PROGRAMAS COMPUTACIONALES APLICADOS A SECUENCIAS EN LA BIOLOGIA

TESIS DE LICENCIATURA

EN INVESTIGACION BIOMEDICA BASICA

por Jaime Lagunez Otero
Instituto de Investigaciones Biomédicas

Universidad Nacional Autónoma de México
Enero de 1983

PROGRAMAS COMPUTACIONALES APLICADOS A SECUENCIAS EN LA BIOLOGIA

OBJETIVOS:

1) Considerar ideas sobre nuevos métodos para análisis de secuencias.

2) Presentar los programas sobre secuencias de macromoléculas y de conducta desarrollados aquí.

a) Generación de secuencias
b) Estudio de secuencias

3) Presentar resultados obtenidos con los programas.

INDICE / SUMARIO

I. Introducción

Aplicación de las computadoras a la Biología. Las secuencias de macromoléculas y de eventos conductuales se pueden tratar en forma similar.

A. Conducta. Metodos tradicionales no son adecuados para buscar patrones comunes entre diferentes secuencias de conducta.

B. Biología Molecular. El metabolismo, la ingeniería genetica y los programas computacionales.

Características de la microcomputadora utilizada.

II. Programas para el análisis de secuencias de conducta exploratoria en ratones.

Presentación de los metodos utilizados para traducir series de eventos conductuales a secuencias de carácteres.

A) Caracteristicas de la conducta.

B) La obtención de secuencias repetidas permite el estudio posterior sobre su relacion con factores geneticos y emocionales.

C) El problema de la representación de información relevante.

1) Definición de unidades conductuales.
2) Consideraciones sobre coordenadas espacio-temporales en la representación de la información.

D) Utilizacion de programas para busqueda de homologias.

E) Resultados obtenidos.

III. Programas para el análisis de A.D.N. y proteina; aplicación a problemas especificos.

A) Busqueda de secuencias reconocidas por enzimas de restricción en el genoma de Klebsiella pneumoniae.

Programa utilizado:

BUSEC: Búsqueda de secuencias cortas dentro de secuencias mayores definidas por el investigador.

CRADNA: Creacion de archivos para secuencias en forma de tripletes.

B) Busqueda de zonas participantes en la regulación de glutamino sintetasa en Escherichia coli.

Programas utilizados:

COMP: Obtención de cadena complementaria de A.D.N.

TRAD: Traducción de A.D.N. a proteína en las tres fases posibles.

BUPAL: Busqueda de secuencias invertidas repetidas (palindromes).

C) Comparación entre cuatro secuencias de Glutamato Deshidrogenasa Nadph dependientes y determinación de distancias mutacionales.

Programas utilizados:

HOMOLOG: Busqueda de homologia entre secuencias largas de A.D.N., entre proteínas o entre secuencias conductuales.

GRAFPOL: Graficacion de las caracteristicas químicas mas relevantes de los aminoácidos de una proteina.

MMD: Determinacion de distancias evolutivas entre secuencias peptídicas.

IV. Discusiones

Mejorías para los programas. Posibilidades para el futuro.

Referencias

Indice de figuras

repetidas invertidas en un segmento de mas de 80 pares de bases.

3.2.6 Compaginación de secuencias palindrómicas en la región de control de glnA en E. Coli.

3.2.7 Dos posibles estructuras secundarias encontradas por BUPAL en la region de control de glnA (bases 565 a 685).

3.3.1 Coincidencias dentro de 80 elementos de GDH nadph dependientes de E.coli, N.crassa, pollo, y bovino.

3.3.2 Comparación absoluta entre las diferentes secuencias. A. E.coli y N.crassa, B. Bovino y N.crassa, C. Bovino y E.coli.

3.3.3 Tabla de correspondencia entre tripletes y aminoacidos usada para determinar diferencias de una base.

3.3.4 Matriz mostrando los aminoacidos que difieren en solo una base.

3.3.5 Comparacion entre N.crasa y E.coli tomando en cuenta el cambio de una base.

3.3.6 Histogramas de hidrofobicidad/drofilicidad de las enzimas GDH nadph dependientes de E.coli, N.crassa, Pollo, y Bovino. Cerca de la posición 60 se observa un pico de acidez.

3.3.7 Comparación entre E.coli y N.crassa tomando en cuenta polaridad de los aminoacidos.

3.3.8 Matriz indicando distancias minimas mutacionales entre los diferentes aminoacidos.

3.3.9 Distancias minimas mutacionales entre las cuatro secuencias.

3.3.10 Dendrograma obtenido a partir de de las distancias mutacionales.

4.1 Dos ejemplos de secuencias conductuales representadas en forma de histogramas.

4.2 Reconcimiento de patrones usando la tecnica de programacion dinamica.

INTRODUCCION

El uso de sistemas computacionales en la biología ha crecido de manera explosiva en los ultimos anos debido principalmente a la disminución de costos de las unidades procesadoras digitales. Esto ha tenido lugar por la creacion de tecnologia para alta concentracion de circuitos electronicos. Asimismo los investigadores se han visto obligados a utilizar las mas avanzadas tecnicas matemáticas y computacionales para resolver problemas biológicos complejos como son los ecologicos, los regulatorios, y los de diferenciacion celular.

En este trabajo se revisan dos aplicaciones biologicas: A. en el estudio de conducta y B. en la biologia molecular.

A) Conducta

En la etología participan varias disciplinas, entre ellas la bioquimica, la fisiología, y la psicología. Sin embargo no se pueden utilizar los mismos métodos de adquisición de información de estas ciencias. Requiere de técnicas creadas ad-hoc para ella. Para encontrar la correlacion entre la expresion conductual con los fenomenos fisiologicos internos se ha visto que los metodos ya utilizados no son lo suficientemente informativos (1,2,3). Se han usado diversos 'indices' para encontrar la relacion entre los estados emocionales/fisiologicos y la conducta observada. Estos indices son fundamentalmente de tipo cuantitativo en forma de frecuencias con que aparece un evento dado o la transición de un evento a otro. Como se ve en practicamente cualquier conducta como el cortejo, la definición de territorios o quizas la exploración de una medio extrano, la secuencia de eventos con que suceden es muy importante para determinar su significado. Las pautas conductuales aisladas no son suficientes para obtener un significado satisfactorio que refleje el funcionamiento interno de la caja negra que es el sistema nervioso central. Ejemplos de las tecnicas que deshebran las secuencias conductuales son (5):

1. Procesos de Markov
2. Escalamiento Multidimensional
3. Análisis de cúmulos
4. Teoría de la información

Los metodos que emplean el análisis de cúmulos y los tratamientos Markovianos enfocan el problema desde el punto de vista de la probabilidad de suscitarse una transicion de un evento a otro. Se encadenan aquellos eventos que se suceden o preceden frecuentemente. Esto es una reconstruccion de una realidad y no un estudio de los hechos como unidades integras. La analogia del problema seria intentar determinar la similitud entre dos cadenas de ADN solo con las frecuencias con que se suceden unas bases a

otras y no con la secuencia en si. Una desventaja en particular del tratamiento de tipo Markoviano es que el numero de datos requeridos crece en forma exponencial conforme se alarga la secuencia en cuestión. Por otra parte es muy probable que la ocurrencia de un evento sea dependiente de la serie anterior de eventos conductuales y no solamente del evento anterior. Esto es contrario a la definicion de un proceso Markoviano. Un sistema como el que se utiliza en este trabajo para detectar homologías entre secuencias macromoleculares mantiene la integridad de la informacion sin los problemas ya mencionados.

Revisando los programas para el analisis de secuencias de bases nucleicas utilizados en la Universidad de Berkel se encuentra uno que genera una matriz con dos secuencias en los ejes X y Y para determinar homologías. Este programa fue el predecesor de el programa HOMOLOG que permite visualizar facilmente zonas homologas entre dos secuencias ya sean proteícas, nucleícas o conductuales. HOMOLOG presenta las secuencias en común en forma diagonal u horizontal. La forma diagonal es un plano cartesiano con cada eje correspondiendo a una secuencia como la matriz del programa original. En las coordenadas determinadas por las posiciones de los elementos se presentan los ementos homologos o blancos si no hubo coincidencia. Esto es diferente al progrma de Berkely ya que este no imprime el elemento explicito sino solo un punto. Con la nueva presentacion se pueden ver secuencias en sentido diagonal crrespondientes a las zonas homólogas (fig 1.1)

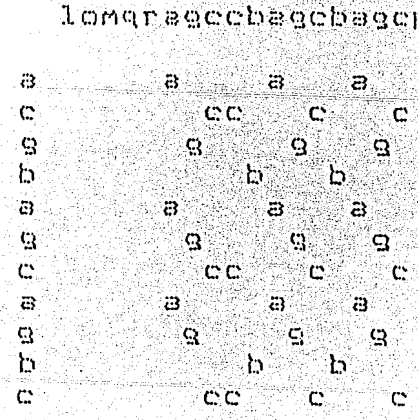


Fig. 1.1 Presentacion diagonal de HOMOLOG.

Variando algunas características como la presentación, dando opciones parametricas como tomar por acierto aquellos elementos que no son identicos pero pertenecen a una misma categoría y acumulando los resultados automaticamente en archivos se pueden generalizar las funciones de este programa y otros similares a todo tipo de secuencias con una

obtencion mayor de informacion.

Aqui vemos la presentacion horizontal que tiene la misma finalidad que la de la fig. 1.1 pero permite ver con mas claridad las zonas homologas. Ademas el programa puede dar una lista de las secuencias similares mas importantes. El programa se aplico tanto a secuencias conductuales como a secuencia macromoleculares (Fig 1.2).

```
BUSQUEDA DE HOMOLOGIA

ABSOLUTO

 A$=EGFFDRGASIVEDKLVEGLRTPDSKEQPRHRVRGILRIIKPCNHVLSVSFPIKPDDGZWEVIEGYRADHSHQRTPCKGGIRYSLDVSVDEVK
SD$=SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI
1)                  L                       V              R
2)                    S    H                      DD                  G
3)            K       S  Q    R  L       V       DDG   V  GYR  Q     P KGG R    V
4)             L  L                  P            D           S      G
5)                           R      I       V                                S
6)          E                        I       V
```

LA HOMOLOGIA MAXIMA OCURRIO EN LA FASE 3 ,FUE DE 20 COINCIDENCIAS EN UN TOTAL DE 89 ELEMENTOS

EL PORCENTAJE DE SIMILITUD = .224719101

Fig. 1.2 Presentacion horizontal de HOMOLOG generada a base de deslizamientos de una secuencia sobre otra.

B) Biologia Molecular

En el control del metabolismo existen muy diversos mecanismos para obtener la cantidad optima de una cierta sustancia dada la situacion energetica del organismo, su fase reproductiva, y la cantidad de sustrato o sustratos presentes. La celula requiere de diversos mecanismos de control para obtener una adecuada regulacion metabolica (19). Estos mecanismos incluyen a la modificacion de la eficiencia de sintesis de ARN a partir de su cadena complementaria en ADN (transcripcion) la modificacion de la eficiencia de la sintesis proteica a partir del templado de tripletes en ARN (traduccion), la eficiencia de union enzima-sustrato y la velocidad de su separacion. Para determinar con una mayor precision a que nivel se lleva a cabo la regulacion en la celula es importante estudiar la estructura primaria del A.D.N. en la zona de interes , es decir la secuencia de bases que lo componen. Se requiere determinar la localizacion de los sitios de union de proteinas activadoras, represoras de sintesis de ARN (ARN

polimerasa), y los sitios de union de ribosoma necesarios para la sintesis de proteina. Estos sitios son identificados como secuencias definidas o como estructuras secundarias (estructuras tridimensionales formadas por la interaccion entre las bases). Por otra parte existen sistemas mas complejos que requieren de una regulacion a nivel de transcripcion y traduccion simultaneamente (26). El caso mejor conocido es el de "atenuacion" en el cual se han identificado dos caracteristicas importante a nivel estructural de A.D.N. Estas son a) la formacion de de estructuras secundarias alternativas en una region anterior al gene que codifica para una enzima biosintetica y b) la presencia de codones en tandem para el aminoacido correspondiente en fase con un peptido hipotetico tambien localizado antes del gene estructural. En la deteccion de estas posibles estructuras es donde intervienen los programas creados.

La ingenieria genetica es una rama de la biologia molecular que reune varias metodologias tales como la clonacion molecular y la secuenciacion de A.D.N. con el fin de entender los sistemas metabolicos celulares y de obtener importantes productos biologicos como son el interferon y la insulina (10,23). Con el advenimiento de ella se hizo posible la lectura de todo tipo de informacion almacenada en las secuencias de A.D.N. obtenidas con tecnicas de laboratorio cada vez mas eficientes (14,15). La computadora acelera este proceso considerablemente Una herramienta clave de la ingenieria genetica son las enzimas de restriccion. Estas son proteinas catalizadoras de la ruptura de ADN de doble cadena (25). Los sitios de corte son comunmente secuencias (repetidas invertidas) denominadas palindromicas de entre 4 y 8 pares de bases (EJ. ATGCA--TGCAT). Debido a secuencias mas largas de este tipo se pueden formar gazas en el espacio al aparearse internamente las bases. Las estructuras pueden darse tanto en ADN como en ARN. La localizacion de los sitios de reconocimiento para diversas enzimas de restriccion reviste importancia cuando el objetivo es una caracterizacion final de alguna region del A.D.N. para facilitar su posterior manipulacion. Dado el gran numero de diferentes enzimas de restriccion descritas a la fecha se hace indispensable la computadora para realizar la tarea repetitiva de buscar los sitios de corte en las secuencias. Otros trabajos igualmente adecuados para la computadora son la cuenta de secuencias de alta repeticion (de interes actual); la determinacion de contenido de A-T/G-C; y la traduccion y uso de tripletes (11).

El programa BUSEC puede encontrar secuencias definidas dentro de secuencias largas ya sean iniciadores que son sitios de union de proteinas activadoras, operadores que son sitios de union a proteinas represoras o bien secuencias reconocidas por enzimas de restriccion. Para detectar estructuras secundarias se requiere el programa BUFAL que

encuentra secuencias repetidas invertidas. El programa HOMOLOG sirve para detectar zonas similares con el mismo sistema de regulacion. Esto implica una relacion evolutiva o funcional cercana.

Objetivos especificos de la tesis.

En el caso del estudio de la conducta el problema principal estriba en traducir los hechos observados a una secuencia de caracteres. La segunda seccion explica como se logro esta traduccion en el estudio de la conducta exploratoria de ratones en un medio extrano y los resultados obtenidos de su analisis. La tercera presenta algunos de los proyectos de biologia molecular en que se requirieron los programas ya mencionados y como fueron desarrollados. Se tiene ademas un apendice con los listados de los programas y algunos de los datos utilizados.

En el trabajo se presentan los ocho programas: BUSEC, BUPAL, COMP, CRADNA HOMOLOG , MMD, TDA Y TRAD. Las funciones de algunos de los relacionados con biomoleculas se encuentran en la literatura. Ya que se desarrollaron con las funciones como guia y no los algoritmos se puede afirmar que son unicos tanto en metodo como en presentacion.

Datos Tecnicos:

Se utilizo una microcomputadora APPLE II con microprocesador 6502 y 48 K de memoria. Como memoria secundaria se usaron diskettes Floppy de 5.75''. Todos los programas se escribieron en APPLESOFT BASIC.

II. TRADUCCION DE INFORMACION CONDUCTUAL OBSERVADA A SECUENCIAS DE SIMBOLOS

1.Antecedentes

La conducta queda definida por series de eventos motores llamados unidades conductuales o pautas que se presentan en forma de secuencias. Existen series que se presentan con regularidad y con cierta variacion. Esta idea es paralela a la de melodia musical (6). Al conjunto de reglas que generan a las melodias se les ha considerado equivalentes a reglas gramaticales o de logica interna (7). Cambios cualitativos en estas reglas podrian reflejar cambios en los estados fisiologicos de los individuos o diferencias geneticas entre subespecies o grupos. En forma de grafica de estados se podria representar este concepto de la manera siguiente (Fig 2.1)

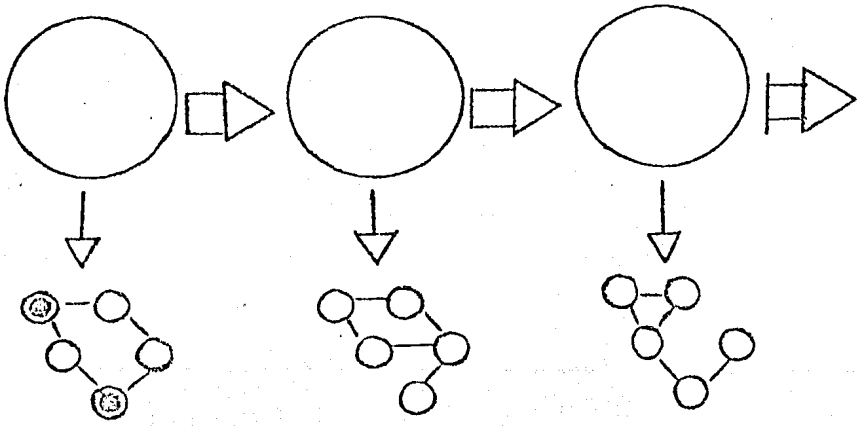


Fig. 2.1 Diagramas de estados en conducta.

