secuencias HLA I y II
HSSP	Alineamientos de proteínas con similitud estructural
IMGT	Inmunogenética
Kabat	Proteínas de interés inmunológico
LiMB	Listado de bases de datos de biología molecular
Lista	Regiones codificadoras de levadura

Methyl	Sitios de metilación específica
Misfolded	Modelos de proteínas con deliberado plegamiento incorrecto
NRL3D	Base de datos secuencia-estructura
NRSub	Genoma de *Bacillus subtilis*, no redundante
Nucleosomal DNA	Secuencia de DNA nucleosomales
P53	Mutaciones P53
PBD	Base de datos de proteínas de Brookhaven
PDB Select	Listado representativo de identificadores de cadena
PIR	Base de datos de proteínas "Internacional"
PKCDC	Dominios catalíticos de protein Kinasas
Prints	Motivos proteicos
Prodom	Dominios de proteínas
Prosite	Base de datos de patrones
PUU	Base de datos de dominios estructurales
REBASE	Base de datos de enzimas de restricción
REILibrary	Listado de enzimas de restricción
RepBase	Secuencias prototipo de DNA repetitivo en humanos
Rhdb	Hibridos de radiación
RLDB	Referencias
rRNA	Secuencias de rRNA
SBASE	Dominios protéicos
SeqAnalRef	Bibliografía de análisis de secuencias
SmallRNA	Secuencias de RNA de pequeño tamaño
SRP	Señales de reconocimiento SRP
SWISS-PROT	Secuencias de proteínas
TFD	Factores de transcripción
TransFac	Elementos cis y trans en eucariotas
Transterm	Señales de terminación de la traducción
tRNA	Secuencias de rRNA
Yeast	Base de datos de los cromosomas de la levadura *Saccharomyces*

*Acceso:**WWW**: http://www.ebi.ac.uk; **E-mail**: netserv@ebi.ac.uk; **FTP anónimo**: ftp.ebi.ac.uk

Todas las búsquedas realizadas a través del National Center for Biotechnology Information (NCBI) se realizan mediante BLAST. Otros métodos pueden ser empleados por medio del European Bioinformatics Institute (EBI), Oak Ridge National Laboratory (ORNL) y EERIE. A través de estas conexiones se tiene acceso a FASTA y también al algoritmo Smith-Waterman (también conocido como SSERCH o BLITZ). En general la mejor opción consiste en empezar por un análisis de BLAST, utilizando por defecto los parámetros y matrices del programa. Si no resultase adecuado se puede acudir a reacondicionar los parámetros iniciales o a emplear otros programas más lentos, pero en determinados casos de mejor resolución, como FASTA o BLITZ. Hay una extensa colección de sistemas de comparación de secuencias y bases de datos para fines específicos que son de dominio público vía E-*mail* y en las páginas WWW.

Un caso especial de búsqueda de homologías, que nos da una información adicional, es la comparación frente a las bases de datos de EST (Expressed Sequence Tags). Estas bases de datos están formadas por secuencias procedentes de cDNA y corresponden por tanto a secuencias no muy largas obtenidas mediante una sola lectura. Puesto que el cDNA se ha obtenido a partir

de un mRNA, la homología de nuestra secuencia con una secuencia de la base de datos EST nos indica que existe una alta probabilidad de que se transcriba y por tanto sea funcional.

5. IDENTIFICACIÓN DE DOMINIOS

Una herramienta más sofisticada para la posible predicción de función consiste en la identificación de dominios. Por ejemplo regiones de la proteína que por sus características puedan reconocerse como regiones de membrana, sitios de unión para determinados coenzimas, regiones de interacción específica con DNA, regiones conservadas evolutivamente en una determinada familia de proteínas etc. El tratamiento clásico de este tipo de búsqueda se basa en alineamientos múltiples como el popular CLUSTAL o en modelos de predicción de la estructura secundaria y terciaria.

Algunas bases de datos son específicas de dominios proteínicos, por ejemplo ProDom, Sbase, Prosite. Una versión especializada de BLAST, a la que se denomina BEAUTY, o los programas BLASTPAT o FASTPAT (Todos ellos disponibles a través del Bailor College of Medicine search Launcher) pueden comparar una determinada secuencia con una base de datos compuesta por patrones basados en homologías. Estas búsquedas basadas en patrones más que en secuencias son especialmente útiles para encontrar homologías dentro de familias extensas.

6. PREDICCIÓN DE LA LOCALIZACIÓN SUBCELULAR

En las células eucariotas, las señales de transporte de proteínas a diferentes compartimentos están presentes en la propia secuencia y se conoce con bastante certeza su naturaleza (Verner y Schatz, 1988). Esto permite hacer predicciones de la localización celular basadas en la secuencia de aminoácidos. Uno de los métodos más utilizados es el diseñado por Nakai and Kanehisa (1992) que permite identificar secuencias que dirigen el tráfico de la proteína hacia el núcleo, la mitocondria, el exterior de la célula etc., y en el que está basado el programa Pshort. El método utilizado por estos autores consiste en adaptar el conocimiento empírico sobre la composición de las señales a un sistema que pueda ser procesado en el ordenador. Se trata de un sistema experto de inteligencia artificial que utiliza una base de conocimiento organizada conforme a unas reglas antecedente-consecuente, si A-entonces B (*if-then*), a través de un árbol lógico de razonamiento.

7. ANÁLISIS FUNCIONALES

En muchos casos, los análisis bio-informáticos basados en las secuencias aminoacídicas deducidas a partir de la secuencia de DNA proporcionan poca información acerca del papel real de la proteína, y en cualquier caso, siempre se requiere demostrar experimentalmente su funcionalidad. A continuación

comentaremos algunas de las metodologías experimentales más comunmente empleadas para este estudio a gran escala.

8. DISRUPCIÓN GÉNICA

Analizar el fenotipo producido al interrumpir o eliminar un gen puede ayudar a conocer en qué procesos celulares se encuentra implicado su producto. Históricamente esta es una de las técnicas más utilizadas para verificar la funcionalidad de genes que habían sido clonados por complemetar alguna mutación en un gen de interés. Al interrumpir el gen, si este era funcional, se obtendría un fenotipo similar al de la mutación que había complementado (o incluso más drástico). En algunos casos se aplica la estrategia opuesta es decir, en lugar de anularlo se sobre-expresa el gen, bien colocándolo bajo un promotor fuertemente inducible o bien clonándolo en un vector que genere alto número de copias en la célula, posteriormente se efectúa el análisis fenotípico. En general se tiende más a analizar el efecto de la disrupción; la sobre-expresión, aunque puede ser útil para una primera aproximación da una idea más indirecta del papel en la célula porque puede estar afectando a "terceros" que no podemos controlar.