En este diagrama cada nodo corresponde a un estado fisiologico modificado por el origen de la cepa bajo estudio y/o efectos de farmacos. Las flechas horizontales en este caso representan la transicion de un estado a otro. Cada nodo se puede expander en un subsistema cuyos nodos corresponden a pautas conductuales y las lineas tambien representan transiciones. Ya que existe una correspondencia entre estados y subsistemas, al identificar a los subsistemas o a las reglas que definen a los subsistemas, debemos poder determinar el estado fisiologico correspondiente.

El resultado del analisis que tome en cuenta estos conceptos es mucho mas informativo que un estudio de frecuencias simples. Motivados por la idea de desarrollar un instrumento para detectar secuencias comunes de comportamiento en ratones (representantes de los

subsistemas) se desarrollaron los programas TDA y SDR. La funcion del primero es traducir la actividad del raton a una secuencia de caracteres. El segundo es una variante de HOMOLOG que construye un archivo para cada una de las secuencias encontradas dentro de dos o mas secuencias exploratorias completas. De esta manera se tiene un registro detallado de los resultados de cada uno de los estudios. En los archivos se tiene la secuencia, en cuantas y cuales observaciones se presento y en que posiciones dentro de las secuencias.

Al analizar un numero considerable de secuencias representativas de los movientos podriamos encontrar un patron tipico exploratorio. Toda desviacion de este patron podria tener un significado fisiologico.

2.Diseno experimental

Se estudiaron ratones explorando un medio novedoso y como control positivo ratones bajo el efecto de la droga Diazepam (5 mg/kg) . Se utilizaron datos obtenidos de la observacion de la conducta exploratoria de ratones albinos (BALB-c) machos durante 15 minutos en un medio clasico de estudio: el 'campo abierto'. Esto es una caja plana de madera sin objetos adicionales, pintada negra de 60cm. por 60 cm dividida en 36 cuadro de 10 cm. por 10 cm.(21) (Fig 2.1) La caja era lavada con amoniaco antes de cada observacion. Como prueba adicional sobre fuente de variacion, despues de diez minutos de exploracion a todos se les expuso a un ruido aversivo con duracion de 2 segundos.

NUMERACION DE LOS CUADROS EN CAMPO ABIERTO

II	1	I	2	I	3	I	4	I	5	I	6	II
II	7	I	8	I	9	I	10	I	11	I	12	II
II	13	I	14	I	15	I	16	I	17	I	18	II
II	19	I	20	I	21	I	22	I	23	I	24	II
II	25	I	26	I	27	I	28	I	29	I	30	II
II	31	I	32	I	33	I	34	I	35	I	36	II

Fig.2.2 Numeracion en campo abierto.

Recordando las secuencias homologas de la figura 1.2 generadas por el programa HOMOLOG a base de deslizamientos,

se ve que el problema relevante a este estudio consiste en representar la conducta observada en forma de secuencias de caracteres que seran comparadas con el programa.

El primer paso consiste en definir los criterios para determinar una unidad conductual dentro de una secuencia. Para ello se debe tomar en cuenta los siguientes puntos:

a. Actividad realizada
b. Coordenada espacial (localizacion)
c. Coordenada temporal (duracion)

a. Actividad realizada

Estas son las cinco actividades basicas:

1. aseo
2. desplazamiento
3. exploracion
4. congelamiento
5. escape

Un aseo incluye el lavar, lamer, rascar, y la agitacion de la cabeza, patas, cuerpo, genitales, y/o cola,. Interrupciones del aseo por pausas con duracion menor a cuatro segundos determinan una nueva pauta. Los desplazamientos se les consideran como tales cuando no existen interrupciones para explorar o asearse. La exploracion se determina cuando el raton husmea una cierta zona. El congelamiento ocurre frecuentemente despues del ruido aversivo y su duracion minima es de cinco segundos pero puede mantenerse en el mismo estado hasta el final de la observacion. Un desplazamiento de mas de dos cuadros por segundo define un escape.

b. Coordenada espacial

Para intentar clasificar zonas equivalentes dentro de la caja y poder extrapolar la informacion obtenida de los campos de exploracion experimentales a otras condiciones se construyeron las siguientes matrices donde se intenta explicar a priori la frecuencia de estancia en estas distintas regiones de el medio.

MATRIZ I

I	0	I	0	I	0	I	1	I	1	I	1	I
I	0	I	0	I	1	I	1	I	1	I	2	I
I	0	I	1	I	1	I	1	I	2	I	2	I
I	1	I	1	I	1	I	2	I	2	I	2	I
I	1	I	1	I	2	I	2	I	2	I	3	I
I	1	I	2	I	2	I	2	I	3	I	3	I

MATRIZ II

II	0	I	0	I	2	I	2	I	0	I	0	II
II	0	I	4	I	4	I	4	I	4	I	0	II
II	2	I	4	I	4	I	4	I	4	I	2	II
II	2	I	4	I	4	I	4	I	4	I	2	II
II	0	I	4	I	4	I	4	I	4	I	0	II
II	0	I	0	I	2	I	2	I	0	I	0	II

MATRIZ RESULTANTE

II	0	I	0	I	2	I	3	I	1	I	1	II
II	0	I	4	I	5	I	5	I	5	I	5	II
II	2	I	5	I	5	I	5	I	5	I	2	II
II	3	I	5	I	5	I	6	I	6	I	4	II
II	1	I	5	I	6	I	6	I	6	I	3	II
II	1	I	2	I	4	I	4	I	3	I	3	II

Fig.2.3 A. Matriz I - Gradiente de distancia con respecto al origen en campo abierto. B. Matriz II - Barreras fisicas en campo abierto. C. Matriz resultante de I y II.

La matriz I es un gradiente discreto con unidades arbitarias de distancia con respecto al punto de partida del raton. La matriz II clasifica la caja en zonas acotadas por dos, una o ninguna pared. La suma punto a punto de las dos genera la tercera matriz.

Consideramos que esta matriz refleja la emocionalidad producida por las diferentes regiones de la zona explorada debido a la lejania del sitio de origen y a la falta de proteccion en forma de barreras. En tercera dimension esta ultima matriz se veria como la siguiente figura. El raton bajo observacion debe escalar montanas de 'tension emocional' para desplazarse.

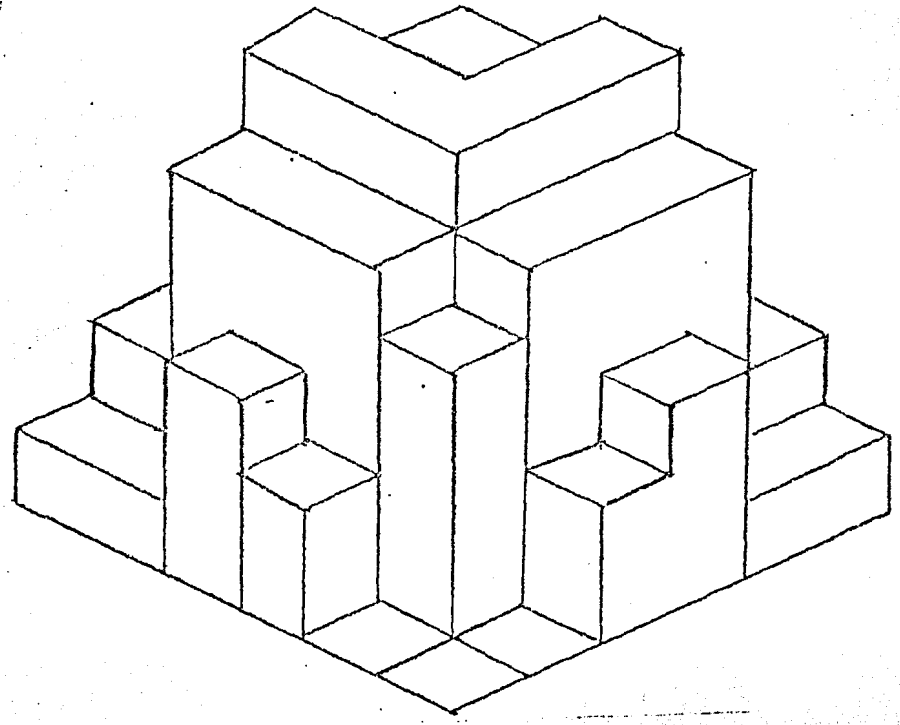


Fig. 2.4 Matriz de 'angustia' en tercera dimension.

Es posible que exista una relacion directa inversa entre el tiempo o frecuencia de permanencia y el la altura de la columna para cada punto de la matriz de tal manera que a en cada punto exista un valor constante producto de esta relacion. Se requerira de un analisis estadistico para determinarlo.

c. Coordenada temporal

Para tomar en cuenta la coordenada temporal, se determina el tiempo de estancia en una zona y se divide entre 3 segundos. El resultado determina el numero de veces que aparece ese estado de manera continua en la secuencia. De este modo, al contrario de de otros estudios, nuestra gramatica permite pasar de un estado a si mismo.

Por otra parte la posicion dentro de la secuencia

corresponde a la aparicion cronologica del evento. Al realizar deslizamientos razonables, es decir cronologicamente equivalentes, de una secuencia sobre la otra se esta tomando en cuenta esta coordenada.

La siguiente tarea es la de definir los caracteres a usar por la actividad y la zona donde se realiza. La figura 2.5 muestra las posibles eventos generales que ocurren dentro de una conducta exploratoria con sus caracteres correspondientes.

CODIGO

@-->F ==>> desplazamiento en zonas 0 a 6
G-->M ==>> exploracion en zonas 0 a 6
N ==>> aseo
O ==>> timbre
P ==>> escape
Q ==>> congelamiento

Fig. 2.5 Codigo utilizado para traducir.

El algoritmo en TDA para realizar la traduccion consiste en los siguientes pasos.

I. Leer el elemento en turno -->>

II. Determinar si corresponde a un cambio de actividad -->>

III. Definir duracion de actividad anterior leyendo los minutos y segundos, en caso de ser un cambio de actividad-->>

IV. Determinar zona de 'ansiedad' en la matriz a la cual corresponde el cuadro leido.

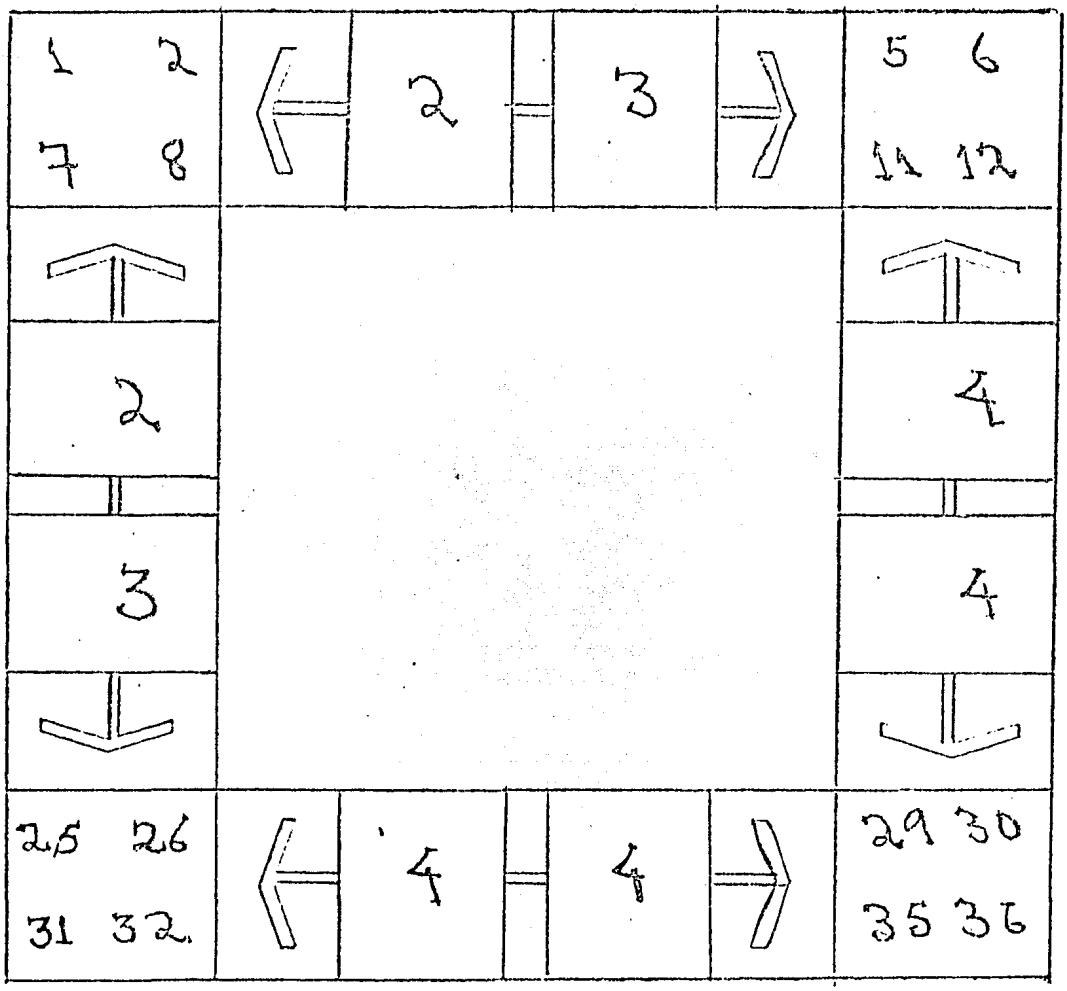
V. Determinar si se ha cruzado una zona no anotada por el observador y en dado caso agregar esta informacion a la secuencia en produccion.

VI. Acumular cadena de eventos y salir si se ha terminado la secuencia.

VII. Regresar al punto I.

Para determinar si se ha cruzado por alguna zona no registrada al introducir los datos se utilizo la siguiente tabla. En las esquinas se se encuentran los numeros de los cuadros encontrados en la informacion sin procesar. Si ocurre un desplazamiento de

Fig. 2.6 Tabla utilizada para determinar las zonas por las que hubo desplazamiento.



esquina a esquina se habra cruzado por las zonas indicadas en la tabla y se concatena con la secuencia en produccion. En las esquinas del esquema se encuentran los valores de la numeracion del campo abierto. Los numeros que se encuentran entre las esquinas corresponden a los valores de emocionalidad por los cuales el raton probablemente paso para desplazarse de unos a otros cuadros en esquina

Con el codigo ya definido se puede hacer una traduccion de la informacion anotada a lapiz a secuencias de caracteres. La siguiente figura muestra esta etapa.

E,0,0,1,D,0,8,1,4,6,8,E,0,30,1,F,1,02,1

v
v
v

G@B@BBCCBGQ

Fig.2.7 Ejemplo de traduccion.