De un modo clásico, Rothstein, 1991, hace una revisión de metodologías de integración y disrupción génica. Este autor propuso un sistema para interrumpir genes en levaduras que implica los siguientes pasos: clonación del gen, construcción de un mapa de restricción, introducción de un marcador seleccionable para interrumpir o reemplazar la secuencia codificadora, y finalmente, obtención de un fragmento lineal que se introduce por tansformación en las levaduras y que, por recombinación con las secuencias genómicas homólogas, da lugar a la interrupción del gen existente con el gen marcador. En sistemas de mamíferos se utiliza una metodología similar y ha sido ampliamente aplicada en células *stem* de embriones de ratón. Las frecuencias de actuación específica sobre el gen de interés son relativamente bajas probablemente debido al hecho de que el DNA transfectado tambien puede integrarse aleatoriamente en cualquier sitio cromosómico (Hasty y Bradley, 1993).

Cabe señalar la variabilidad de las diferentes regiones genómicas en su capacidad para la recombinación y para las frecuencias de integración. Así por ejemplo la región en la que se localiza el gen LEU2 de *S. cerevisiae* exhibe frecuencias de integración 100 veces menores que la de HIS3. La utilización de fragmentos lineales ayuda a forzar la recombinación en la proximidad de los puntos de corte. Esto es especialmente importante cuando el marcador seleccionable es también una secuencia presente en el genoma de la levadura; así por ejemplo, si queremos interrumpir el gen que codifica para citocromo *c* con URA3, debemos linealizar la construcción en el citocromo *c*, con esto evitamos obtener levaduras URA^+ que simplemente han recuperado su alelo URA salvaje por recombinación. En la Figura 1A se muestra el fundamento de este procedimiento.

Entre los inconvenientes de esta metodología destacamos el que se requiere disponer del gen clonado, elaborar el mapa de restricción, la presencia de los sitios de restricción adecuados para incluir el marcador y posteriormente

poder linealizar la construcción con un enzima que permita aislar la construcción de la disrupción entera. La aplicación de este método requiere trabajar con cepas mutantes en varios marcadores, de lo contrario al interrumpir el gen con un marcador, por ejemplo URA3, se obtiene una cepa URA^+, eliminando la posibilidad de trabajos posteriores en esa cepa con plásmidos que lleven dicho marcador. Para solucionar este problema Scherer y Davis en 1979 diseñaron el método de reemplazamiento *Pop-in/pop-out* (salto dentro-salto fuera) que implica dos pasos: integración del plásmido empleando el marcador URA3 como marcador seleccionable y posterior escisión de la secuencia plasmídica conteniendo el marcador URA3. En este caso se introduce una mutación en el gen que se desea interrupir y se utiliza un vector integrativo. La integración de la molécula circular da lugar a la repetición directa del gen de interés, de modo que tendremos una copia salvaje y otra mutante. El emparejamiento de las repeticiones directas puede conducir a la pérdida de la secuencia plasmídica segun se ilustra en la Figura 1B. Este último evento puede ser seleccionado haciendo crecer a las levaduras en medio 5-FOA (ácido 5-fluoro orótico, cuya presencia es tóxica para las levaduras de fenotipo URA^+). Aquellos entrecruzamientos que sucedan en el lado apropiado del sitio mutado conducirán a reemplazar el sitio salvaje por la secuencia mutada (Rothstein 1991).

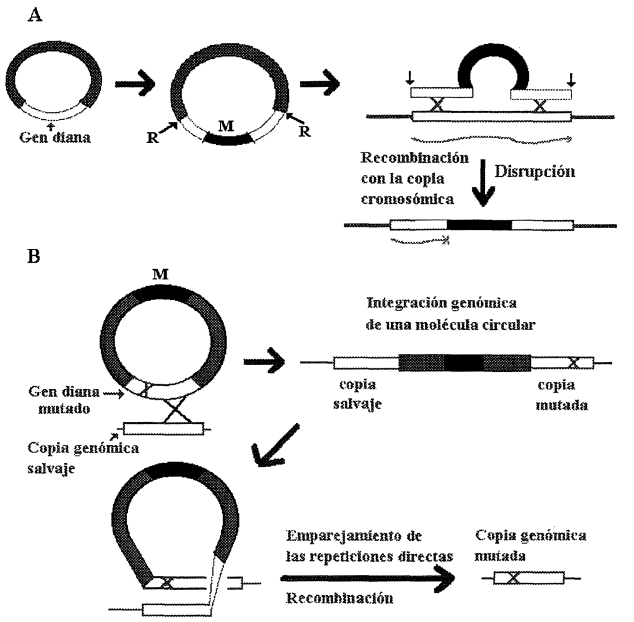

Figura 1.- A: Modelo de disrupción génica según Rothstein (1991). **B:** Modelo de disrupción *pop-in/pop-out* (Sherer and Davis, 1979).

La idea de eliminar el gen marcador por recombinación ha servido como modelo para otros trabajos posteriores, por ejemplo Roca *et al.* (1992) utilizan un sistema basado en una recombinasa sitio-espescífica inducible por galactosa. Para ello situan dos repeticiones directas del sitio reconocido por la recombinasa de *Zygosaccharomyces rouxii* a ambos lados del gen marcador (en este caso URA3) en un plásmido que permite clonar las secuencias génicas flanqueando los dos sitios de la recombinasa. Tras la integración (seleccionable por crecimiento en ura$^-$) se realiza una segunda transformación con un segundo plásmido que lleva la recombinasa expresada a partir del promotor inducible por galactosa (*GAL*1 de *S. cerevisiae*) y un marcador LEU. Al cultivar las levaduras que llevan la integración en un medio con galactosa se expresa la recombinasa y como resultado final obtenemos la disrupción del gen de interés y mantenemos la capacidad de utilizar URA como marcador para posteriores trabajos.

La introducción de las técnicas de PCR y el conocimiento de la secuencia de numerosos genes disponibles en las bases de datos han dado lugar al desarrollo de múltiples metodologías para interrumpir genes a partir de fragmentos generados tras varias rondas de PCR, en ocasiones sin necesidad de pasos de clonación. Estas técnicas se basan en la utilización de cebadores (*primers*) oligonucleotídicos de aproximadamente 55 nt que presentan homología con el gen de interés y con un marcador génico (Baudin *et al.,* 1993, Wach *et al.,* 1994, Amberg *et al.,* 1995). El producto resultante es el gen marcador flanqueado por una región pequeña homóloga al gen que deseamos interrumpir que puede utilizarse directamente para la integración. Entre las limitaciones de este método se incluye la necesidad de trabajar con cepas que posean deleciones totales de estos marcadores para evitar la conversión. Por otro lado la frecuencia de transformación de estos productos de PCR con regiones cortas de homología (30 nt) es baja y finalmente, tras la deleción del gen, el marcador permanece integrado en el genoma y no puede ser utilizado en la cepa modificada.