La primera expresion es la manera en que se registra la informacion. 'E' significa exploracion. Los siguientes dos numeros corresponden al tiempo en minutos y segundos cuando comenzo la actividad y el siguiente numero indica el cuadro o los cuadros en los que se realizo la actividad. En este caso despues de explorar el cuadro uno, el raton se desplazo por los cuadros 1,4,6,y 8. La expresion inferior es la misma informacion pero traducida utilizando los codigos con las zonas clasificadas. Esto es el tipo de secuencia que se puede usar en un programa como HOMOLOG. Nuestra gramatica hipotetica debera contener una aproximacion de la siguiente secuencia como melodia central:

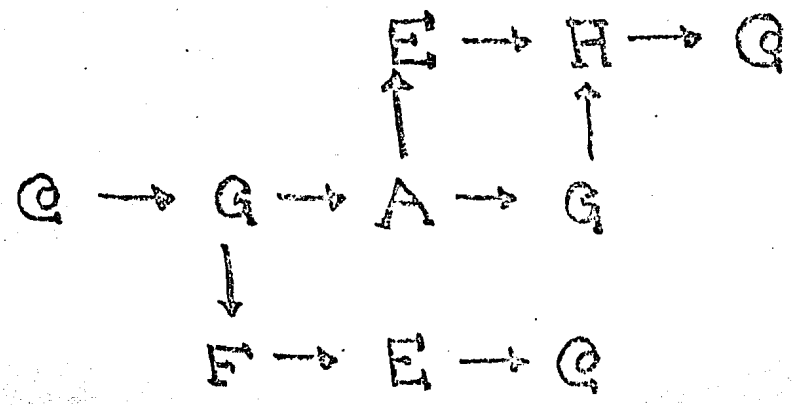
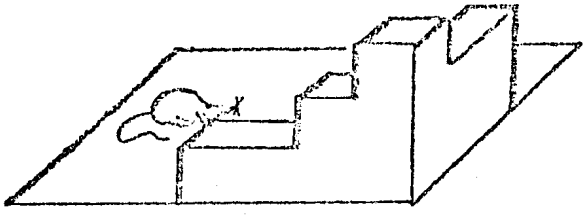


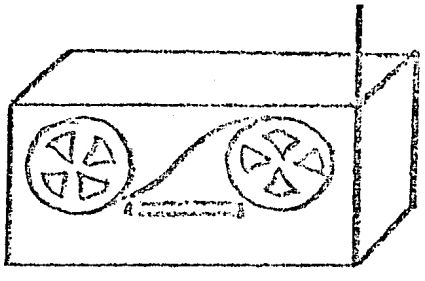
Fig. 2.8 Ciclo basico hipotetico de secuencias de pautas.

A esto se le denomino el ciclo tipico. En una secuencia promedio se debia esperar que se repitiera tres o cuatro veces con valores cada vez mas altos de ansiedad.

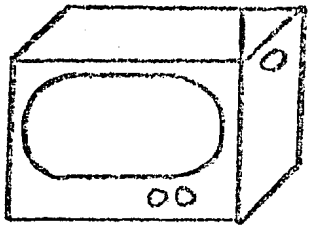
El siguiente diagrama esquematiza el flujo de informacion dentro del sistema montado. Los ultimos tres pasos son los que se corresponden a este trabajo. (Fig 2.9).

OBSERVACION ========>> GRABACION EN VIDEOTAPE ==>>



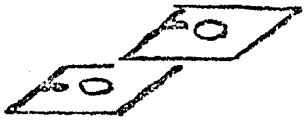


REVISION DE VIDEOTAPE===>>ANOTACION DE PAUTAS GENERALES==>>



E,1,2,4,7,6

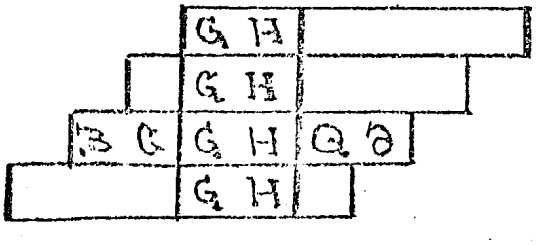
GRABACION DE INF. EN DISCO====>>TRADUCCION A CODIGO ==>>



∂BCGHQ∂A

COMPARACION POR HOMOLOG=========>> GRABACION DE SECUENCIAS

(PAREJA Y POSICION)



3. Resultados

Este proceso se aplico a 14 ratones 7 bajo los efectos de DIAZEPAM y 7 sin la droga. Con HOMOLOG se hicieron las comparaciones de las cadenas conductuales y con los eventos homologos explicitos se crearon los archivos de SDR. El deslizamiento permite una flexibilidad temporal en las homologias encontradas. La restriccion impuesta mas importante fue la de hacer el deslizamiento en ambos sentidos por no mas de siete eventos.

La figura siguiente muestra las secuencias encontradas en dos o mas conductas exploratorias. Encontramos que existen mas secuencias homologas entre los ratones bajo la influencia de la droga que entre los no afectados, un resultado similar a otros donde drogas uniformizan las poblaciones (comunicacion personal Marcela Santis).

La mayoria de las secuencias encontradas no son suficientemente largas o se presentan con frecuencias que no apoyan fuertemente la posibilidad de un patron general exploratorio. Existen dos posibles explicaciones para este resultado preliminar: Las suposiciones requeridas no reflejan al sistema bajo estudio adecuadamente o el patron general no existe. Se requiere ahora realizar estos analisis con mas individuos y en condiciones experimentales variadas. Sera necesario afinar los valores y las categorizaciones hechas antes del estudio. Seria interesante revisar las similitudes dentro de la secuencias homologas ya encontradas usando las mismas tecnicas. Teniendo mas datos experimentales se puede hacer un analisis estadistico minucioso que determine la significancia de cada una de ellas.

Con respecto a la matriz tridimensional existe la posibilidad de que a cada nueva conquista se crea un nuevo sitio de iniciacion y posiblemente un nuevo gradiente a partir de ese punto. En estudios futuros sera necesario revisar esta posibilidad.

II APLICACIONES DE PROGRAMAS PARA EL ANALISIS DE SECUENCIAS EN MACROMOLECULAS

A) Busqueda de secuencias de enzimas de restriccion en el genoma de Klebsiella pneumoniae.

En estudios recientes (16,20,22) se ha hecho un analisis de la homologia entre los genes involucrados en la fijacion de nitrogeno de la celula (nif) en diferentes bacterias. Aprovechando esta informacion es posible extrapolar datos entre estas especies para posteriormente intervenir activamente en la regulacion del metabolismo y para obtener una mejor comprension de los estos mecanismos.

Al utilizar una enzima de restriccion (25) con A.D.N. como sustrato se generan segmentos con longitudes promedio dependientes del numero de bases reconocidas por la enzima. Estos segmentos se pueden visualizar en una placa de electroforesis en gel de agarosa que separe moleculas por tamano. Con bromuro de etidio se pueden ver las bandas fluorescer en luz ultravioleta. Si previamente se desnaturalizan las cadenas y se renaturalizan con A.D.N. proveniente de otras especies se detectaran homologias por las moleculas hibridas obtenidas. En especifico se ha encontrado una gran homologia entre el genoma bien conocido de Klebsiella pneumoniae y el de Rhizobium phaseoli (22). Se encontro que el plasmido pAC-30 conteniendo un inserto con los genes estructurales KD y H (3.1.1) del operon nif de Klebsiella hibridaba con tres bandas de la digestion total por ECO R1 del genoma de Rhizobium. Esto genero la pregunta si cada una de las bandas corresponde a uno de los genes del plasmido o si un gene de K. pneumoniae estaba hibridando con varios de Rhizobium o sea si existia informacion reiterada. La estrategia mas directa para responder a esta pregunta seria obtener secciones totalmente intragenicas de los genes KD y H, clonarlos en diferentes plasmidos e hibridar cada uno contra el genoma total y observar a que bandas corresponde cada plasmido. Con el proposito de encontrar la combinacion mas adecuada de enzimas de restriccion que produjera las deleciones necesarias se desarrollo el programa BUSEC.

1.	AGCT	ALU I
2.	CTCGGG	AVA I
3.	CTCGAG	AVA I
4.	CCCGGG	AVA I
5.	CCCGAG	AVA I
6.	GGATCC	BAM HI
7.	ATATCT	BGL II
8.	GAATTC	ECO RI
9.	GGCC	HAE III
10.	GCGC	HHA I

11.	AAGCTT	HINd III
12.	GTTAAC	HPA I
13.	CCGG	HPA II
14.	GGTACC	KPN I
15.	CTGCAG	PST I
16.	GTCGAC	SAL I
17.	CCCGGG	SMA I
18.	TCGA	TAQ I
19.	CTCGAG	XHO I
20.	CCGCGG	SST II
21.	TGATCA	BCL I

3.1.2 Secuencias optativas dentro de BUSEC.

BUSEC permite detectar secuencias determinadas por el usuario ya sea por impresion en teclado o por seleccion de alguna de las secuencias optativas presentadas en la introduccion del programa. (Fig. 3.1.2) La deteccion se hace de dos posibles maneras:

A. Homologia absoluta o'
B. Similitud porcentual

El primer metodo utiliza tres pasos principales:

1. Comparar la secuencia buscada contra la analizada elemento por elemento .
2. Almacenar en vector de valores enteros la localizacion de secuencias encontradas.
3. Incrementar apuntador de la secuencia en revision para repetir comparacion.

El segundo algoritmo da flexibilidad a la definicion de secuencias similares. La mejoria consiste en hacer una suma de los elementos atinados y dividir entre el total de elementos para seleccionar aquellas secuencias que tienen una similitud mayor al umbral determinado por el usuario (un cierto porcentaje). Para hacer eficiente el procedimiento en la comparacion de cada elemento se determina el codigo ASCII de los dos elementos a prueba. Se realiza una comparacion booleana de estos valores y se considera acierto si la diferencia es cero o falla si el resultado es diferente de cero. Por ejemplo el valor ASCII de la letra 'A' es 65 y el de la letra 'C' es de 67, la diferencia entre estos dos valores es diferente de cero, por tanto los elementos son diferentes. Si estamos comparando ACTCG contra AGTAG encontramos que el porcentaje de similitud se obtiene dividiendo tres entre cinco, o sea: 60%.

Dado que en el lenguaje BASIC se realiza la compilacion del programa simultaneamente con la ejecucion, el tiempo de procesamiento es relativamente largo. Seria recomendable trancribir este algoritmo a un lenguaje como PASCAL. Que aumenta la velocidad de ejecucion siete a diez veces.

Para detectar una secuencia definida en forma mas eficiente en ocasiones es preferible dar los elementos mas

Fig. 3.1.1 Inserto de K D y H en plasmido pAC-35.

Fig. 3.1.3 Parte de la secuencia del inserto de KDH que se analizo con BUSEC.

```
GAATTCAACGCGTTATGAAGAGAGTCGCCGCGCAGCGCGCCAAGAGATTGCGTGGAATAAGACACAGGGGGCGACAAGCTGTTGAACAGGCGACAAAGCGCCCATGG
                                        -250                                                        -200

CCCCGGCAGGCGCAATTGTTCTGTTTCCCACATTTGGTCGCCTTATTGTGCCGTTTTGTTTTACGTCCTGCGCGGCGACAAATAACTAACTTCATAAAAATCATAAG
                                    -150                                                    -100

                                                                                            1
                                                                                    Met Thr Met Arg Gln Cys Ala Ile
AATACATAAACAGGCACGGCTGGTATGTTCCCTGCACTTCTCTGCTGGCAAACACTCAACAACAGGAGAAGTCACCATG ACC ATG CGT CAA TGC GCT ATT
                            -50                                                             1
                                                20
Tyr Gly Lys Gly Gly Ile Gly Lys Ser Thr Thr Thr Gln Asn Leu Val Ala Ala Leu Ala Glu Met Gly Lys Lys Val Met
TAC GGT AAA GGC GGT ATC GGT AAA TCC ACC ACC ACG CAG AAC CTC GTC GCC GCG CTG GCG GAG ATG GGT AAG AAA GTG ATG
                                                                                                        100
                    40                                                                          60
Ile Val Gly Cys Asp Pro Lys Ala Asp Ser Thr Arg Leu Ile Leu His Ala Lys Ala Gln Asn Thr Ile Met Glu Met Ala
ATC GTC GGC TGC GAT CCG AAG GCG GAC TCC ACC CGT CTG ATT CTG CAC GCC AAA GCA CAG AAC ACC ATT ATG GAG ATG GCC
                                                                    80
Ala Glu Val Gly Ser Val Glu Asp Leu Glu Leu Glu Asp Val Leu Gln Ile Gly Tyr Gly Asp Val Arg Cys Ala Glu Ser
GCG GAA GTC GGC TCG GTC GAG GAC CTC GAA CTC GAA GAC GTG CTG CAA ATT GGC TAC GGC GAT GTG CGC TGC GCG GAA TCC
                        200
                                            100
Gly Gly Pro Glu Pro Gly Val Gly Cys Ala Gly Arg Gly Val Ile Thr Ala Ile Asn Phe Leu Glu Glu Glu Gly Ala Tyr
GGC GGC CCG GAG CCA GGC GTC GGC TGC GCG GGA CGC GGC GTG ATC ACG GCG ATC AAC TTT CTT GAA GAA GAA GGC GCC TAC
                                                300
                    120                                                                         140
Glu Asp Asp Leu Asp Phe Val Phe Tyr Asp Val Leu Gly Asp Val Val Cys Gly Gly Phe Ala Met Pro Ile Arg Glu Asn
GAG GAC GAT CTC GAT TTC GTG TTC TAT GAC GTG CTC GGC GAC GTG GTC TGC GGC GGC TTC GCC ATG CCG ATC CGC GAA AAC
                                                                    400
                                                                160
Lys Ala Gln Glu Ile Tyr Ile Val Cys Ser Gly Glu Met Met Ala Met Tyr Ala Ala Asn Asn Ile Ser Lys Gly Ile Val
AAA GCC CAG GAG ATC TAC ATC GTC TGC TCC GGC GAA ATG ATG GCG ATG TAC GCG GCC AAC AAT ATC TCC AAA GGG ATC GTT
                                                                                            500
                                        180
Lys Tyr Ala Lys Ser Gly Lys Val Arg Leu Gly Gly Leu Ile Cys Asn Ser Arg Gln Thr Asp Arg Glu Asp Glu Leu Ile
AAA TAC GCC AAA TCC GGC AAG GTG CGC CTC GGC GGC CTG ATC TGT AAC TCA CGT CAG ACC GAC CGT GAA GAC GAA CTG ATT
            200                                                                         220
Ile Ala Leu Ala Glu Lys Leu Gly Thr Gln Met Ile His Phe Val Pro Arg Asp Asn Ile Val Gln Arg Ala Glu Ile Arg
ATT GCC CTG GCG GAA AAG CTC GGT ACC CAG ATG ATC CAC TTT GTG CCC CGC GAC AAC ATC GTG CAG CGC GCG GAG ATC CGC
              600
                                                240
Arg Met Thr Val Ile Glu Tyr Asp Pro Ala Cys Lys Gln Ala Asn Glu Tyr Arg Thr Leu Ala Gln Lys Ile Val Asn Asn
CGC ATG ACG GTT ATC GAG TAC GAC CCC GCC TGT AAA CAG GCC AAC GAA TAC CGC ACC CTG GCG CAG AAG ATC GTC AAC AAC
                                        700
                                    260
Thr Met Lys Val Val Pro Thr Pro Cys Thr Met Asp Glu Leu Glu Ser Leu Leu Met Glu Phe Gly Ile Met Glu Glu Glu
ACC ATG AAA GTG GTG CCG ACG CCC TGC ACC ATG GAT GAG CTG GAA TCG CTG CTG ATG GAG TTC GGC ATC ATG GAA GAG GAA
                                                    800
    280                                                                     1
Asp Thr Ser Ile Ile Gly Lys Thr Ala Ala Glu Glu Asn Ala Ala ***                 Met Met Thr Asn Ala Thr Gly
GAC ACC AGC ATC ATT GGC AAA ACC GCC GCC GAA GAA AAC GCG GCC TGA  GCACAGGACAATT  ATG ATG ACC AAC GCA ACG GGC
                                                                                    900
                                        20
Glu Arg Asn Leu Ala Leu Ile Gln Glu Val Leu Glu Val Phe Pro Glu Thr Ala Arg Lys Glu Arg Arg Lys His Met Met
GAA CGT AAT CTG GCG CTG ATC CAG GAA GTC CTG GAG GTG TTC CCG GAA ACC GCG CGA AAA GAG CGC AGA AAG CAC ATG ATG
                    40                                                                          60
Val Ser Asp Pro Lys Met Lys Ser Val Gly Lys Cys Ile Ile Ser Asn Arg Lys Ser Gln Pro Gly Val Met Thr Val Arg
GTC AGC GAT CCG AAA ATG AAG AGC GTC GGC AAG TGC ATT ATC TCT AAC CGC AAA TCA CAA CCC GGC GTA ATG ACC GTA CGC
  1000
                                                                    80
Gly Cys Ala Tyr Ala Gly Ser Lys Gly Val Val Phe Gly Pro Ile Lys Asp Met Ala His Ile Ser His Gly Pro Ala Gly
GGC TGC GCC TAC GCC GGT TCC AAA GGG GTG GTA TTT GGG CCG ATT AAG GAT ATG GCC CAT ATT TCG CAC GGA CCG GCT GGC
                            1100
                                            100
Cys Gly Gln Tyr Ser Arg Ala Glu Arg Arg Asn Tyr Tyr Thr Gly Val Ser Gly Val Asp Ser Phe Gly Thr Leu Asn Phe
TGC GGC CAG TAT TCC CGC GCC GAA CGA CGC AAC TAC TAC ACC GGA GTC AGC GGC GTC GAT AGC TTC GGC ACG CTG AAC TTC
                                                1200
                    120                                                                         140
Thr Ser Asp Phe Gln Glu Arg Asp Ile Val Phe Gly Gly Asp Lys Lys Leu Ser Lys Leu Ile Glu Glu Met Glu Leu Leu
ACC TCT GAT TTT CAG GAG CGC GAC ATC GTC TTC GGC GGC GAT AAA AAG CTC AGC AAG CTG ATT GAA GAG ATG GAG TTG CTG
                                                                1300
                                                            160
Phe Pro Leu Thr Lys Gly Ile Thr Ile Gln Ser Glu Cys Pro Val Gly Leu Ile Gly Asp Asp Ile Ser Ala Val Ala Asn
TTC CCG CTC ACC AAA GGG ATC ACC ATT CAG TCG GAA TGC CCG GTG GGG CTG ATC GGT GAT GAT ATC AGC GCG GTG GCC AAC
                                                                                            1400
                                        180
Ala Ser Ser Lys Ala Leu Asp Lys Pro Val Ile Pro Val Arg Cys Glu Gly Phe Arg Gly Val Ser Gln Ser Leu Gly His
GCC AGC AGC AAG GCG CTG GAT AAA CCG GTG ATC CCG GTA CGC TGC GAA GGC TTT CGC GGC GTG TCG CAG TCT CTG GGG CAC
            200
His Ile Ala Asn Asp Val Val Arg Asp Trp Ile
CAT ATC GCC AAC GAC GTG GTG CGC GAC TGG ATC C
                    1500
```

conservados o mas importantes como secuencia a buscar y posteriormente filtrar del resultado las secuencias no relevantes repitiendo la busqueda con los elementos menos determinantes. Como ejemplo es preferible hacer una busqueda de 'GTAC' y luego filtrar haciendo una busqueda completa sobre las secuencias encontradas que hacer una busqueda del 80% de similitud de 'TGTACA'.