Schneider y colaboradores (1996) diseñaron un *cassette* "reciclable" para disrupciones génicas en *S. cerevisiae* su trabajo es una combinación de la técnicas de PCR y las de eliminación por recombinación de repeticiones directas (este último paso sucede según el esquema de la Figura 1B. Se trata de diseñar cebadores con 50 bases del extremo 5´ homólogas a la secuencia diana y 18 bases del extremo 3´ homólogas al vector (pMPY-ZAP) que lleva el marcador genético flanqueado por dos repeticiones. Tras la amplificación por PCR se consigue un fragmento de 2,1kb que es utilizado para la integración en el genoma. Posteriormente, al igual que en casos anteriores, se selecciona por crecimiento en 5-FOA aquellos que sufran recombinación y la consiguiente eliminación de URA.

Recientemente Maftahi y colaboradores han descrito la metodología necesaria para las disrupciones sistemáticas de genes de *S. cerevisiae* (Maftahi *et al.*, 1996). En la Figura 2 se esquematiza el procedimiento.

Se trata de realizar dos rondas de PCR.

1ª- Amplificación de los extremos 5´y 3´del gen con cebadores que incluyen "colas" de 16 nt en el extremo 5´y 3´de los fragmentos de inicio y terminación respectivamente. Estas son capaces de hibridar entre sí por emparejamiento de bases. Además contienen la secuencia de reconocimiento para un enzima de restricción de corte poco frecuente como AscI.

Tras la segunda amplificación por PCR se consigue un fragmento que contiene las regiones 5´y 3´ en la misma orientación separadas por la secuencia de reconocimiento AscI.

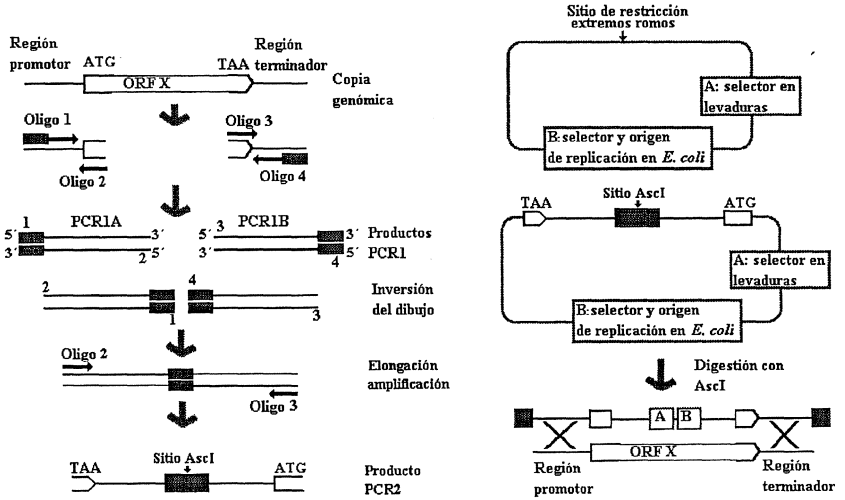

Figura 2..- Modelo de disrupción génica propuesto por Maftahi y colaboradores (1996).

El siguiente paso es la clonación del producto de PCR en un vector con un gen selector para levaduras y un selector y origen de replicación en *E. coli.*

Este paso tiene una doble utilidad:

a) Linearizar el plásmido con AscI para dirigir la disrupción génica por doble entrecruzamiento.

b) Introducir un gen informador bajo el control del promotor del gen que estamos analizando para estudios de expresión.

La disrupción se comprueba por PCR con dos cebadores situados en el genoma a ambos lados de la región interrumpida (que amplificarían toda la región modificada) y también mediante amplificaciones en las que intervienen cebadores que solo anillan en la región integrada.

En los casos en los que se desconoce si se trata de un gen esencial es necesario efectuar la disrupción en una célula diploide. El análisis de las tétradas

producidas nos pemitirá conocer si el gen es esencial (sólo dos tétradas viables) o no, en cuyo caso las cuatro tétradas serán viables.

9. LIBRERÍAS DE FUSIÓN A GENES INFORMADORES

Se trata de librerías genómicas con un gen informador situado bajo el control de elementos reguladores de levaduras. Los genes informadores suelen codificar para enzimas que catalizan reacciones cuyo producto puede ser visualizado fácilmente (enzimas informadoras). Los genes más empleados, la mayoría de ellos pertenecientes a *E. coli*, incluyen aquellos que codifican para la β-galactosidasa (*lacZ*), cloramfenicol acetil transferasa (CAT), galactokinasa (GALK), glucuronidasa (GUS) y también otros cuyos productos presentan actividad bioluminiscente, como el gen de la luciferasa bacteriana o el de luciérnaga (Docherty y Clark, 1993).

La construcción de librerías de expresión resulta de gran utilidad para la identificación de genes implicados en procesos celulares específicos. Dang *et al.* (1994) utilizaron este sistema para clonar genes que se encuentren regulados por un activador transcripcional específico, la proteína Hap2, que reconoce la secuencia CCAAT en promotores de genes implicados en el ciclo de Krebs, la cadena de transporte de electrones y la biosíntesis de hemo.

Se transforma con la librería de fusión una cepa de levadura en la que el gen que codifica para el activador transcripcional se encuentra bajo el control de un promotor inducible; con ello se regula la expresión del factor. Esto permite comparar la coloración de las colonias en condiciones de inducción y no inducción del factor transcripcional y únicamente en aquellos casos en que el factor se una al promotor del gen se apreciarán cambios de color. El esquema del procedimiento seguido se muestra en la Figura 3.

Estos autores construyeron dos tipos de librerías de fusión:

1.- Librería construida en el plásmido YEp366 con el gen *lacZ* (sin ATG) situado 3´respecto al MCS donde clonaron los fragmentos de DNA genómico (procedentes de una digestión parcial con Sau3A). Se calcula que una de cada 6 inserciones se encuentra en la orientación y posición correcta de pauta de lectura.

2.- Una segunda librería construída con el transposón mini-Mu (derivado de MudIIZZ4). Este transposón contiene elementos necesarios para replicación en levaduras y bacterias así como el gen *lacZ* de *E. coli* sin ATG; conduce a la formación de proteínas de fusión con β-galactosidasa cuando el transposón se inserta en la correcta pauta de lectura en una ORF. Para ello se transforma la cepa de *E. coli* portadora del transposón con una genoteca de levaduras y posteriormente se seleccionan los transformantes.

Con el DNA plasmídico procedente de ambas librerías se transformó una cepa de levaduras que contenía el gen HAP2 bajo la dirección de un promotor inducible por galactosa en un plásmido URA^+. Los transformantes crecidos en placas LEU^+ fueron replicados a placas X-Gal para seleccionar aquellos que

portaban la proteína de fusión, que crecerán como colonias azules. Cada colonia azul fué posteriormente sembrada en placas 5-FOA con objeto de seleccionar a las células que pierden el activador HAP2. El objeto de éste último paso es tener un control de la misma levadura con la proteína de fusión pero sin el activador transcripcional. De esta manera se puede comparar la producción de β-galactosidasa en medios con y sin galactosa. Aquellas levaduras que incrementen la producción de β-galactosidasa al crecer en galactosa sólo cuando HAP2 está presente contendrán genes presumiblemente regulados por este factor. Cabe señalar que en cualquiera de los dos casos la representación del genoma es sólo parcial, pero con todo permitió caracterizar nuevos genes regulados por este factor. Como se puede apreciar en la Figura 3, la librería construída en el plásmido permitió obtener mayor número de genes regulados por HAP2. Entre los genes analizados por este sistema encontraron que HAP2 regula procesos mitocondriales como transporte de proteínas al interior de la mitocondria, o maduración de tRNAmitocondrial.