Para almacenar la secuencia ya publicada del inserto KDH (1981) se hizo el programa CRADNA que crea un archivo de tipo acceso aleatorio con registros de 24 caracteres cada uno. Como accesorio el programa facilita la revision y correccion presentano en pantalla los registro de tres en tres bases. En este caso el archivo contiene 76 registros. En el momento de ejecucion de BUSEC se pregunta que archivo se va a revisar y de cuales son los primeros y ultimos registros de interes.

El resultado parcial (Sobre el gene 'H' y parte del 'D') se encuentra en las siguientes paginas. El estudio se realiza en tres pasos: con las primeras siete secuencias a detectar, luego de la ocho a la dieciseis y finalmente de la 17 a la 21. Se pueden ver dos tipos de presentaciones. La primera da los numeros de las bases donde se encuentran cada una de las secuencias, luego el numero de secuencia optativa; la secuencia; el nombre de la enzima que la reconoce; el nombre del archivo que se reviso; y los primer y ultimo registros. En la segunda presentacion se encuentran seis columnas correspondiendo al numero de base donde se encontro; en cual de los setentaiseis registros se encontro; la secuencia del registro correspondiente; el numero de base dentro del registro donde comienza el encuentro,; y finalmente la secuencia y un asterisco si se encontro en sentido contrario en la cadena.

Se encontro posteriormente por otros metodos (18) que efectivamente si existe reiteracion de informacion.

Otras aplicaciones de BUSEC serian determinar el uso de codones y de pares de bases A-T/C-G, y la deteccion de secuencias repetidas

MATRIZ DE COORDENADAS:
517 535 673 922 1003 1104 1524 1529 1596 1605 NO.SEC: 1   SEC.: AGCT   ENZIMA.
LU I   ARCHIVO: AR   1ER REG.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 NO.SEC: 2   SEC.: CTCGGG   ENZIMA: A VA I   ARCHIVO: AR   1
R REG.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 NO.SEC: 3   SEC.: CTCGAG   ENZIMA: A VA I   ARCHIVO: AR   1
R REG.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 NO.SEC: 4   SEC.: CCCGGG   ENZIMA: A VA I   ARCHIVO: AR   1
R REG.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 NO.SEC: 5   SEC.: CCCGAG   ENZIMA: A VA I   ARCHIVO: AR   1
R REG.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 NO.SEC: 6   SEC.: GGAATC   ENZIMA: BAM HI   ARCHIVO: AR
ER REG.: 1   ULT. REG.: 76
754 0 0 0 0 0 0 0 0 0 NO.SEC: 7   SEC.: AGATCT   ENZIMA: BGL II   ARCHIVO: AR
1ER REG.: 1   ULT. REG.: 76

NO. BASE	NO. REG.	SECUENCIA	NO. EN REG.	SEC. ENZ
517	21	AAGTCGGCTCGGTCGACGACCTCG	13	AGCT *
535	22	AACTCGAAGACGTGCTGCAAA	7	AGCT *
673	28	TCGATTTCGTGTTCTATGACGTGC	1	AGCT *
922	38	TGGCGGAAAAGCTCGGTACCCAGA	10	AGCT
1003	41	GCCGCATGACGGTTATCCAGT	19	AGCT *
1104	46	AGCTGGAATCGCTGCTGATGCAGT	0	AGCT
1524	63	ACCGGACTCAGCGGCGTCGATAGC	12	AGCT *
1529	63	ACCGGACTCAGCGGCGTCGATAGC	17	AGCT
1596	66	TTCGGCGGCGATAAAAAGCTCAGC	12	AGCT
1605	66	TTCGGCGGCGATAAAAAGCTCAGC	21	AGCT

754	31	AAGCCCAGGACATCTACATCGTCT	10	AGATCT

MATRIZ DE COORDENADAS: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2? 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ENZIMA: ECO RI   ARCHIVO: AR   1ER REG.:
0 0 0 0 0 0 NO.SEC: 8   SEC.: GAATTC
1   ULT. REG.: 76
94 130 134 348 399 497 583 588 591 772 800 838 858 1025 1188 1270 1369 1402 1426
1441 1460 1472 1509 1704 1735 1744 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 NO.S
EC: 9   SEC.: GGCC   ENZIMA: HAE III   ARCHIVO: AR   1ER REG.: 1   ULT. REG.: 76
33 53 54 59 60 61 122 141 201 202 331 392 393 500 569 607 608 615 650 735 797 84
8 954 971 972 973 974 1050 1185 1240 1278 1279 1280 1290 1386 1393 1484 1485 14
7 1568 1697 1698 1723 1765 1812 1813 NO.SEC: 10   SEC.: GCGC   ENZIMA: HHA I   A
RCHIVO: AR   1ER REG.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ENZIMA: HIND III   ARCHIVO: AR   1ER RE
0 0 0 0 0 0 NO.SEC: 11   SEC.: AAGCCTT
G.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
141 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ENZIMA: HPA I   ARCHIVO: AR   1ER REG.
0 0 0 0 0 0 0 NO.SEC: 12   SEC.: GTTAAC
: 1   ULT. REG.: 76
94 130 134 348 399 497 583 588 591 772 800 838 858 1025 1188 1270 1369 1402 14?
1441 1460 1472 1509 1703 1735 1744 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 NO.S
EC: 13   SEC.: CCGG   ENZIMA: HPA II   ARCHIVO: AR   1ER REG.: 1   ULT. REG.: 76

126 927 1095 1746 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 NO.SEC: 14 SEC.: GGTACC ENZIMA: KPN I ARCHIVO: AR
1ER REG.: 1 ULT. REG.: 76
669 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 NO.SEC: 15 SEC.: CTGCAG ENZIMA: PST I ARCHIVO: AR 1ER REG.
: 1 ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 NO.SEC: 16 SEC.: GTCGAC ENZIMA: SAL I ARCHIVO: AR 1ER REG.:
1 ULT. REG.: 76

NO. BASE	NO. REG.	SECUENCIA	NO. EN REG.	SEC. ENZ
25	1	GAATTCAACGCGTTATGAAGAGAG	1	GAATTC
94	3	TGCGTCGAATAAGACACAGGGGGC	22	GGCC
130	5	AGCGCCCATGGCCCCGGCAGGCGC	10	GGCC
134	5	AGCGCCCATGGCCCCGGCAGGCGC	14	GGCC *
348	14	TTTACCGTAAAGGCCGTATCGGTA	12	GGCC
399	16	TCGTCGCCGCGCTGCCGGAGATGG	15	GGCC *
497	20	ACACCATTATGGAGATGGCCGCGG	17	GGCC
583	24	AATCCGGCGGCCCCGGAGCCAG	7	GGCC *
[illegible]88	24	AATCCGGCGGCCCCGGAGCCAG	12	GGCC
591	24	AATCCGGCGGCCCCGGAGCCAG	15	GGCC *
772	32	GCTCCGGCGAAATGATGGCCATGT	4	GGCC *
800	33	ACGCGGCCAACAATATCTCCA	8	GGCC
838	34	GGATCGTTAAATACGGCCAAATCCG	22	GGCC *
858	35	GCAAGGTGCGCCTCGGCGGCCTGA	18	GGCC
1025	42	ACCCCGCCTGTAAACAGGCCAACG	17	GGCC
1188	49	CCGAAGAAAACGCGGCCCGAGCAC	12	GGCC
1270	52	ATCCAGGAAGTCGTGGAGGTGTTC	22	GGCC *
1369	57	CCACCCGGCGTAATGACCGTACGC	1	GGCC *
1402	58	GGCTGCGCCTACGCCGGTTCCAAA	10	GGCC *
1426	59	GGGGTGGTATTTCGGCCCATTAG	10	GGCC
1441	60	GATATGGCCCATATTTCGCACGGA	1	GGCC
1460	60	GATATGGCCCATATTTCGCACGGA	20	GGCC *
1472	61	CCGGCTGGCTGCGGCCAGTATTCC	8	GGCC
1509	~~62~~	CGCGCCGAACGACGCAACTACTAC	21	GGCC *
1704	71	GCGGTGGCCAACGCCAGCAGCAAC	0	GGCC
1735	72	GCGCTGGATAAACCGGTGATCCCG	7	GGCC *
1744	72	GCGCTGGATAAACCGGTGATCCCG	16	GGCC *
33	1	GAATTCAACGCGTTATGAAGAGAG	9	GCGC *
53	2	TCGCCGCGCAGCGCGCCAAGAGAT	5	GCGC *
54	2	TCGCCGCGCAGCGCGCCAAGAGAT	6	GCGC
59	2	TCGCCGCGCAGCGCGCCAAGAGAT	11	GCGC
60	2	TCGCCGCGCAGCGCGCCAAGAGAT	12	GCGC *
61	2	TCGCCGCGCAGCGCGCCAAGAGAT	13	GCGC
122	5	AGCGCCCATGGCCCCGGCAGGCGC	2	GCGC
141	5	AGCGCCCATGGCCCCGGCAGGCGC	21	GCGC
201	8	TACGTCCTGCGCGGCGACAAATAA	9	GCGC
202	8	TACGTCCTGCGCGGCGACAAATAA	10	GCGC *
331	13	CCATGACCATGCGTCAATGCGCTA	19	GCGC
392	16	TCGTCGCCGCGCTGCCGGAGATGG	8	GCGC *
393	16	TCGTCGCCGCGCTGCCGGAGATGG	9	GCGC
500	20	ACACCATTATGCAGATGGCCGCGG	20	GCGC *
569	23	GCTACGGCGATCTGCGCTGCG	17	GCGC
607	25	TCGGCTGCGCCGGACGCGCCGTCA	7	GCGC
608	25	TCGGCTGCGCCGGACGCGCCGTCA	8	GCGC *
615	25	TCGGCTGCGCCGGACGCGCCGTCA	15	GCGC *
650	27	AACAACGCGCCTACCAGGACG	10	GCGC

735	30	TCCCCATGCCCATCCCCGAAAACA		15	GCGC	*
797	33	ACGCCCCCAACAATATCTCCA	5		GCGC	*
848	35	GCAACCTGCCCCTCCCCCCCTCA		8	CCGC	
954	39	TGATCCACTTTGTGCCCCCCGACA		18	CCGC	*
971	40	ACATCGTGCAGCCCCCCACATCC		11	CCGC	
972	40	ACATCGTGCAGCCCCCCCAGATCC		12	CCGC	*
973	40	ACATCGTGCAGCCCCCCCAGATCC		13	CCGC	
974	40	ACATCGTGCAGCCCCCCCCACATCC		14	CCGC	*
1050	43	ACGAATACCGCACCCTGCCCCAGA		18	CCGC	
1185	49	CCCAACAAAACCCCCCCCCACCAC		9	CCGC	*
1240	51	ACCGCCCAACGTAATCTGCCCCTC		16	CCGC	
1278	53	CCCCAAACCCCCCCAAAACACCCC		6	CCGC	*
1279	53	CCCCAAACCCCCCCAAAACACCCC		7	CCGC	
1280	53	CCCCAAACCCCCCCAAAACACCCC		8	CCGC	*
1290	53	CCCCAAACCCCCCCAAAAGACCCC		18	CCGC	
1386	57	CCACCCCCCCTAATCACCCTACCC		18	CCGC	*
1393	58	CCCTCCCCCTACCCCCCTTCCAAA		1	CCGC	
1484	61	CCCCCTCCCTCCCCCCCAGTATTCC		20	CCGC	*
1485	61	CCCCCTCCCTCCCCCCCAGTATTCC		21	CCGC	
1567	65	CATTTTCAGGAGCCCCACATCGTC		7	CCGC	
1568	65	CATTTTCAGCAGCCCCACATCGTC		8	CCGC	*
1697	70	CCCCTCATCCCTCATCATATCAGC		17	CCGC	
1698	70	CCCCTCATCCCTCATCATATCAGC		18	CCGC	*
1723	71	CCCCTCCCCAACCCCAGCAGCAAG		19	CCGC	
1765	73	CTACCCTCCCAAACCCTTTCCCCCC		13	CCGC	*
1812	75	ATCCCCAACCACCTCCTCCCCAC		12	CCGC	
1813	75	ATCCCCAACCACCTCCTCCCCAC		13	CCGC	*
144	6	AATTCTTCTCTTTCCCACATTTCC		0	CTTAAC	*
94	3	TCCCTCCAATAAACACACAGCCCC		22	CCCC	*
130	5	ACCCCCCATCCCCCCCCCCAGCCC		10	CCCC	*
134	5	AGCCCCCATCCCCCCCCCCAGCCC		14	CCCC	
348	14	TTTACCCTAAAACCCCTATCCCTA		12	CCCC	*
399	16	TCCTCCCCCCCCTCCCCCAACATCC		15	CCCC	
497	20	ACACCATTATCCACATCCCCCCCC		17	CCCC	*
583	24	AATCCCCCCCCCCCCAACCCAC	7		CCCC	
588	24	AATCCCCCCCCCCCCAACCCAC	12		CCCC	*
591	24	AATCCCCCCCCCCCCAACCCAC	15		CCCC	
772	32	CCTCCCCCCAAATCATCCCCATCT		4	CCCC	
800	33	ACCCCCCCAACAATATCTCCA	8		CCCC	*
838	34	CCATCCTTAAATACCCCAAATCCC		22	CCCC	
858	35	CCAAGCTCCCCCTCCCCCCCTCA		18	CCCC	*
1025	42	ACCCCCCCTCTAAACAGCCCAACC		17	CCCC	*
1188	49	CCCAACAAAACCCCCCCCCACAC		12	CCCC	*
1270	52	ATCCACCAAGTCCTCCAGCTCTTC		22	CCCC	
1369	57	CCACCCCCCCTAATCACCCTACCC		1	CCCC	
1402	58	CCCTCCCCCTACCCCCCTTCCAAA		10	CCCC	
1426	59	CCCCTCCTATTTCCCCCCATTAC		10	CCCC	*
1441	60	CATATCCCCCATATTTCCCACCCA		1	CCCC	*
1460	60	CATATCCCCCATATTTCCCACCCA		20	CCCC	
1472	61	CCCCCTCCCTCCCCCCAGTATTCC		8	CCCC	*
1509	62	CCCCCCCAACCACCCAACTACTAC		21	CCCC	
1704	71	CCCCTCCCCAACCCCAGCAGCAAG		0	CCCC	*
1735	72	CCCCTCCATAAACCCCTCATCCCC		7	CCCC	
1744	72	CCCCTCCATAAACCCCTCATCCCC		16	CCCC	

```
    MATRIZ DE COORDENADAS:
 0  0  0  0  0  0  0  0  0 NO.SEC: 17   SEC.: CCCGGG   ENZIMA: SMA I    ARCHIVO: AR   1E
 REG.: 1    ULT. REG.: 76
1Z 535 673 922 1003 1104 1524 1529 1596 1605 NO.SEC: 18   SEC.: TCGA   ENZIMA:
       ARCHIVO: AR   1ER REG.: 1    ULT. REG.: 76
 0  0  0  0  0  0  0  0  0 NO.SEC: 19   SEC.: CTCGAG   ENZIMA: XHO I    ARCHIVO: AR   1E
 REG.: 1    ULT. REG.: 76
99 657  0  0  0  0  0  0  0  0 NO.SEC: 20   SEC.: CCGCGG   ENZIMA: SST II   ARCHIVO: AR
   1ER REG.: 1    ULT. REG.: 76
22  0  0  0  0  0  0  0  0  0 NO.SEC: 21   SEC.: TGATCA   ENZIMA: BCL I    ARCHIVO: AR
ER REG.: 1    ULT. REG.: 76

0. BASE     NO. REG.          SECUENCIA                NO. EN REG.          SEC. EN..

 517          21          AAGTCGGCTCGGTCGAGGACCTCG          13              TCGA
 535          22          AACTCGAAGACGTGCTGCAAA          7              TCGA
 673          28          TCGATTTCGTGTTCTATGACGTGC          1              TCGA
 922          38          TGGCGGAAAAGCTCGGTACCCAGA          10              TCGA *
1003          41          GCCGCATGACGGTTATCGAGT          19              TCGA
1104          46          AGCTGGAATCGCTGCTGATCGAGT          0              TCGA *
1524          63          ACCGGAGTCAGCGGCGTCGATAGC          12              TCGA.
1529          63          ACCGGAGTCAGCGGCGTCGATAGC          17              TCGA *
1596          66          TTCGGCGGCGATAAAAAAGCTCAGC          12              TCGA *
1605          66          TTCGGCGGCGATAAAAAAGCTCAGC          21              TCGA *


 499          20          ACACCATTATGGAGATGGCCGCGG          19              CCGCGG
 517          27          AAGAAGGCGCCTACGAGGACG          9              CCGCGG *

 642          25          TCGGCTGCCCGGGACGCGGCGTGA          22              TGATCA


 26           5          AGCGCCCATGGCCCCGGCAGGCGC          6              GGTACC *
 27          38          TGGCGGAAAAGCTCGGTACCCAGA          15              CGTACC
 95          45          TGCCGACGCCCTGCACCATGCATG          15              CGTACC *
 46          72          GCGCTGCATAAACCGGTGATCCCG          18              CGTACC

 69          27          AAGAAGGCGCCTACGAGGACG          21              CTGCAG *
```