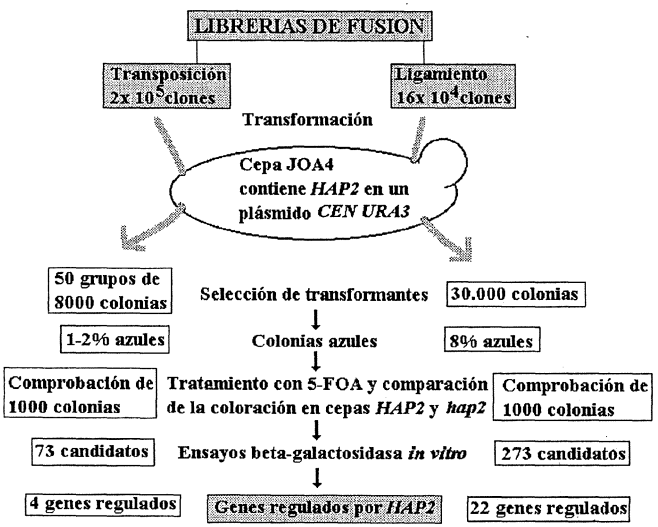

Figura 3.- Ejemplo de utilización de dos librerías de fusión para la identificación de genes regulados por un factor transcripcional específico (Hap2) según Dang y colaboradores (1994).

10. EL SISTEMA DEL DOBLE HÍBRIDO

Utilizado para el estudio de interacciones proteína-proteína en el proceso de activación transcripcional, fué propuesto por Fields y Song en 1989. Se trata de trabajar con dos dominios de Gal4. Esta proteína es un potente activador transcripcional del gen *GAL1* cuando las células crecen en medio con galactosa.

El dominio N-terminal se une a la secuencia de DNA de un modo específico pero no produce activación de la transcripción. Por su parte, el dominio C-terminal que contiene la región responsable de la activación no es capaz de ejercer su papel por sí solo puesto que no es capaz de reconocer la secuencia específica del promotor.

Basándose en esta propiedad Fields y Song diseñaron un sistema con dos plásmidos que codifican para proteínas híbridas (Figura 4A). El plásmido con el híbrido A contiene el dominio de unión al DNA del factor transcripcional de levadura Gal4 fusionado a la secuencia de la proteína conocida X. El plasmido B contiene el dominio de activación de Gal4 fusionado a la proteína Y. Cuando ambos plásmidos se expresan en la misma célula se obtienen dos proteínas híbridas. Si los factores X e Y interaccionan entre sí se obtiene un dímero que contiene los dominios de unión a DNA y activación transcripcional de Gal4. Para verificar esta interacción se recurre a un plásmido con un promotor que contenga el dominio de unión de GAL4 seguido de un gen informador (*lac*Z) podremos verificar fácilmente esa interacción:

— Si las proteínas X e Y interaccionan se producirá el dímero híbrido que a su vez activará la transcripción de *lacZ*, con la consiguiente producción de β-galactosidasa; por tanto las colonias de levaduras crecerán, en medio X-Gal, de color azul.

— Si no interaccionan no se forma el dímero y por consiguiente no hay activación de *lac*Z. Las levaduras crecerán de color blanco en el citado medio.

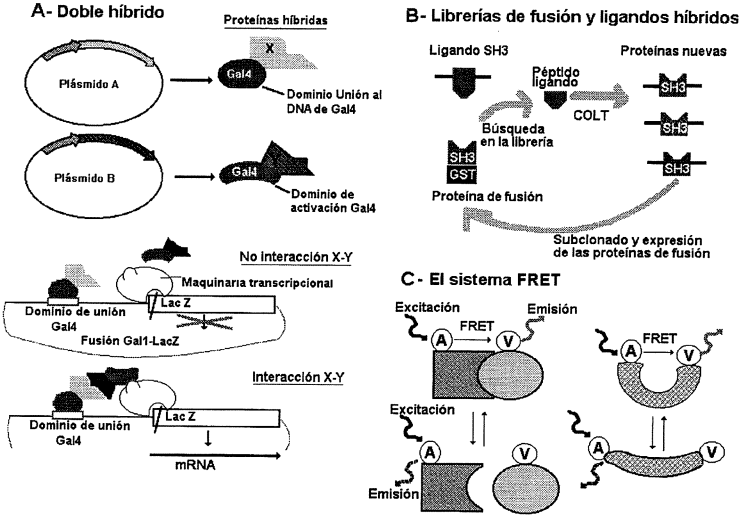

Figura 4.- Técnicas para el estudio de interacciones proteína-proteína que nos permiten clonar los genes involucrados. **A:** El sistema del doble híbrido (Fields and Song, 1989) **B:** El sistema COLT (Sparks *et al.,* 1996) **C:** Aplicaciones de GFP, el sistema FRET (Cubitt *et al.*, 1995).

Es bastante común encontrar la combinación del dominio de unión a DNA de Gal4 y el dominio de activación de la proteína VP16 de herpes virus. La fusión GAL4-VP16 es eficiente en la activación transcripcional de genes de levaduras, por ello también se utiliza en el sistema de doble híbrido colocando cada dominio en un vector diferente (Marcus *et al.*, 1994).

El sistema del doble híbrido puede ser utilizado al igual que las librerías de fusión para seleccionar aquellos genes de una librería cuyos productos interaccionan con un determinado factor (en este caso la interacción ha de ser proteína-proteína). Para ello de construye la librería en un vector que fusione el DNA genómico con el dominio de activación de Gal4. Se transforma la levadura con un vector que contenga la fusión GAL4 -gen conocido y, tras asegurar la presencia del plásmido en la levadura, se efectúa la segunda transformación, esta vez con la genoteca de fusión. Aquellas levaduras transformadas con un híbrido que interaccione con la proteína conocida serán capaces de expresar β-galactosidasa dirigida desde el pomotor *GAL1* (Chien *et al.*, 1991). La fusión *GAL1-lacZ* puede estar integrada en el genoma de la levadura lo que elimina un paso de transformación (Bendixen *et al.*, 1994).

11. FUSIONES A GFP (GREEN FLUORESCENT PROTEIN)

El grupo de Jonhston (Universidad de Washington) en colaboración con Hegeman (Universidad de Giessen, Alemania) han desarrollado un sistema de fusión a la proteína GFP, (*Green Fluorescent Protein*), que está ya siendo aplicado en el análisis funcional sistemático del genoma de levadura, concretamente para todas las ORFs de función desconocida del cromosoma VIII.