B) Busqueda de zonas participantes en la regulacion de Glutamino Sintetasa en E.coli.

Glutamino sintetasa es una enzima cuya funcion es central en el metabolismo nitrogenado de E. coli. (24), y por tanto se encuentra altamente regulada; De ahi la importancia de conocer en detalle la region de control del gene que la codifica. En esta zona se habian encontrado 3 probables sitios de iniciacion de transcripcion (operadores). Dos de los cuales se localizan muy lejos de la iniciacion de traduccion. Esto hizo pensar en la posible existencia de una region atenuadora en este fragmento de A.D.N.. Aunque metodos geneticos no sugerian esto se requeria un estudio mas detallado para confirmarlo. En la siguiente figura se pueden observar los posibles sitios de reconocimiento para iniciacion de transcripcion dentro de cajas y el sitio de anclaje ribosomal (para traduccion) superscrita con puntos. Si existiese una regulacion de tipo atenuacion como en el operon de triptofano (26) debe haber alguno o todos los componentes siguientes: A) La codificacion para un peptido iniciado despues de los posibles promotores mas lejanos. B) Dos o mas codones para glutamina en tandem y en fase con el peptido hipotetico. C) Estructuras secundarias de tipo haza sobrepuestas en la misma zona.

El programa TRAD utiliza una matriz tridimensional para traducir secuencias de A.D.N. o A.R.N. a secuencias de aminoacidos en las posibles tres fases. Las coordenadas de la matriz corresponden al valor ASCII de los tres componentes del triplete (Fig.3.2.2). Para detectar el peptido y los codones en tandem se hizo una traduccion de la zona en tres fases (Fig. 3.2.3)

Para buscar estructuras secundarias se uso el programa BUPAL. El programa permite visualizar con facilidad a secuencias invertidas repetidas dentro de D.N.A y R.N.A. y da todas las secuencias mayores a 4 bases con sus respectivas posisiciones. El programa consiste en hacer comparaciones de la secuencia de interes contra su secuencia complementaria generada en direccion opuesta.

BUPAL es una variante de HOMOLOG que genera una matriz donde se ven todas las posibles combinaciones de apareamientos dentro de una misma cadena. Revisando esta matriz primaria (antes de filtrar las secuencias menos significativas), en sentido vertical se puede observar que se generan dos eje palindromico en forma diagonal de derecha a izquierda. Al comienzo del proceso de deslizamiento de una secuencia con respecto a la otra y al agregar las ultimas bases de una de las secuencias al comienzo de la misma, se pueden observar palindromes cortos en las dos mitades de la matriz. Conforme se ejecuta el programa, los palindromes de la derecha de la secuencia ocupan la mayor

Fig. 3.2.1 Secuencia anterior al gene glnA indicando sitios de iniciacion de transcripcion y traduccion.

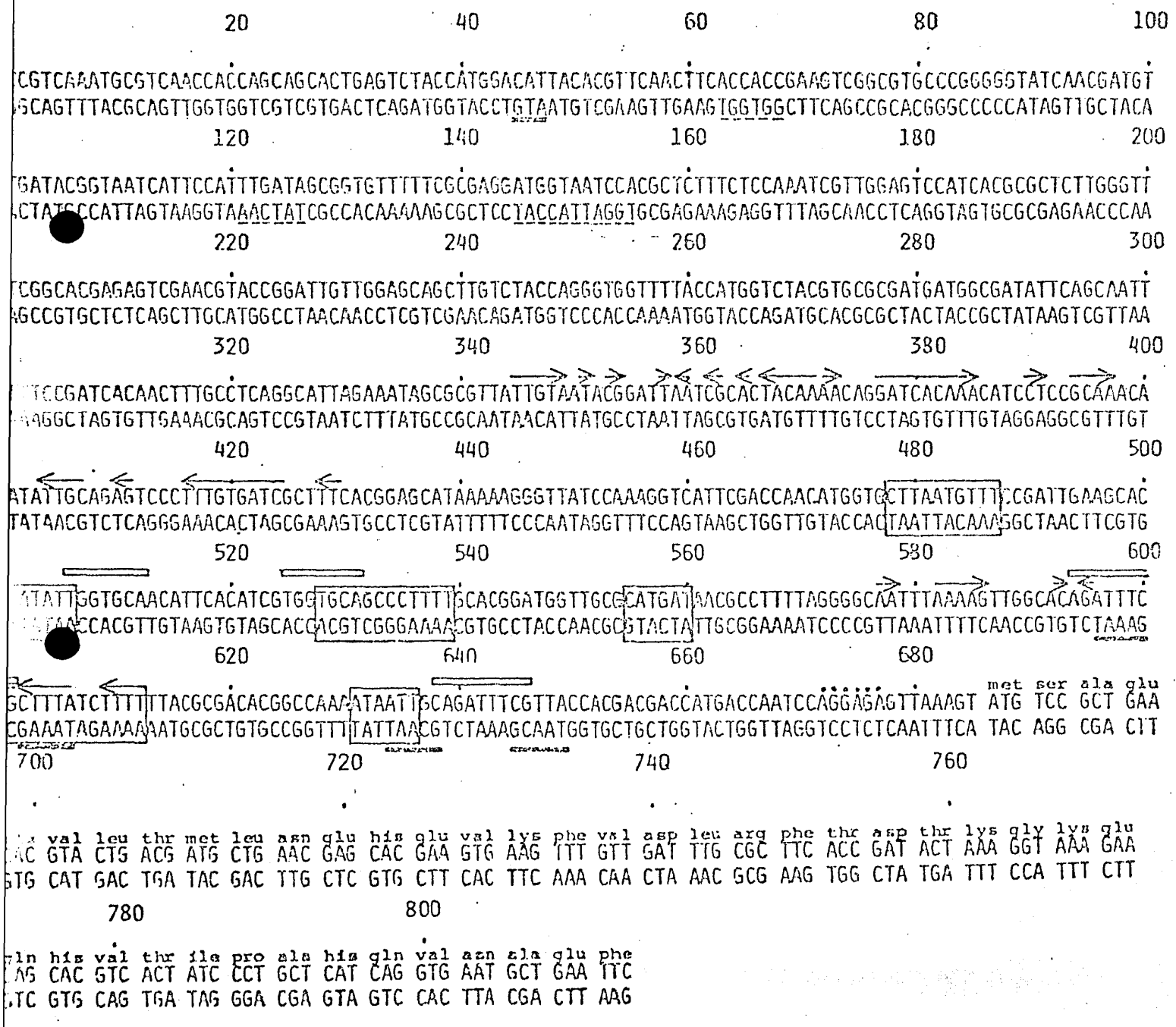


Fig. 3.2.2 Espacio cartesiano representando el metodo de traduccion de tripletes a aminoacidos.

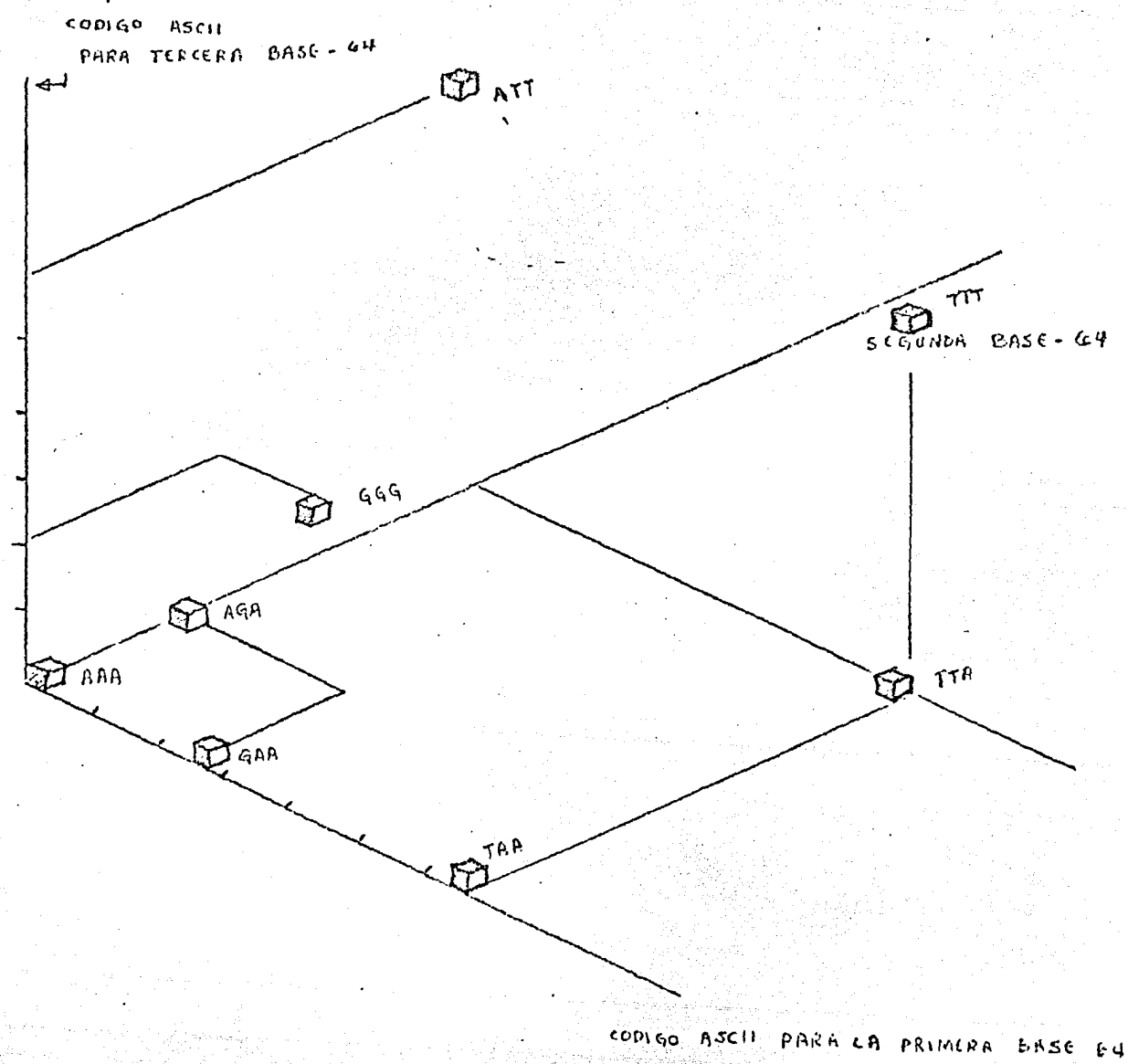


Fig. 3.2.3 Traduccion en tres fases de region de control de glnA.

```
IR  LEU THR LEU LEU ASP TRP SER TRP SER SER TRP FIN ARG ASN LEU GLN LEU PHE TR
:O  CYS ARG VLN LYS LYS ILE LYS ARG ASN LEU CYS GLN LEU LEU ASN CYS PRO FIN LYS
.A  LEU SER CYS ALA THR ILE ARG ALA LYS GLY LEU HIS HIS ASP VLN ASN VLN ALA PRO
.E  FIN CYS PHE ASN ARG LYS HIS FIN ALA PRO CYS TRP SER ASN ASP LEU TRP ILE THR
EU  PHE MET LEU GLU SER ASP HIS LYS GLY THR LEU GLN TYR CYS LEU ARG ARG MET PHE
.N  ILE LEU PHE CYS SER ALA ILE ASN PRO TYR TYR ASN ASN ALA LEU PHE LEU MET PRO
.U  A LYS LEU FIN SER

EU  FIN LEU SER TRP ILE GLY HIS GLY ARG ARG GLY ASN GLU ILE CYS ASN TYR PHE GLY
:G  VLN ALA FIN LYS ARG FIN SER GLU ILE CYS ALA ASN PHE FIN ILE ALA PRO LYS ARG
:S  TYR HIS ALA GLN PRO SER VLN GLN LYS GLY CYS THR THR MET FIN MET LEU HIS GLN
YR  SER ALA SER ILE GLY ASN ILE LYS HIS HIS VLN GLY ARG MET THR PHE GLY FIN PRO
HE  LEU CYS SER LYS ALA ILE THR LYS GLY LEU CYS ASN ILE VLN CYS GLY GLY CYS LEU
IN  SER CYS PHE VLN VLN ARG LEU ILE ARG ILE THR ILE THR ARG TYR PHE FIN CYS LEU
:G  GLN SER CYS ASP ARG

HE  ASN SER PRO GLY LEU VLN MET VLN VLN VLN VLN THR LYS SER ALA ILE ILE LEU ALA
.N  SER ARG LYS LYS ASP LYS ALA LYS SER VLN PRO THR PHE LYS LEU PRO LEU LYS GLY
.N  ILE MET ARG ASN HIS PRO CYS LYS ARG ALA ALA PRO ARG CYS GLU CYS CYS THR ASN
.E  VLN LEU GLN SER GLU THR LEU SER THR MET LEU VLN GLU FIN PRO LEU ASP ASN PRO
    TYR ALA PRO LYS ARG SER GLN ARG ASP SER ALA ILE LEU PHE ALA GLU ASP VLN CYS
    PRO VLN LEU FIN CYS ASP FIN SER VLN LEU GLN FIN ARG ALA ILE SER ASN ALA FIN
    YS VLN VLN ILE GLY
```

```
ADUCION DE SECUENCIA: TTTCCGATCACAACTTTGCCTCAGGCATTAGAAATAGCGCGTTATTGTAATACGC
TTAATCGCACTACAAAACAGGATCACAAACATCCTCCGCAAACAATATTGCAGAGTCCCTTTGTGATCGCTTTCACGGA
CATAAAAAGCGTTATCCAAAGGTCATTCGACCAACATGGTGCTTAATGTTTCCGATTGAAGCACTATATTGGTGCAACA
TCACATCCTGGTGCAGCCCTTTTGCACCGATGGTTGCCCATGATAACGCCTTTTACGGGCAATTTAAAAGTTGGCACAG
TTTCGCTTTCATCTTTTTTACGCGACACGGCCAAAATAATTGCAGATTTCGTTACCACGACGACCATGACCAATCCAGGA
AGTTAAAGT
```

Fig. 3.2.4 Ejes de simetria para deteccion de palindromes en matriz generada por BUPAL.

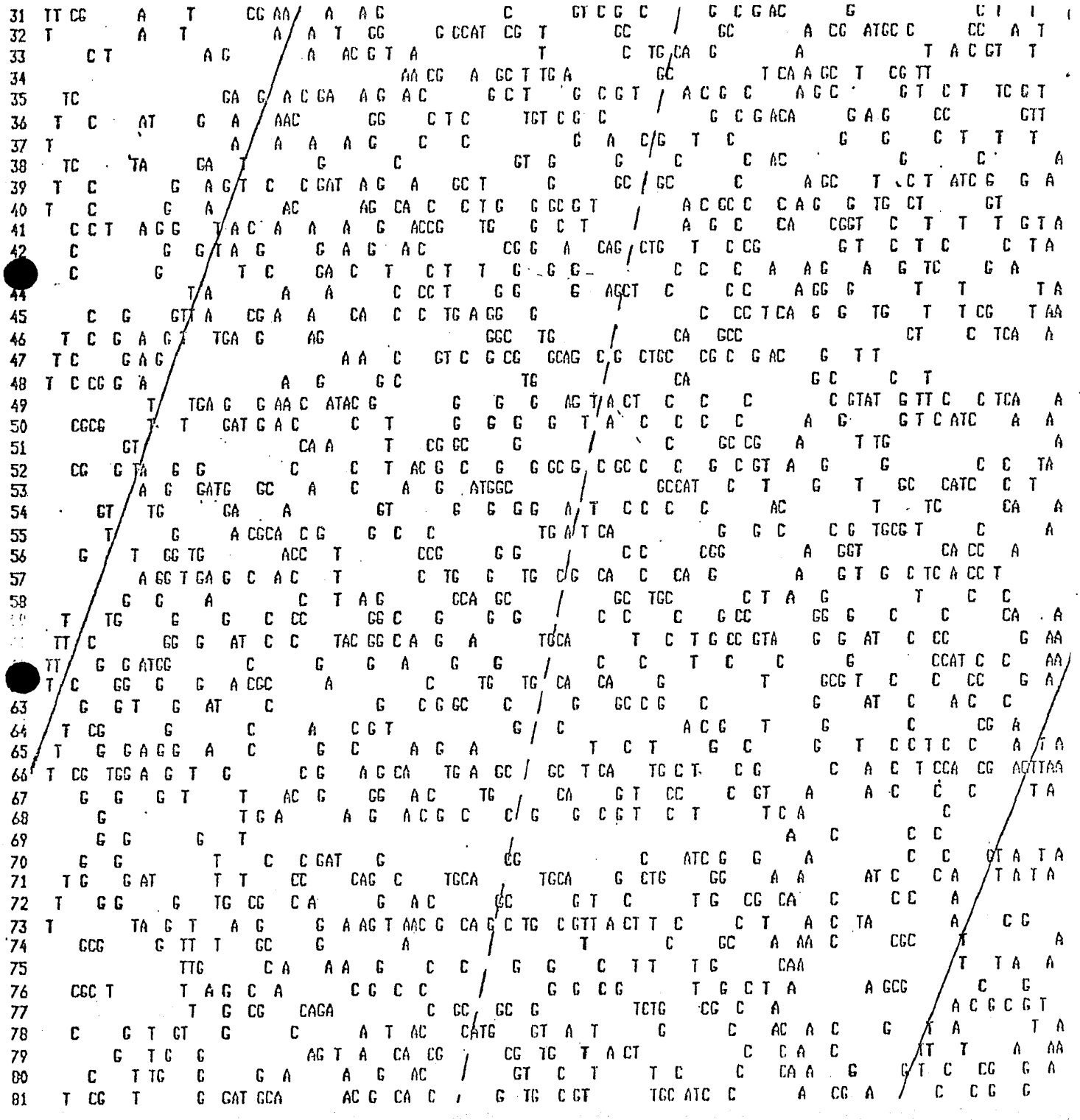


Fig. 3.2.6 Optima estructura para la busqueda de secuencias repetidas invertidas en un segmento de mas de 80 pares de bases.

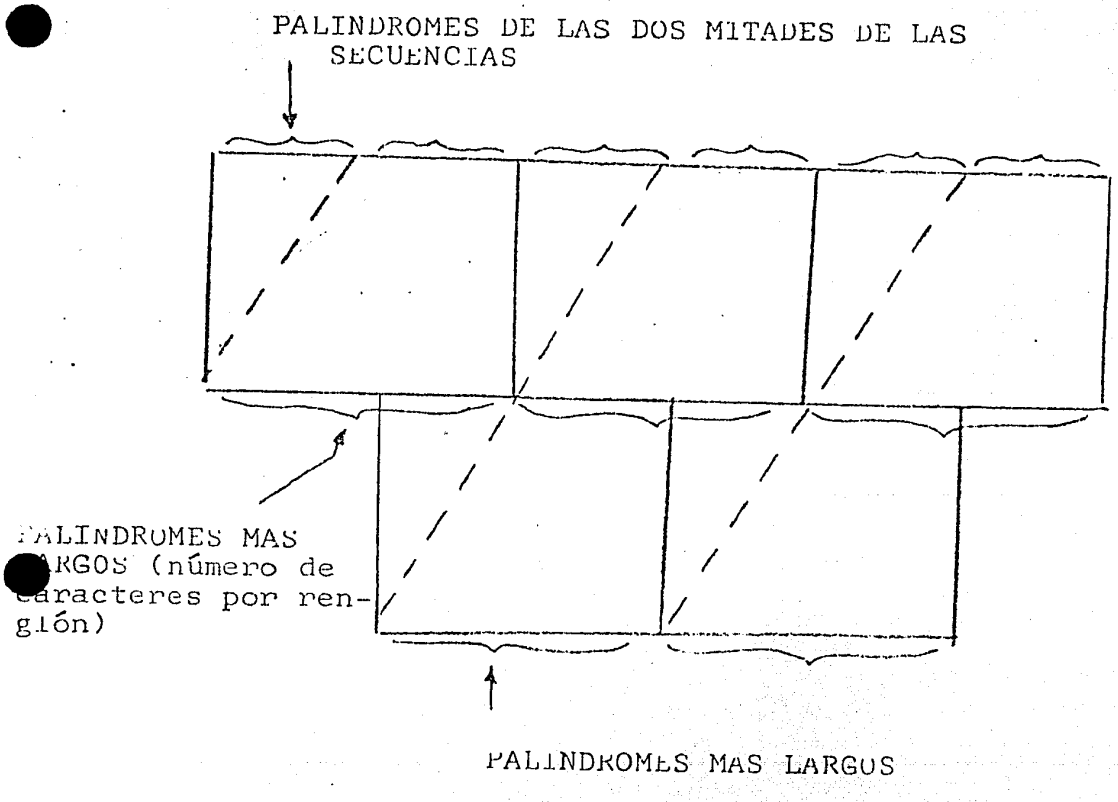
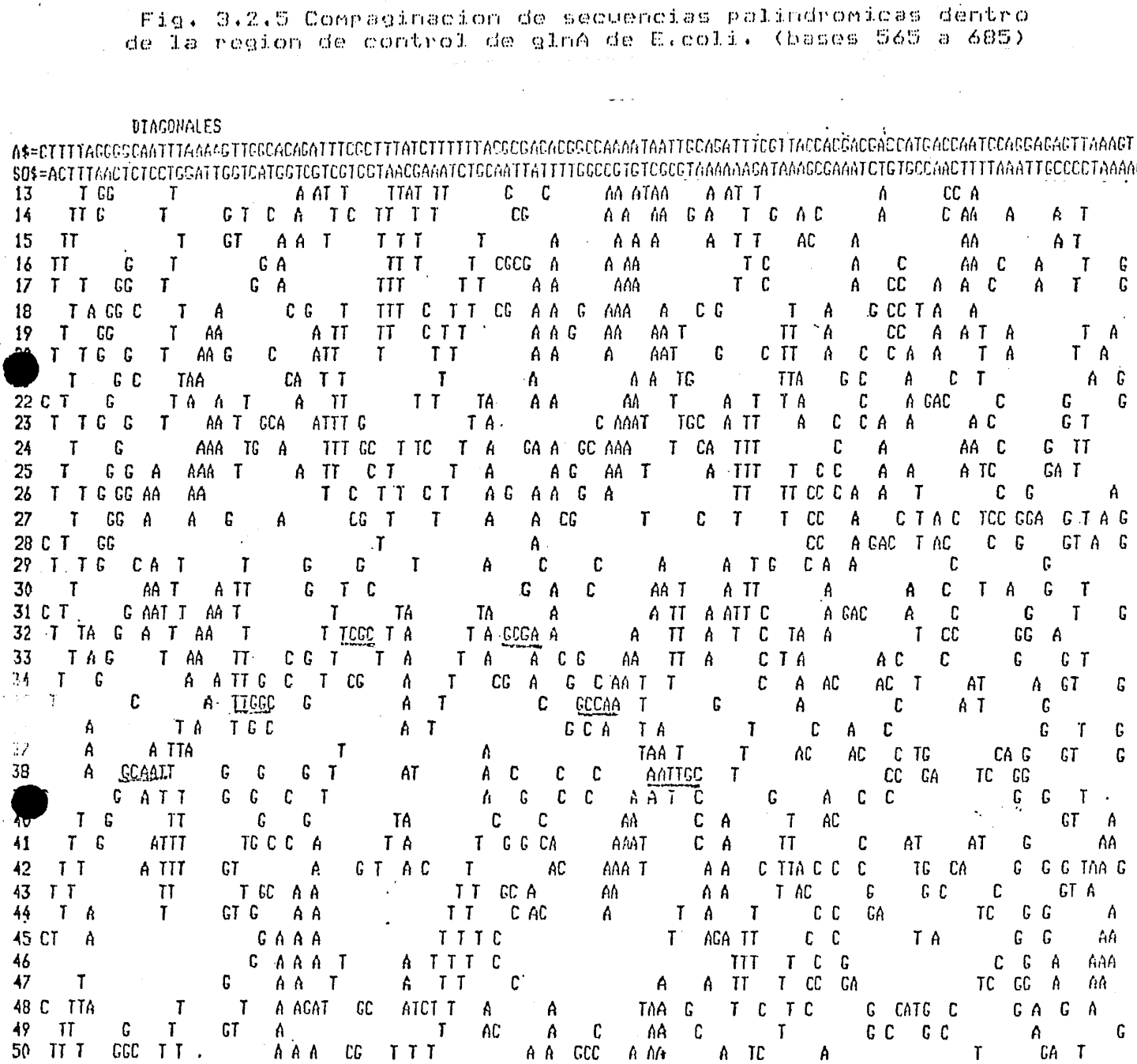


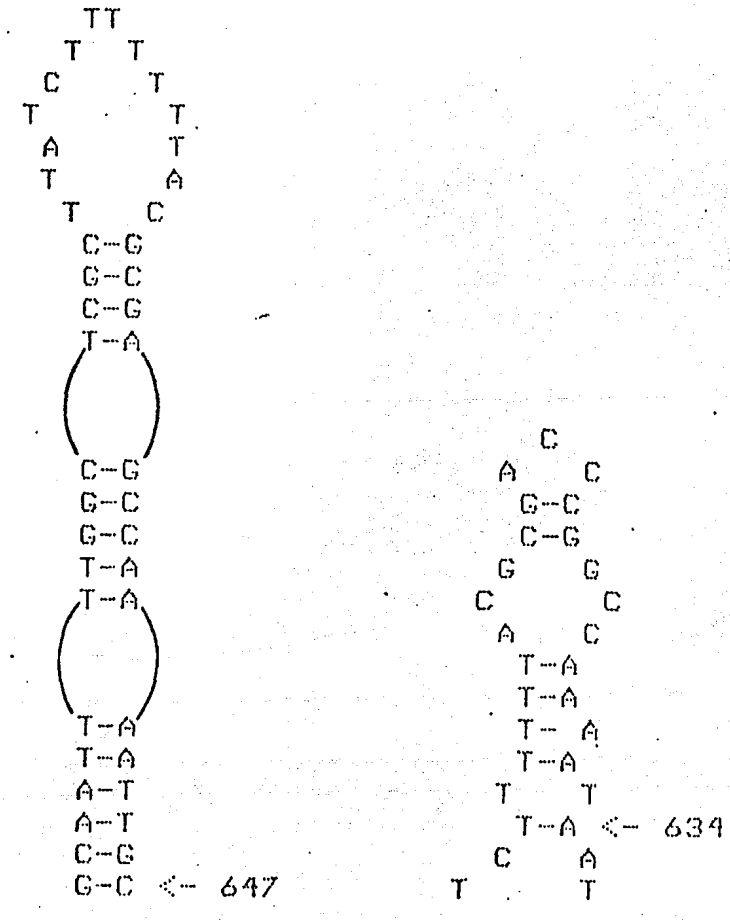
Fig. 3.2.5 Compaginacion de secuencias palindromicas dentro de la region de control de glnA de E.coli. (bases 565 a 685)



parte de la matriz hasta llegar al punto en que se compara toda la secuencia contra toda su complementaria sin agregar bases a los extremos. En este renglon se obtienen los palindromes de mayor longitud. Teniendo esto en mente, el analisis de una secuencia mas larga que el numero de caracteres por renglon se requiere hacer el estudio como indica la figura 3.2.5. Para obtener posibles estructuras secundarias se seleccionan las secuencias de mayor longitud y se compaginan con secuencias encontradas en los renglones superiores e inferiores. En la figura 3.2.6 se ve un ejemplo de las posibles combinaciones que se pueden producir.

Este fenomeno sugiere como se generan estructuras secundarias de gran complejidad y de utilidad regulatoria para la maquinaria de replicacion, trancripcion, y traduccion.

Se buscaron, utilizando BUPAL, posibles estructuras secundarias, las mas importantes se esquematizan en la figura 3.2.7. Esta zona forma dos gazas alternativas. La primera tiene mayor probabilidad de formarse ya que es termodinamicamente mas estable.



3.2.6 Dos posibles estructuras secundarias encontradas por

BUPAL en la region de control de glnA.

Revisando la traduccion de la zona encontramos que los peptidos posibles son relativamente cortos y que no existen los codones en tandem de glutamina. Por otra parte las estructuras secundarias que se observan no tienen conformaciones clasicas del mecanismo de atenuacion. Podrian, sin embargo, estas estructuras funcionar como sitios de reconocimiento para proteinas operadoras o activadoras.

Este estudio se realizo sobre la secuencia complementaria a la introducida por el investigador se uso el programa uando el programa COMP. COMP consulta una tabla en la cual se encuentra la letra correspondiente a la base complementaria. El indice de la tabla esta dado por el codigo ASCII de las las letras 'A', 'C', 'G', Y 'T', o sea (65, 67, 71, 84). Conforme se lee la secuencias se genera la secuencia complementaria haciendo la consulta necesaria.

E) Comparacion entre cuatro secuencias de glutamato deshidrogenasa Nadph dependientes y determinacion de distancias minimas mutacionales

Haciendo estudios sobre las secuencias de proteinas equivalentes en diferentes especies se ha visto una correlacion con las clasificaciones hechas por taxonomistas clasicos (8,9). En especial se han analizado las estructuras de citocromo C y de la histona H1. Este proyecto pretende realizar un estudio similar con la enzima glutamato deshidrogenasa-nadph dependiente de E.Coli, N.Crassa, Pollo, y Bovino (4).

El estudio se hizo sobre los primeros 87 amino-acidos de la proteina de Neurospora y poniendo en fase las de las tres secuencias restantes con respecto a los tres aminoacidos:

Alanina, Triptofano, Arginina de E.Coli (AWR) y
Glicina, Tirosina, Arginina (GYR) de las otras secuencias. (Fig. 3.3.1).