La proteína GFP obtenida del pez *Aequorea victoria* emite una fluorescencia verde cuando es excitada por luz azul de 395 nm. A diferencia de otros sistemas de fluorescencia, no necesita de la adición de otras proteínas, cofactores o sustratos lo que hacen este sistema especialmente adecuado para estudios *in vivo*. El gen *GFP* codifica para una proteína de 238 aminoácidos que sufre una única modificación postranscripcional que es precisamente la responsable de la formación del cromóforo. Esto sucede por ciclación de un pequeño tripéptido (entre las posiciones 65-67) formado por Ser-Tyr-Gly, que es posteriormente oxidado a nivel de la Tyr por el oxígeno atmosférico. Las fusiones a esta proteína fluorescente se han utilizado en distintos sistemas biológicos, desde bacterias a *Drosophila*, para comprobar expresión génica y localización celular (Cubitt *et al.* 1995).

Los vectores de expresión recientemente desarrollados, permiten obtener fusiones N-terminales o C-terminales del gen *GFP* al gen de interés y seguir la localización celular en células vivas de levadura. La proteína GFP ha sido también utilizada para construir un *cassette* de reemplazamiento que, después de recombinación homóloga, se integra en el genoma de la levadura a nivel del locus genómico de interés, de forma que la expresión de *GFP* queda bajo el control del promotor en estudio. El *cassette* de reemplazamiento está constituido por el gen *GFP* seguido de un gen marcador que es *HIS3*. El *cassette* se amplifica utilizando oligonucleótidos complementarios con colas 5' homólogas a

la región donde se quiere realizar la recombinación, según la técnica desarrollada por Baudin (Baudin *et al.* 1993). Una vez conseguida la integración y verificada, la cantidad de GFP producida a partir de este promotor puede después cuantificarse en células vivas por citometría de flujo (Niedenthal *et al.,* 1996).

Otra de las aplicaciones de la GFP es el estudio de interacciones proteína-proteína gracias al sistema conocido como FRET (*Fluorescence Resonance Energy Transfer*). Este es un sitema espectroscópico para monitorizar la asociación dinámica de macromoléculas en células vivas. Requiere un método general para marcar una de las macromoléculas con un fluoróforo donante y la otra con un fluoróforo aceptor, normalmente se puede hacer con los vectores de fusión a GFP disponibles. Además es necesario que el espectro de emisión del donante se solape con el espectro de absorción del receptor y que los solapamientos de los dos espectros de absorción y los dos espectros de emisión sean mínimos. En la Figura 4C se muestra el mecanismo para determinar la proximidad proteína-proteína mediante FRET utilizando 2 mutantes de GFP con dos coloraciones diferentes. La transferencia de energía desde el mutante que emite azul hasta un segundo mutante que emite verde depende de la distancia entre los dos fluoróforos, permitiendo conocer la proximidad de dos subunidades o dominios de una proteína (Cubitt *et al.,* 1995). Este sistema presenta la ventaja, frente al sistema del doble híbrido, de que no se requiere el transporte al núcleo para que se produzca la interacción proteína-proteína.

12. SISTEMA COLT PARA ESTUDIO DE INTERACCIONES PROTEÍNA-PROTEÍNA

Las librerías combinatorias de péptidos han sido utilizadas para definir ligandos que se unen específicamente a un determinado dominio de un factor proteico. Así por ejemplo, Rickles y colaboradores crearon una librería que consistía en 6 residuos de aminoácidos aleatorios flanqueados en sus extremos N y C terminal por gliccocola y unidos a GST (glutatión S-transferasa) para estudiar la especificidad de los ligandos que se unen a SH3 (Src Homology 3, presente en proteínas que median procesos intracelulares de transducción de señales). Estos ligandos aleatorios eran insertados entre las secuencias del gen III del bacteriofago M13, generando proteínas híbridas que incluían estos 8 aminoácidos. Posteriormente se enriquecía la población de fagos con ligandos específicos mediante experimentos de unión al híbrido GST-Src inmovilizado sobre poliestireno. Normalmente al cabo de 10 ciclos conseguían un enriquecimiento de 1.000 veces, tras lo cuál aislaban el DNA de los bacteriofagos y secuenciaban la región de la inserción (Rickles *et al.,* 1994).

En ocasiones los dominios de los ligandos estudiados dan "reacciones cruzadas" terminología utilizada en este caso para señalar que un determinado péptido selecciona dominios protéicos que difieren de la secuencia consenso y que en principio, no están relacionados con SH3; Sparks y colaboradores (1996), aprovecharon esta circunstancia y desarrollaron un sistema para clonar nuevos genes o porciones de genes cuyo producto proteico es el ligando de un dominio conocido. Este fué denominado sistema COLT (*Cloning of Ligand Targets*) por

analogía con el sistema CORT (*Cloning of Receptor Targets*) procedimiento en el que se utilizaban librerías de expresión de cDNA para buscar proteínas ligando de una región específica del factor receptor de crecimiento epidérmico (EGFR).

Para probar su sistema utilizaron el dominio SH3 presente en proteínas que median procesos intracelulares de transducción de señales (Figura 4B). Prepararon sondas de péptidos sintéticos biotinilados capaces de unirse a SH3, pero que también dan reacciones cruzadas con otros dominios no relacionados con SH3. Para mejorar la fuerza de la señal acomplejaron los péptidos con streptadivina-fosfatasa alcalina (SA-PA). Con esta sonda analizaron librerías de expresión de ratón y humanos. El sistema de búsqueda fue la hibridación de filtros. Tras la identificación de los clones positivos se extraía el DNA plasmídico o bien en el caso de las librerías en lgt11 y lgt22 se amplificaba por PCR. Estos clones fueron posteriormente secuenciados y caracterizados.

La comprobación *in vitro* de la especificidad para la unión a SH3 se realiza en pocillos con proteína de fusión SH3-GST inmovilizada. Los ligandos se incubaban como complejos SA-AP, o bien como péptidos biotinilados monovalentes (la diferencia es la cantidad de señal obtenida). De los 74 clones analizados 69 contenían al menos un dominio SH3. Estos clones codificaban para 18 proteínas diferentes con dominio de unión a SH3, 10 de las cuales eran desconocidas hasta el momento. El sistema COLT permite identificar con cierta rapidez cDNAs que codifiquen virtualmente para cualquier dominio funcional de interés. COLT evita las limitaciones de los métodos convencionales de análisis de genotecas utilizando DNA puesto que permite identificar proteínas con secuencias altamente divergentes pero que poseen actividad funcional equivalente.