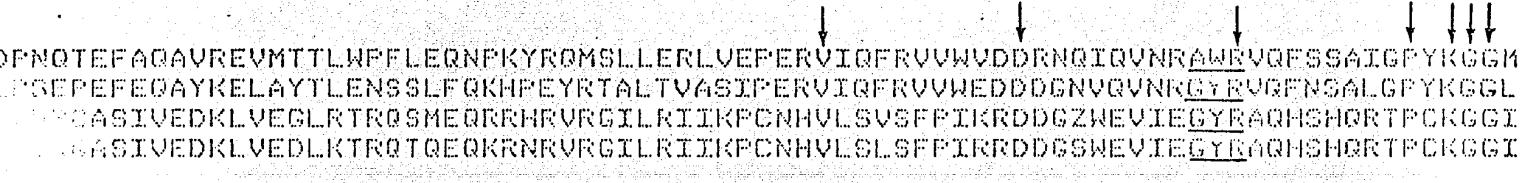


Fig. 3.3.1 Coincidencias dentro de 80 elementos de GDH nadph dependientes de E.coli, N.crassa, pollo, y bovino.

Para detectar homologia entre las proteinas se hizo una clasificacion de las posibles similitudes entre secuencias de la manera siguiente:

A) Homologia Absoluta
B) Cambio de una Base
C) Estructura o Polaridad Similar

Una comparacion de tipo absoluto implica que tiene que

haber el mismo amino acido en ambas secuencias. Por ejemplo tenemos que entre las secuencias de Escherichia coli y de Neurospora crass se encontro como indica la siguiente figura una similitud cercana al 50% y entre bovino y E. coli alrededor de 17%..

```
BUSQUEDA DE HOMOLOGIA

ABSOLUTO

 A$=KRDPNQTEFAQAVREVMTTLWPFLEQNPKYRQMSLLERLVEPERVIQFRVVWVDDRNQIQVNRAWRVQFSSAIGPYKGGHRFHPSVN
SO$=EGFFDRGASIVEDKLVEGLRTRQSMEQRRHRVRGILRIIKPCNNVLSVSFPIKRDDGZWEVIEGYRAQHSHQRTPCKGGIRYSLDVSVDEVK
1)                E       Q   R     L         V
2)                  L               R  P                          S      G
3)              V           R   L       V        D    V   R Q S    P KGG R    V  <--
4)                      EQ      L                DD                      G
5)  D         E  T          R              F    D                              SV
6)          V   T                          V          I   RA

LA HOMOLOGIA MAXIMA OCURRIO EN LA FASE 3 ,FUE DE 15 COINCIDENCIAS EN UN TOTAL DE 87 ELEMENTOS

EL PORCENTAJE DE SIMILITUD = .172413793



BUSQUEDA DE HOMOLOGIA

ABSOLUTO

 A$=KRDPNQTEFAQAVREVMTTLWPFLEQNPKYRQMSLLERLVEPERVIQFRVVWVDDRNQIQVNRAWRVQFSSAIGPYKGGMRFHPSVN
SO$=SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDQGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI
1)      N                                L          V          R
2)                          K                    V  D                  G
3)    P  EF QA  E  TL              L     PERVIQFRVVW DD    QVNR  RVQF SA GPYKGG R HPSVN  <--
4)                T    L Q        L           V  DD N          S      G
5)       E A               P YR        E          D  Q
6)    P               F                                 R

LA HOMOLOGIA MAXIMA OCURRIO EN LA FASE 3 ,FUE DE 44 COINCIDENCIAS EN UN TOTAL DE 87 ELEMENTOS

EL PORCENTAJE DE SIMILITUD = .505747126
```

Fig. 3.3.2 Comparacion absoluta entre las secuencias de E.coli y N.crassa, y entre Bovino y E.coli.

A continuacion se buscaron coincidencias de aminoacidos que diferian en una sola base de su codon de

traduccion. Para ello se construyo una matriz en la cual se viera esta relacion. Utilizando la tabla de la siguiente figura (3.3.3) se puede observar que todos los codones que se encuentran en una misma caja difieren en solo la ultima base. Los codones en un mismo renglon difieren en la base central y todos los codones que se encuentran en la misma posicion dentro de los diferentes cuadros pero en la misma columna difieren solo en la primera base.

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	End	UGA	End
UUG	Leu	UCG	Ser	UAG	End	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Fig. 3.3.3 Tabla de correspondencia entre tripletes y aminoacidos usada para determinar diferencias de una base.

La matriz que contiene la relacion entre los diferentes aminoacidos se muestra en la figura 3.3.4.

A	1	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	1	0	0
R	0	1	0	0	0	1	1	0	0	1	1	0	1	1	1	0	1	1	1	1	0
N	0	0	1	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	0	1
D	1	0	1	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	1
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	1	0	1	1
Q	0	1	0	0	0	0	1	1	0	0	1	0	1	1	0	0	1	0	0	0	0
E	1	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
H	0	1	1	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1
I	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	1	0	1	1	0	0
L	0	1	0	0	0	0	1	0	0	0	1	1	1	0	1	1	1	1	0	1	0
K	0	1	1	1	0	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0
M	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	0	0
F	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	1
P	1	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	1	1	0	0
S	1	1	1	0	0	1	0	0	0	1	0	1	1	0	0	1	1	1	1	1	1
T	1	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	0
W	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0
Y	0	0	1	1	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1

Fig. 3.3.4 Matriz mostrando los aminoacidos que difieren en solo una base.

Tomando en cuenta esta matriz el porcentage de homologia entre E.coli y N. crassa aumenta al 74% com se ve en la siguiente figura.

```
BUSQUEDA DE HOMOLOGIA

CAMBIO DE UNA BASE

A$=KRDPNQTEFAQAVREVMTTLWPFLEQNPKYRQMSLLERLVEPERVIQFRVVWVDDRNQIQVNRAWRVQFSSAIGPYKGGMRFHPSVN
SD$=SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI
1)      PNQT  AQ VRE M    PF   NP  RQ      L EP    QF  V VD RN     RA RV     IGP    M  H S
2)    DP      Q V EVMT  W  L Q FKY Q     RLV   R I    V  DDRN I   R W    SSA      G   HPS
3)  R PNQTEFAQA REV  TL  FL      R MS L     PERVIQFRVVWVDDR  IQVNRA RVQFSSAIGPYKGGMRFHPSVN
4) K  P Q       R V T   PFL QNP   QMSL E LV  E V    V  VDD N    N       SS      G M  HP
5)      Q E A  V E M   WPFLEQNFKYR  S LE L E E  IQF  V VDDR QI      RV FSSA       M    S
6)    DP   E  QAV   M TLWPFLEQ    R MSL   LVE  RV  F VV VDD   IQVNRA     SSA         F

LA HOMOLOGIA MAXIMA OCURRIO EN LA FASE 3 ,FUE DE 64 COINCIDENCIAS EN UN TOTAL DE 87 ELEMENTOS

EL PORCENTAJE DE SIMILITUD = .735632184
```

3.3.5 Comparacion entre N.crasa y E.coli considerando el cambio de una base.

Para detectar caracteristicas quimicas seleccionadas por presiones funcionales a nivel de estructura primaria (la secuencia de bases) GRAFPOL presenta un histograma donde la longitud de las barras corresponde a la hidrofobicidad de cada uno de los aminoacios de la secuencia peptidica.

La clasificacion es la siguiente:

* --->	HIDROFOBICO
** --->	CICLICOS
*** --->	NEUTROS
**** --->	POLARES BASICOS
***** --->	POLARES ACIDOS

En la figura 3.3.6 se pueden apreciar los resultados obtenidos del analisis de las secuencias de Glutamato Deshidrogenasa de las cuatro ramas filogeneticas. Se detecto un pico de region acida seguido por otro de hidrofobicidad ciclica para finalizar con otra acida cerca de la posicion 60.

Con el criterio de hidrofobicidad/hidrofilicidad (valores de cero a cuatro) encontramos una homologia del 61% en la secuencias de E. coli y N. crassa (Fig. 3.3.7)

```
BUSQUEDA DE HOMOLOGIA

POLARIDAD

A$=KRDPNQTEFAQAVREVMTTLWPFLEQNPKYRQMSLLERLVEPERVIQFRVVWVDDRNQIQVNRAWRVQFSSAIGPYKGGMRFHPSVN
SO$=SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI
1)      PN    A AV    T      E       M   ERLVEP   I      V      QVN   R  F  AI P       HP
2)   R    T  A   R V            K R MS          I  R V   D  QI   N AWR    S  A    Y   G   FH S N
3)     P   EF QA   E M TL   F     KY  M   LE    EPERVIQFRVVWVDD    IQVNR   RVQF  SA  GPYKGG  R HPSVN
4)     PN T     A     MT   PFL Q     R   SL ER V    V  F V    DD N    V  A      SS    P   G  M    P VN
5)  KR     E A  V     T  P     P YR M    R  EPE       V  D  NQI    RA     FS           F    VN
6)   R PN T FA    R V TT    F       Y         V  E      V        I    R WRV F   A     K    RF    VN

LA HOMOLOGIA MAXIMA OCURRIO EN LA FASE 3 ,FUE DE 53 COINCIDENCIAS EN UN TOTAL DE 87 ELEMENTOS

EL PORCENTAJE DE SIMILITUD = .609195402
```

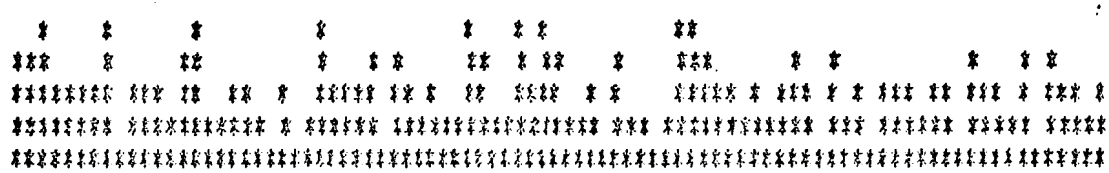
3.3.7 Comparacion entre E.coli y N.crassa tomando en cuenta la polaridad de los aminoacidos.

Es importante considerar ambos tipos de similitudes ya que una homologia puede implicar tanto evolucion convergente o como divergente dependiendo de la presion funcional para una caracteristica definida.

Para encontrar una relacion filogenetica mas clara se utilizo un sistema similar al que se uso con citocromo C. Se requiere de la matriz que se muestra en la figura 3.3.8. La tabla es similar a la publicada por Fitch y Margoliash y otros grupos y da la distancia minima mutacional entre los

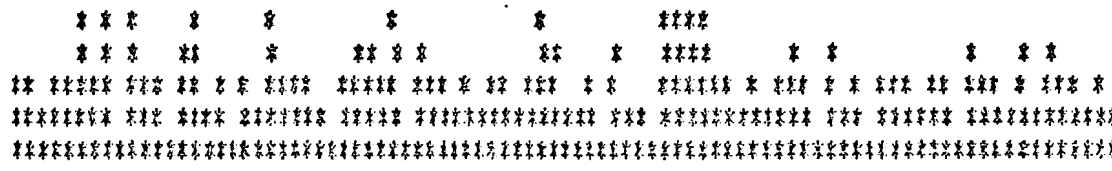
Fig. 3.3.6 Histogramas de hidrofobicidad/hidrofilicidad de las enzimas GDH nadph dependientes de E.coli, N.crassa, Pollo, y Bovino. Cerca de la posicion 60 se observa un pico de acidez.

E COLI



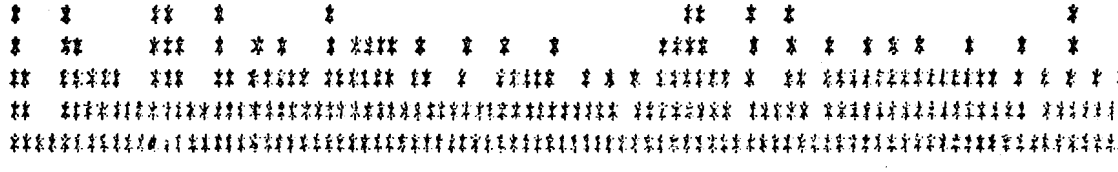
KRDPNQTEFAQAVREVMTTLWPFLEQNPKYRQMSLLERLVEPERVIQFRVVWVDDRNQIQVNRAWRVQFSSAIGPYKGGMRFHPSVN

N.CRASSA



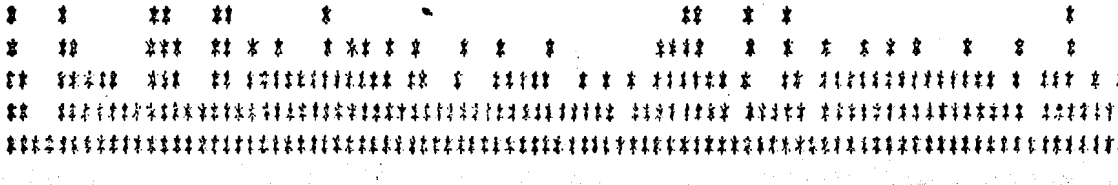
SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI

POLLO



EGFFGRGASIVEDKLVEGLRTRQSKEQRKHRVRGILRIIKPCNHVLSVSFPIKRDDGZWEVIEGYRAQHSHQRTPCKGGIRYSLDVSVDEVK

BOVINO



EGFFGRGASIVEDKLVEDLKTRQTGEQKRNRVRGILRIIKPCNHVLSLSFPIRRDDGSWEVIEGYRAQHSHQRTPCKGGIRYSTDVSVDEVK

diferentes aminoacidos. Esta se define como el numero de bases que requieren ser sustituidas para que una secuencia se transforme en otra. Por ejemplo tenemos que para que un codon que codifica para alanina como GCG pase a ser uno que codifique para glicina (GGG) la modificacion minima seria en la segunda base y por tanto la distancia minima mutacional es de 1.

	D	C	T	F	E	H	K	A	M	N	Y	P	Q	R	S	W	S	V	I	G
D	0	2	2	2	1	1	2	1	3	1	1	2	2	2	2	3	2	1	2	1
C	2	0	2	1	3	2	3	2	3	2	1	2	3	1	1	1	2	2	2	1
T	2	2	0	2	2	2	1	1	1	1	2	1	2	1	1	2	2	2	1	2
F	2	1	2	0	3	2	3	2	2	2	1	2	3	2	1	2	1	1	1	2
E	1	3	2	3	0	2	1	1	2	2	2	2	1	2	2	2	2	1	3	1
H	1	2	2	2	2	0	2	2	3	1	1	1	1	1	2	3	1	2	2	2
K	2	3	1	3	1	2	0	2	1	1	2	2	1	1	2	2	2	2	2	2
A	1	2	1	2	1	2	2	0	2	2	2	1	2	2	1	2	2	1	2	1
M	3	3	1	2	2	3	1	2	0	2	3	2	2	1	2	2	1	1	1	2
N	1	1	2	1	2	1	2	2	3	1	0	2	2	2	1	2	2	2	2	2
Y	2	2	1	2	2	1	2	1	2	2	2	0	1	1	1	2	1	2	2	2
P	2	3	2	3	1	1	1	2	2	2	2	1	0	1	2	2	1	2	3	2
Q	2	1	1	2	2	1	1	2	1	2	2	1	1	0	1	1	1	2	2	1
R	2	1	1	1	2	2	2	1	2	1	1	1	2	1	0	1	1	2	1	1
S	3	1	2	2	2	3	2	2	2	3	2	2	2	1	1	0	1	2	3	1
W	2	2	2	1	2	1	2	2	1	2	2	1	1	1	1	1	0	1	1	2
L	1	2	2	1	1	2	2	1	1	2	2	2	2	2	2	2	1	0	1	1
V	2	2	1	1	3	2	2	2	1	1	2	2	3	2	1	3	1	1	0	2
I	1	1	2	2	1	2	2	1	2	2	2	2	2	1	1	1	2	1	2	0

Fig. 3.3.8 Matriz indicando distancias minimas mutacionales entre los diferentes aminoacidos.

Se determinaron las distancias minimas mutacionales entre las cuatro secuencias. Estos valores se encuentran en la tabla de la figura 3.3.9. Para ello se llevaron a cabo dos pasos.:

I) Deslizar una secuencia contra la otra recortando segmentos de las secuencias para buscar maxima homologia absoluta.

La necesidad de efectuar esto implica que ocurrieron inserciones y deleciones a lo largo de la evolucion.

II) Usar la tabla de la figura anterior.

En la siguiente figura se presentan las distancias minimas de mutacion obtenidos.

	E	N	P	B
E I		42	54	62
N I			52	58
P I				11

Fig. 3.3.9 Distancias minimas mutacionales entre las cuatro secuencias.

Como se puede apreciar la distancia entre E.Coli y N.Crassa es 30% menor que la distancia entre Neurospora y Bovino. Tenemos entonces que este eucarionte primitivo esta mas cercano filogeneticamente a los procariontes que a los eucariontes modernos.

Con estos datos se genero un arbol filogenetico muy parecido al clasico.

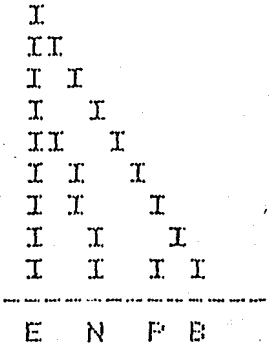


Fig. 3.3.10 Dendrograma obtenido apartir de de las distancias mutacionales.

En el dendrograma las distancias minimas mutacionales corresponden a la suma de de las distancias corresponden a las sumas de las longitudes de las ramas. Asi la suma de las dos ramas de Pollo y Bovino es igual a 11 y entre N. crassa y E. coli es de 42.

Esta representacion se debe tomar con mucha cautela por los siguientes motivos:

a) Se esta analizando una sola proteina.

b) Solo se analizan los primeros 87 aminoacidos de ella.

c) Se esta tomando la distancia minima y no la real por la degeneracion del codigo genetico. No podemos definir la historia verdadera de divergencia.

d) Ciertas partes de la estructura tienden a conservase mas que otras. (Entre Neurospora, Pollo, y Bovino existen veintiun glicinas en comun implicando posiblemente una estructura de tipo alfa-helice.

e) La velocidad de divergencia varia en diferentes ramas.

- Aumenta con la diminucion en el tiempo de generacion.

Algunas de estas objeciones son validas tambien en el modelo de Fitch y Margoliash como ellos mencionan.

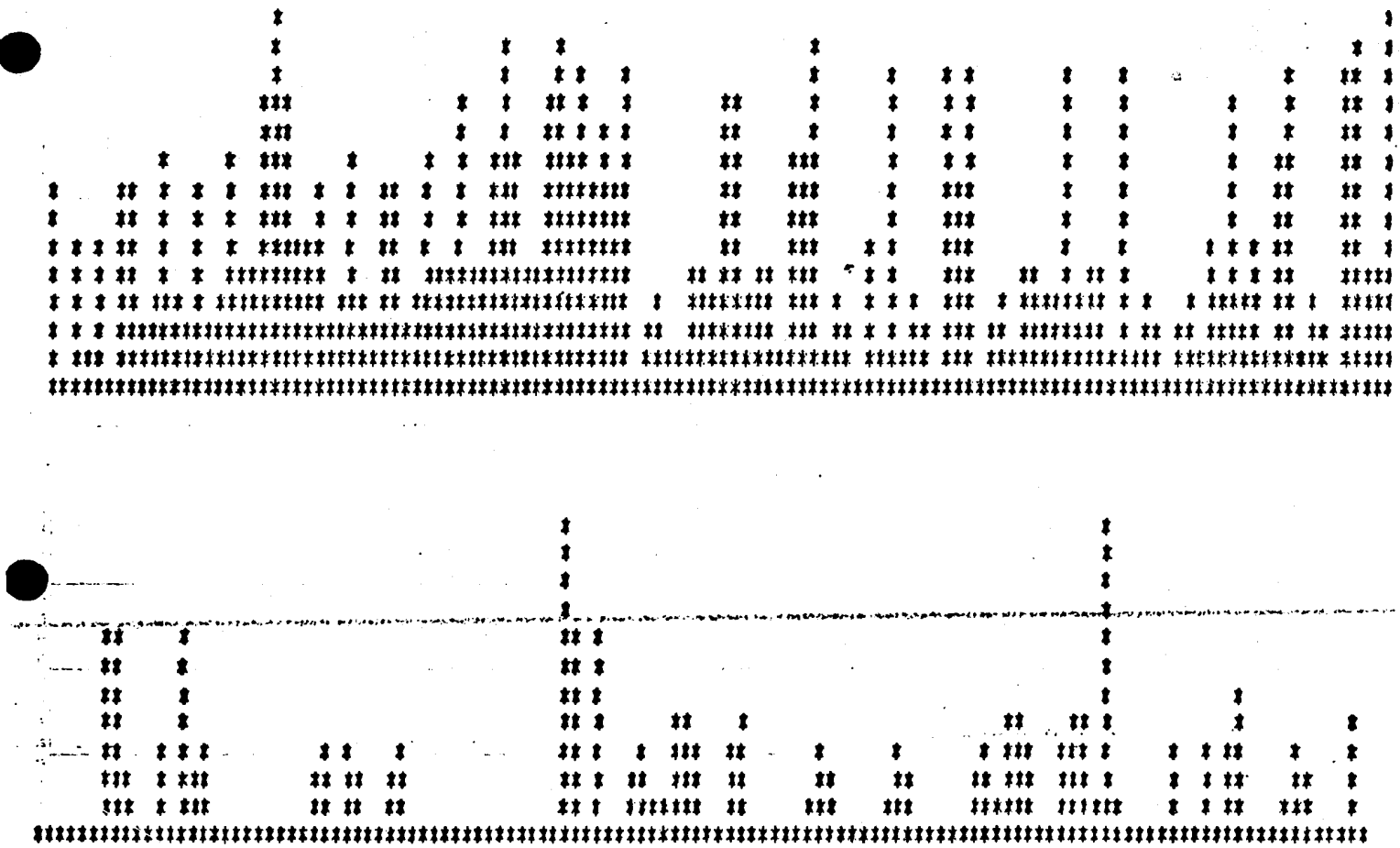
DISCUSIONES:

A. Para encontrar la gramatica general de la conducta exploratoria sera necesario trabajar con mas datos y en condiciones variadas.

B. Seria interesante aplicar el program MMD que representa relaciones arboreas a las secuencias conductuales. Con el posiblemente se detectarian categorias dentro de los diferentes tipos de ratones analogas a una relacion filogenetica. Para aplicar el programa basicamente se usaria en la misma forma que con las secuencias peptidicas. El problema se encuentra en como definir las distancias entre las distintas unidades conductuales. Una posibilidad seria definirlas a partir de un analisis de cumulos usando uno de los siguientes criterios: precedencia, gasto energetico, 'gasto emocional', o morfologia de cada una de las pautas.

C. La Programacion Dinamica es una de las tecnicas utilizadas en el reconocimiento de patrones (12). Fue desarrollada en el campo de la Investigacion de Operaciones. Se basa en comparar todos los puntos de una grafica contra todos los relevantes del segundo patron generando un valor que debe ser minimizado usando probabilidad y estadistica para encontrar el templado equivalente. La figura 4.1 indica que haciendo una comparacion punto a punto entre dos fonogramas de las palabras 'MASSES' y 'MASHES' la computadora rechaza la tesis de que son homologas. En cambio al usar la tecnica, el procesador puede declarar a ambos patrones como correspondiendo a la misma palabra pero pronunciada a velocidades diferentes.

Para revisar homologias visualmente se penso representar las secuencias como histogramas usando GRAFPOL y posteriormente buscar las similitudes utilizando alguna tecnica de reconocimiento de patrones como es la de programacion dinamica. Consideramos que es adecuada ya que permite una flexibilidad temporal en el reconocimiento. La siguiente figura muestra dos histogramas de secuencias de conducta donde cada estado corresponde a una barra cuya magnitud se determina por el valor ASCII del elemento de las secuencias. A simple vista se podria declarar que son dos estados fisiologicos diferentes los que definieron a las secuencias de eventos. Sin embargo requiere de una comprobacion formal.

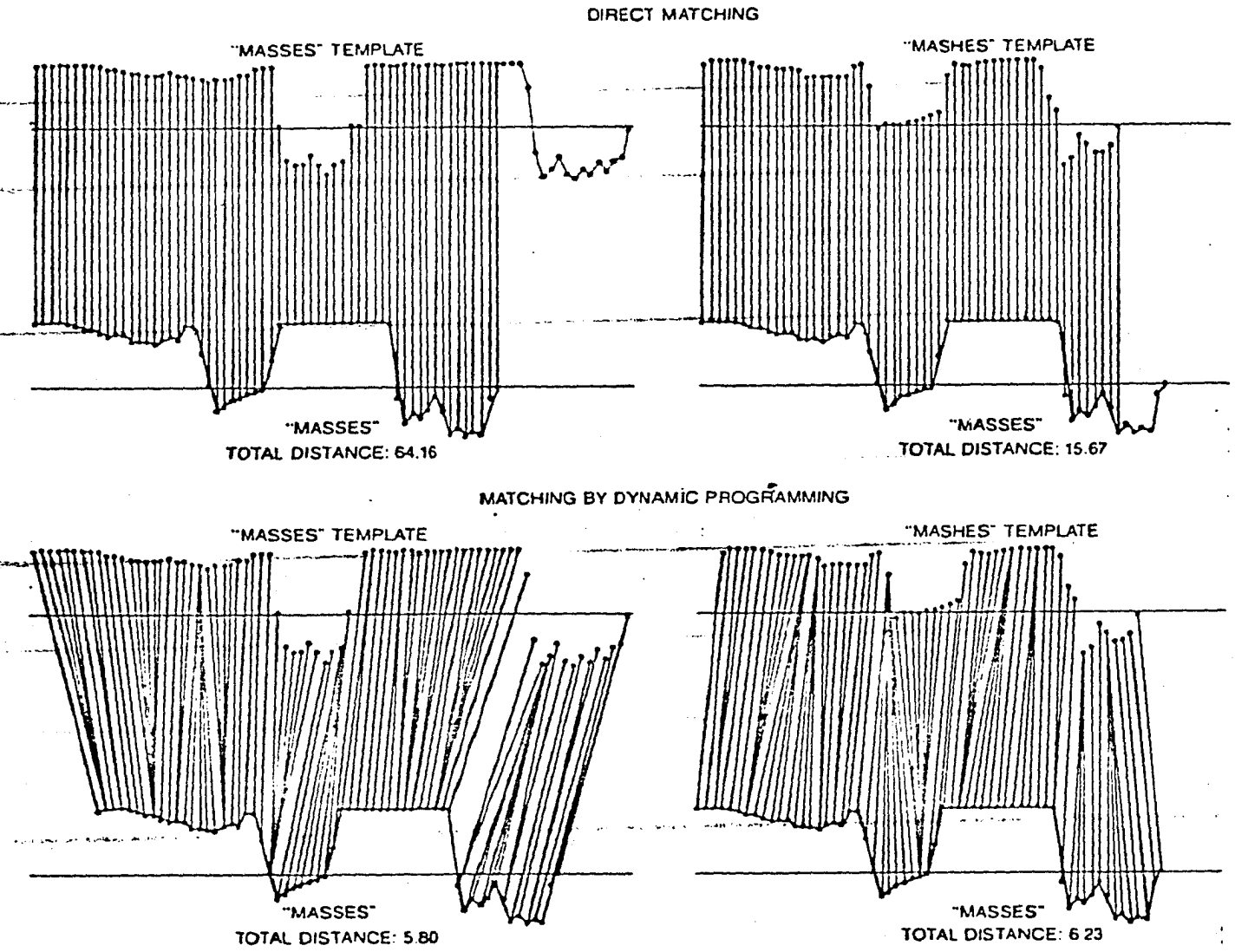


SECUENCIA 21 :

GGGG@3GKKG@C0KGKKG@BCACB0B@BCACB@KGK@@@@0K@@@@3KGBCACB@BCADAHACB@GK@BCADDCCDDACB@BCADDCDDEDDCDDAHBBHADDCDDACB@BCADDCEE@BCL

Fig. 4.1 Dos ejemplos de secuencias conductuales

Fig. 4.2 Reconcimiento de patrones usando la tecnica de programcion dinamica (12).



representadas en forma de histogramas.

D. Existen analogias adicionales al hecho de que las secuencias conductuales y macromoelculares se pueden tratar en forma similar:

ADN + MEDIO ---> PROTEINA ----> FUNCION. BIOQUIMICA

ADN + MEDIO ---> CONDUCTA ----> FUNCION BIOLOGICA

En el primer caso el factor genetico mas el medio interno, principalmente, determinan la manera en que se producira una proteina. Esta proteina tendra una funcion bioquimica. En el segundo sistema el mismo factor genetico mas el medio externo determinaran que conducta se presentara y esta tendra una funcion biologica. Actualmente las relaciones entre los primeros dos elementos del primer sistema y la regulacion de ellas son mucho mejor conocidas que las correspondientes del segundo sistema. Con la creacion de nuevas tecnicas de adquisicion de conocimiento, este aspecto tendra mayor comprension.

El trabajar con simbolos y no con su significado como sucede en el calculo proposicional tiene la ventaja de generalizar la funcion de los programas. Lenguajes funcionales como LISP (List Procesor) que han sido utilizados para trabajar con calculo proposicional deben ser igualmente efectivos en el tratamiento de los problemas aqui expuestos.

REFERENCIAS

1. Archer, J. Test for Emotionality in Rats and Mice: a review. Anim. Behav. 21:205-235, (1973)

2. Barnett, S.A., Cowan P.E. Activity, Exploration, Curosity, and Fear: an Ethological study. Interdisc. Sci. Rev. 1:43-62, (1976).

3. Blanchard, R.J., M.J.Kelley and D.C.Blanchard, Defensive Reactions and Exploratory Behavior in Rats. J. Comp. Physiol. Psychol. 87: 1129-1133, (1974).

4. Blumenthal Kenneth M., Moon Kenneth, Smith Emil, Nicotinamide Adenine Dinucleotide Phosphate-specific Glutamate Dehydrogenase of Neurospora. The Journal of Biol. Chem. Vol. 250, No. 10, pp.3644-3651

5. Colgan, Peter W., (ed.) Quantitative Ethology ED. E. Wiley & Sons Interscience Publications (1971).

6. Diaz, Jose Luis, Analisis Estructural de la Conducta , (en imprenta).

7. Fentress, J.C., Stilwell F.P. Grammar of a Movement Sequence in inbred mice. Nature, Lond. 244: 52-53, (1973).

8. Dickerson, Richard E., The Stucture and History of an Ancient Protein., Scientific American (1972), Vol.226, No.4

9. Fitch Walter M., Margoliash Emanuel, Construction of Phylogenetic Trees. Science Vol.155, p.279.

10. Glover S.W. Aspects of Genetic Engineering in Micro-organisms, Adv. in Microb. Phys. (1978) v.18, 235-271

11. Gouy M., Gautier C., Codon Usage in Bacteria: Correlation with Gene Expressivity.

12. Levinson Stephen E., Liberman Mark, Speech Recognition by Computer, Scientific American, April (1981)p.64

13. Lehninger Albert L., Biochemistry,. ED. Worth Publishers Inc. NY. (1976).

14. Maxam A.M., Gilbert W. (1977) PNAS vol.74. no.5, p.560-564.

15. Messing J., Crea R., Seeburg P. A System for Shotgun DNA Sequencing (1981) Nuc. Acids Res. 9, 309-321.

16. Puhler A., Klipp w., "Fine Structure analisis of the Gene Region for Nitrogen Fixation (nif) of Klebsiella pneumoniae."

17. Bothe H., Trebst A. (eds.) Biology of Inorganic Nitrogen and Sulfur. Springer-Verlag Berlin Heidelberg (1981).

18. Quinto C., De La Vega, Humberto, Flores Margarita, Fernandez Leonor, Ballado Teresa, Soberon Gloria, Palacios Rafael, Reiteration of nitrogen fixation gene sequences in Rhizobium phaseoli

19, Rosenberg, Genetic Regulation

20. Ruvkun Mary B., Ausbel Frederick M., Interspecies Homology of nitrogenase genes PNAS, Vol.77, No.1, pp.191-195 Sanger E., Nicklen S., Conson R., DNA Sequencing With Chain Terminating Inhibitors. PNAS vol. 74,No. 12, p.5463-5467.

21. Santis M., Diaz J. L. Location Response to a Startling Noise on the Preferred Grooming Site in Mice (bajo revision).

22. Scott K.F. et al, Klebsiella nif Structural Gene Sequences, J.Mol.Appl.Genet. Vol.1, No.1, 1981.

23. Timmis Kenneth N. "Gene Manipulation in vitro", Max Planck Institute for Molecular Genetics, Glover S.W., Hopwood D.A. (eds.) Genetics as a tool in Microbiology, Society for General Microbiology Symposium 31, Cambridge University Press. (1981).

24. Tyler, Bonnie, Regulation of the Assimilation of Nitrogen Compounds. Ann. Rev. Biochem. (1978) :47:pp. 1127-1162

25. Werner Arber, DNA Modification and Restriction, Prog. Nuc. A. Res. Mol. Biol. Vol.14:1 (1974)

26. Yanofsky C., Attenuation in the Control of Expression of the Bacterial Operon. Nature Vol. 289, Feb. 26 (1981) p.751.