13. EL PROYECTO EUROFAN: OBJETIVOS Y ORGANIZACIÓN

Podemos afirmar que el proyecto EUROFAN (*European project for Funtional Analysis of the S. cerevisiae genes*) es el primer proyecto de investigación sistemática de la dotación génica completa de un organismo. Teniendo en cuenta que la secuenciación total del genoma de la levadura ha revelado la existencia de aproximadamente 5.800 genes, de los que 1/3 son genes cuya función había sido ya abordada por métodos experimentales y era por tanto conocida con anterioridad, 1/3 tienen una función sólo asignada mediante análisis de homologías, y el resto son de función desconocida, el número de genes para los que es imprescindible realizar una caracterización funcional desde un abordaje experimental se eleva a 2.000. La comunidad científica europea a través de una extensa red de laboratorios ha abordado en una primera fase, que se inició a finales de 1995 y tiene una duración de dos años, el análisis funcional sistemático de 1.000 ORFs. En una segunda fase, que abarcará los dos años siguientes 98-99, se analizarán los otros 1.000 genes restantes.

La organización de Eurofan es jerárquica, de modo que los análisis esenciales son realizados por los denominados "consorcios" de recursos. La

información recabada a través de estos estudios permite ir dirigiendo las distintas ORFs hacia un análisis más exhaustivo realizado por los denominados "nodos" de análisis funcional que están especializados en una determinada línea de investigación.

El objetivo del análisis funcional no es reemplazar la metodología clásica de investigación sino acelerarla creando bases de datos con información clave sobre la función de los genes, y creando un banco genético de líneas mutantes y genes.

Dentro del amplio abanico de investigaciones desplegado en EUROFAN I la prioridad reside en el consorcio B0, ya que este consorcio es el responsable de preparar las herramientas genéticas que van a ser utilizadas por el resto de los nodos. Todos los laboratorios implicados en alguno de los nodos o consorcios de EUROFAN están también comprometidos a realizar este análisis básico que, al ser distribuido en lotes de 6 ORFs, se conoce con el nombre de *Six-Pack Analysis*. Los análisis que se desarrollan dentro de este consorcio aparecen representados en la Figura 5.

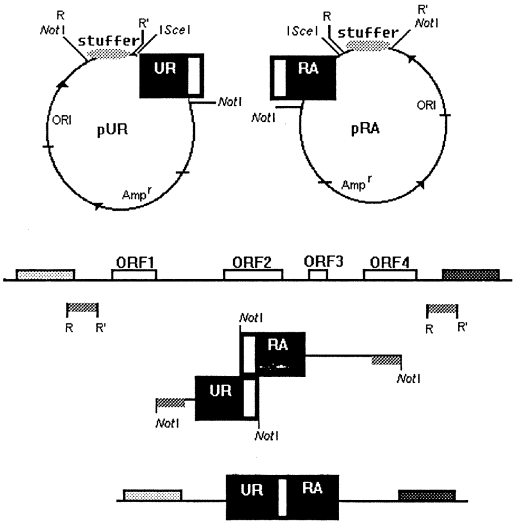

Figura 5.- Esquema del empleo de vectores con marcadores fragmentados (*split-markers*) en las técnicas de deleción múltiple mediante*mass-murder* (Fairhead *et al.*, 1996).

La descripción de las distintas actividades del resto de los consorcios y nodos que fueron incluidos en la primera fase de EUROFAN se esquematizan el la Tabla III.

Tabla III.- Organización de EUROFAN I (1995)

COORDINACION GENERAL
A1.-Gerencia
A2.-Bioinformática
A3.-Relación con la Industria YIP (*Yeast Industry Plataform)*
A4.-Archivo genético
CONSORCIOS
B0.-Generación de deleciones y herramientas genéticas
B1.-Analisis fenotípico cualitativo
B2.-Expresión-RNA. Mapas de transcripción
B4.-Expresión-proteína. Fusiones a genes marcadores
B5.-Interacciones génicas. Análisis del doble híbrido
B6.-Estructura subcelular y organelos
B7.-Relación con otros genomas
B8.-Nuevas metodologías
NODOS
N1.- Síntesis de DNA y ciclo celular
N2.- Síntesis de RNA y procesamiento
N3.- Traducción
N4.- Respuestas a estrés
N5.- Pared celular y Morfogénesis
N6.- Transporte
N7.- Metabolismo energético y carbohidratos
N8.- Metabolismo de lípidos
N9.- Metabolismo especial
N10- Desarrollo
N11- Mutagénesis, reparación, meiosis
N12- Estructura de los cromosomas
N13- Arquitectura celular
N14- Secreción y tráfico de proteínas

14. EL PROYECTO EUROFAN Y EL DESARROLLO DE NUEVAS TECNOLOGÍAS

Diversos grupos de investigación europeos y también de fuera de la comunidad, motivados por el mismo espíritu competitivo que ha sobrevolado sobre la realización del proyecto de secuenciación, han desarrollado y están desarrollando nuevas tecnologías más potentes tanto para la secuenciación a gran escala, como para un análisis funcional sistemático. Algunas de estas técnicas se perfilan brevemente a continuación.

15. DISRUPCIONES MÚLTIPLES (*MASS MURDER*: ASESINATO EN MASA)

El análisis funcional tal como se plantea en el consorcio B0 (*Six-Pack Analysis*) de EUROFAN I es un análisis gen a gen. El laboratorio de Dujon (Institut Pasteur, París) viene desarrollando una nueva técnica de disrupción múltiple que ha bautizado con el terrorífico nombre de *Mass-Murder*. La principal ventaja de esta estrategia es que permite estudiar de una manera

conjunta la función de varios genes mediante un planteamiento de tipo combinatorio.

El análisis combinatorio de los genes de levadura ha sido posible gracias al desarrollo de nuevos vectores con un marcador fragmentado, *split-marker*, (Fairhead *et al.*, 1996). La idea para el desarrollo de estos vectores está basada en la observación experimental de que las moléculas de DNA transformante interaccionan entre sí, en sus secuencias homólogas, antes de integrarse en el cromosoma de la levadura.

Los vectores desarrollados funcionan en parejas y cada uno de los miembros contiene sólo una parte del marcador de selección con una zona solapante respecto del fragmento de marcador que porta su pareja. De esta forma en el proceso de cotransformación se produce recombinación entre estas zonas homólogas y se recupera la funcionalidad del marcador (Figura 6). La serie de vectores desarrollados presenta varias opciones de marcador, URA3, LYS2, KAN, y también hay plásmidos necesariamente integrativos, carentes de secuencias de origen de replicación en levaduras, y replicativos que pueden utilizarse para rescatar genes de otras líneas de levaduras mediante la técnica de *gap-repair*. Además los vectores integrativos contienen sitios I-*Sce* I que permiten la excisión de los marcadores de selección.

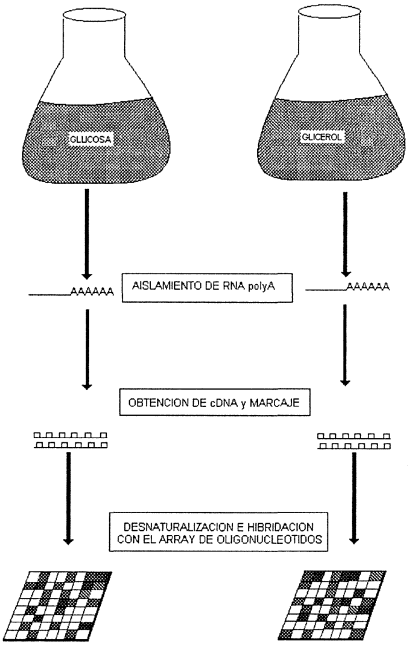

Figura 6.- Aplicación de la técnica de hibridación múltiple a los análisis de expresión.

La estrategia de disrupción en masa, o *mass-murder*, consiste en eliminar una región amplia de ORFs contiguas (generalmente de 2 a 10) sustituyendo cada uno de los cassettes de reemplazamiento de la pareja, *stuffers*, por las zonas adyacentes a las ORFs extremas de la región a interrumpir. Esto es posible ya que las ORFs de función desconocida no parecen estar distribuidas al azar en el genoma, sino que se ve una tendencia a encontrarlas agrupadas. Las primeras experiencias de disrupción génica por el método individual han revelado que muchas disrupciones no llevan emparejado un fenotipo fácilmente reconocible. La confirmación de cuál es la ORF que está causando el fenotipo, entre todas las que han sido simultáneamente delecionadas, se realiza por transcomplementación utilizando la ORF salvaje obtenida mediante *gap-repair*. (Fairhead *et al.*, 1996).

Acudir a esta nueva técnica tiene una doble ventaja ya que un análisis jerárquico bien organizado podría ofrecer una reducción de hasta la mitad del número de líneas mutantes por disrupción que es necesario construir y analizar. Concretamente se está ya aplicando a las ORFs de función desconocida del cromosoma XI y la previsión, si se extendiese su aplicación al total de ORFs desconocidas del genoma de levadura, es que permitiría una exploración preliminar de todos los genes huérfanos con tan sólo 1000 mutantes. Otra ventaja es la detección de un posible fenotipo causado por interacción de los productos de dos genes.

16. *HYBRIDIZATION ARRAY*. DE LA HIBRIDACIÓN MÚLTIPLE AL CHIP

El grupo del Profesor Hoheisel (Heidelberg, Alemania) viene desarrollando desde hace algunos años el sistema de hibridación múltiple que tiene aplicaciones muy diversas, entre ellas la rápida elaboración de mapas físicos y la reducción de redundancia en las estrategias de secuenciación (Hoheisel, 1994). En esta técnica varios clones son comprobados simultáneamente respecto a su respuesta a una sonda y por tanto la idea básica no está conceptualmente muy lejos de otras técnicas ampliamente utilizadas como pueden ser la hibridación de colonias, el *dot* y *slot-blot*. Si damos la vuelta a la idea podemos cambiar las cosas de sitio y poner ahora las sondas, oligonucleótidos u oligopéptidos, ancladas en un soporte, así nace el concepto de *hybridisation array*. Esta técnica ha potenciado el desarrollo de la secuenciación por hibridación (SBH), y de nuevas estrategias para el mapeo de epítopos (Fodor *et al.*, 1993) o mapeo genómico de librerías. Como veremos, tiene también aplicaciones en la elaboración de mapas transcripcionales, una técnica que se está desarrollando dentro del marco de EUROFAN.

El principio básico de todas estas aplicaciones consiste en fabricar una pequeña placa matriz o plantilla que contiene físicamente enlazados una serie de oligonucleótidos que funcionan como sondas. Los oligonucleótidos pueden ser sintetizados *in situ* sobre la propia plantilla, por tanto la escala de síntesis es muy baja. El sistema representa básicamente dos grandes ventajas: ahorro de tiempo,

ya que se puede obtener rápidamente información sobre el elevado número de sondas que se pueden incluir en una sola plantilla, y también ahorro en costes.

Esta técnica puede ser aplicada a la secuenciación. Aunque la idea básica de secuenciación es algo más antigua, el primer fragmento de DNA secuenciado por esta técnica está fechado en 1991 (Strezoska *et al.*, 1991). Desde el punto de vista teórico la técnica de secuenciación por hibridación está basada en la comparación, mediante hibridación, de la secuencia que va a ser secuenciada con una plantilla de oligonucleótidos que contiene todas las variaciones posibles de secuencias de oligonucleótidos de un determinado tamaño, generalmente octámeros.

Para optimizar este sistema es necesario recurrir a la hibridación desarrollada por el grupo de Stephen Fodor (Fodor *et al.*, 1991; Pease *et al.*, 1994). La técnica permite obtener una plantilla miniaturizada. El método empleado utiliza la luz para dirigir la síntesis química a determinadas zonas de un soporte sólido. Esta técnica está basada por tanto en otras ya ampliamente desarrolladas en otros campos de la ciencia, la fotolitografía de semiconductores y la síntesis química en fase sólida. El primer paso consiste en la unión de los primeros reactantes a un soporte sólido y en la protección de los grupos reactivos de éstos mediante un agente que pueda ser eliminado al incidir la luz sobre el (susceptible de fotodesprotección). La luz se dirige a través de una máscara fotolitográfica hacia un área específica del soporte con la resultante desprotección de esa área que es ahora puesta en contacto con el primer grupo que se incorpora. Al repetir este proceso un determinado número de ciclos, se van desprotegiendo distintas zonas y se van consiguiendo síntesis dirigidas en las distintas zonas delimitadas por la plantilla fotolitográfica.

Matemáticamente el número de variaciones con repetición de 4 elementos A,T,G,C, tomados de 8 en ocho es de 48 es decir 65.536. Este elevado número podría sin embargo acoplarse en una plantilla cuadrada de 256x256 entradas que opuparía tan sólo 0,64 x 0,64 cm^2 con la resolución fotolitográfica actual. Si suponemos ahora que queremos secuenciar un fragmento de 1.000 pb el número de hibridaciones positivas que esperamos obtener con nuestra plantilla, en el supuesto de que este fragmento no contuviese repeticiones internas, es de 1000-8+1= 993. Si tenemos un programa capaz de ordenarnos todas estas subsecuencias tendremos la secuencia del inserto.

Los tres principales problemas que es necesario superar y optimizar para conseguir el adecuado funcionamiento de esta técnica son:

1.- Se requiere tener mapas físicos desarrollados hasta el nivel de cada extensión de lectura, es decir de aproximadamente 1 Kb.

2.- Es necesario conseguir que la estabilidad del híbrido formado entre el molde que se está secuenciando y cada uno de los oligos sea la misma, independientemente de la secuencia específica de cada oligo. Este problema está intentando ser subsanado en el equipo de Hoheisel mediante la utilización de nucelótidos modificados en los que el enlace fosfodiester del esqueleto azucar-fosfato es sustituido por un enlace de tipo peptídico de aquí que este tipo de oligos modificados reciban el

nombre de PNA (P, de péptido) en oposición a los oligos DNA (D de desoxirribosa).

3.- La existencia de regiones de repetición interna dentro del fragmento que se está secuenciando origina lecturas cortas. Este problema desaparece si la secuencia es en parte conocida. Por este motivo la técnica de SBH parece que en el futuro tendrá más aplicación en la secuenciación de regiones mayoritariamente conocidas, como puede ser el caso de la secuenciación utilizada como herramienta para detectar determinadas mutaciones en relación con un diagnóstico clínico o en los estudios de variabilidad genética.

La técnica de *hybridisation array* puede ser también aplicada a la elaboración de mapas de transcripción (Figura 7). Supongamos que queremos conocer los niveles de transcripción de 1.000 ORFs de función desconocida y cómo su expresión varía en diferentes condiciones (diferentes tejidos, diferentes condiciones de cultivo, diferentes etapas del desarrollo *etc.*). El procedimiento básico consistiría en preparar una plantilla o *array* de oligonucleótidos de unos 16 nt de longitud y que fuesen específicos para cada una de las sondas. Por otra parte aislaríamos mRNA de las diferentes condiciones a analizar. A partir del mRNA obtendremos el cDNA correspondiente, incorporando en la síntesis el marcaje adecuado. Una hibridación de plantillas idénticas con cada uno de los cDNAs y su posterior cuantificación nos llevaría a obtener información completa en un corto espacio de tiempo. Dentro del consorcio B2, de elaboración de mapas de transcripción para las ORFs de función desconocida del genoma de la levadura, estamos intentando optimizar este protocolo para aplicarlo en la fase II del proyecto EUROFAN. Este estudio se lleva a cabo en colaboración con Hoheisel, participante en el nodo B9 de desarrollo de nuevas tecnologías para el análisis de genomas.

17. REFERENCIAS

Amberg, D., Botstein, D. and Beasley, E. (1995). Precise gene disruption in *Saccharomyces cerevisiae* by Double Fusion Polymerase Chain reaction. *Yeast* 11: 1275-1280.

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

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

Bendixen, C., Ganglof, S., and Rothstein, R. (1994). A yeast mating selection scheme for detection of protein-protein interactions. *Nucl. Acids Res.* 22: 1778-1779.

Brenner, S. E. (1996). BLAST, Blitz, BLOCKS and BEAUTY: sequence comparison on the Net. *Trends Genet.* 11: 230-231.

Cubitt, A. B., Heim, R., Adams, S. R., Boyd, A. E., Gross, L. A. and Tsien, R. Y. (1995). Understanding, improving and using green fluorescent proteins. *Trends Biochem. Sci.* 20: 448-455.

Chien, C-T., Bartel, P.L., Sternglanz, R., and Fields, S. (1991). The two hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad Sci. USA*. 88: 9578-9582.

Dang, V.-D., Valens, M., Bolotin-Fukuhara,M. and Daing-Forner, B. (1994). A genetic screen to isolate genes regulated by yeast CCAT-box binding protein Hap2p. *Yeast* 10: 1273-1283.

Docherty , K. and Clark, A.R. (1993). Transcription of exogenous genes in mammalian cells. *Gen transcription a practical approach.* Hames, B.D. and Higgings S.J. (eds.). pp. 65-123. IRL Press London.

Fairhead, C., Llorente, B., Denis, F., Soler, M., Thierry, A. and Dujon, B. (1996). Combinatorial Deletions of yeast genes: Mass-murder of chromosome XI ORFs. *EUROFAN First Meeting. Louvain la Neuve,* March 28-31.

Fickett, J. W. (1996) Finding genes by computer, the state of the art. *Trends Genet.* 12: 316-320.

Fields, S. and Song, O-K. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340: 245-246.

Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., Solas, D. (1991). Light-Directed spatially addressable parallel chemical synthesis. *Science* 251: 767-773.

Fodor, S. P. A., Rava, R. P., Huang, X. C., Pease, A.C., Holmes, C. P. and Adams, C. L. (1993). Multiplexed biochemical assays with biological chips. *Nature* 364: 555-556.

Hasty, P. and Bradley, A. (1993). Gene targeting vectors for mammalian cells. En *Gene targeting a practical approach*. Ed. A. L. Joyner. IRL Press. Oxford.

Hoheisel, J. D. 1994. Application of hybridisation techniques to genome mapping and sequencing. *Trends Genet* 10: 79-83.

Maftahi , M., Gaillardini, C. and Nicaud, J-M. (1996). Sticky-end polymerase chain reaction method for systematic gene disruption in *Saccharomyces cerevisiae. Yeast* 12: 859-868.

Marcus, G.A., Silverman, N., Berger, S., Horiuchi, J., and Guarente, L. (1994). Functional similarity and physical association between GCN5 and ADA2: putative transcriptional adaptors. *EMBO J.* 13: 4807-4815.

Nakai, K. and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14: 897-911.

Niedenthal, R. K., Riles, L., Johnston, M. and Hegemann, J. H. (1996). Green fluorescent protein as a marker for gene expression and subcellular localization in budding yeast. *Yeast* 12: 773-786.

Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M.T., Holmes, C. P. and Fodor, S. P. A. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* 91: 5022-5026.

Rickles, R.J., Botfield, M.C. Weng, Z., Taylor, J.A. Green, O.M. Brugge, J.S. and Zoller, M.J. (1994). Identification of Src, Fyn, Lyn, PI3K and Abl SH3 domain ligands using phage display libraries. *EMBO J.* 13: 5598-5604.

Roca, J. Gartemberg, M.R. Oshima, Y. and Wang J.C. (1992). A hit and run system for targeted genetic manipulations in yeast. *Nucl. Acids Res.* 17: 4671-4672.

Rotstein, R. (1991). Targeting disruption, replacement and allele rescue: integrative DNA transformation in Yeast. *Methods in Enzymol* 194: 281-301.

Schneider, B.L. Steiner, B., Seufert, W. and Futcher, A.B. (1996). pMY-ZAP: A reusable polymerase chain reaction-directed gene disruption cassette for *Saccharomyces cerevisiae. Yeast* 12: 129-134.

Sparks, A.B., Hoffman, N.G. Mcconnell, S.J. Fowlkes, D.M. and Kay, B.K. (1996). Cloning of ligand targets: systematic isolation of SH3 domain-containing proteins. *Nature Biotechnology* 14: 741-744.

Strezoska, Z., Paunesku, T., Radosavljevic, D., Labat, i., Drmanac, R. and Crkvenjakov, R., 1991. DNA sequencing by hybridization: 100 bases read by a non-gel-based method. *Proc. Nat. Acad. Sci.* 88: 10089-10093.

Verner and Schatz, (1988). Protein translocation across membranes. *Science* 241: 1307-1313.

Wach, A., Brachat, A., Pöhlmann, R. and Philippsen, P. (1994). New Heterologous modules for classical or PCR-based Gene disruptions in *Saccharomyces cerevisiae. Yeast* 10: 1793-1808.