leg. 1

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO
COLEGIO DE CIENCIAS Y HUMANIDADES
UNIDAD ACADEMICA DE LOS CICLOS PROFESIONAL Y DE POSGRADO
INSTITUTO DE INVESTIGACIONES BIOMEDICAS

TECNICAS COMPUTACIONALES APLICADAS A SECUENCIAS EN LA BIOLOGIA

T E S I S

QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN INVESTIGACION BIOMEDICA BASICA

P R E S E N T A

JAIME LAGUNEZ OTERO

MEXICO D.F. 1982

Bashe,
Dulce ser;
Porqué te perdiste en el verdor ?
Ya eras Uno con la Naturaleza, con la Vida.
Quizas es parte de tu incansable búsqueda por la bondad y el saber.
Y tu sonrisa y mirada misteriosa, que escondían ?
Bashe, con la carcajada de niña, eres una cascada en la selva.

En memoria de Beatriz Zagorin Hass

Para mis padres, Mario y Conchita, a quienes no puedo describir en palabras todo lo que siento por ellos.

A mi hermano Mario.
A mis hermanas Yvonne y Blanca María.

A Jeny

AGRADECIMIENTOS

Al Dr. José Negrete por las discusiones tan motivantes en el campo de la investigación, tanto de Inteligencia Artificial como de filosofía de la ciencia.

A la M.C. Marcela Santis por su amistad y por haberme permitido colaborar en el trabajo de análisis de conducta exploratoria.

Al Dr. José Luis Díaz por su invaluable consejo y crítica sobre la presentación del trabajo.

A la Dra. Alejandra Covarrubias por el tiempo que tan paciente y amablemente le dedicó a la supervisión del trabajo y por haberme permitido colaborar en su investigación sobre regulación.

A la M.C. Elvira Sanvicente y al Dr. Francisco Bolívar por permitirme participar en el trabajo de la comparación de las secuencias de glutamato deshidrogenasa.

Al Dr. Mario Lasúnez, un excelente asesor en todos los aspectos, y de quien tuve la fortuna de ser hijo.

INDICE

I. INTRODUCCION

El uso de sistemas computacionales en la biología ha aumentado de manera explosiva en los últimos años debido principalmente a la disminución en los costos de las unidades procesadoras. A su vez, esto ha sido causado por la creación de tecnología de alta concentración de circuitos electrónicos. Asimismo los investigadores se han visto obligados a utilizar las más avanzadas técnicas matemáticas y computacionales para resolver problemas biológicos complejos como son los ecológicos, los regulatorios, y los de diferenciación celular.

En este trabajo se revisan dos aplicaciones biológicas: A) en el estudio de la conducta y B) en la biología molecular.

A) Conducta

En la etología participan varias disciplinas, entre ellas la zoología, la fisiología, y la psicología. Sin embargo no se pueden utilizar los mismos métodos de adquisición de información que se usan en estas ciencias por lo que requiere de técnicas creadas ad-hoc para ella. Para encontrar la correlación entre la expresión conductual con los fenómenos neurofisiológicos se ha visto que los metodos utilizados no son lo suficientemente informativos (1,2,3). Se han usado diversos 'índices' para encontrar la relación entre estados emocionales y/o fisiológicos y la conducta observada. Estos índices son fundamentalmente de tipo cuantitativo y consisten en frecuencias con el que aparece un evento dado o la transición de un evento a otro. El orden completo con que suceden los eventos es muy importante para determinar el significado de la pauta conductual. Esto se ha visto en otros estudios etológicos como en la definición de territorios, el cortejo, y la comunicación entre infantes y adultos. Las unidades conductuales aisladas no son suficientes para obtener el significado satisfactorio que refleje el funcionamiento interno de la 'caja negra' que es el sistema nervioso central. Ejemplos de técnicas (5) que disecan las secuencias conductuales son:

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

Todos estos métodos enfocan el problema desde el punto de vista de la probabilidadad de suscitarse una transición de un evento a otro. Se concatenan aquellos eventos que se suceden o preceden frecuentemente. Esto es una reconstrucción de la realidad y no un estudio de los hechos como unidades íntegras. La analogía del problema sería

intentar determinar la similitud entre dos cadenas de A.D.N. solo con las frecuencias con que ocurren las bases y no con las secuencias en si. Una desventaja en particular del tratamiento de tipo Markoviano es que el número de datos requeridos crece de manera exponencial conforme se alarga la secuencia. Por otra parte, es muy probable que la ocurrencia de una unidad conductual sea dependiente de la serie anterior de eventos y no solamente del evento precedente. Este fenómeno es contrario a la definición 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 información sin los problemas ya mencionados.

Entre los programas para el análisis de secuencias de bases nucleicas utilizados en la Universidad de California en San Francisco, existe uno que genera una matriz con dos secuencias en los ejes X y Y para determinar homologías. Este programa fue el predecesor del programa HOMOLOG, aquí expuesto. Este programa permite visualizar zonas en común entre dos secuencias, ya sean nucleicas, proteicas, o conductuales y en forma diagonal u horizontal. La forma diagonal es un plano cartesiano donde cada eje corresponde a una secuencia como en la matriz del programa de la Universidad de California. En las coordenadas determinadas por las posiciones de los elementos se presentan los elementos homólogos o espacios si no hubo coincidencia. Esto es diferente al programa de San Francisco ya que éste no imprime el elemento explícito sino solo un punto. Con la nueva presentación se pueden ver secuencias en sentido diagonal corresepondientes a las zonas homólogas. (Fig. 1.1)

```
  lomqragccbagcbagcpok
a       a     a     a
c         cc    c     c
g          g     g     g
b            b     b
a       a     a     a
g          g     g     g
c         cc    c     c
a       a     a     a
g          g     g     g
b            b     b
c         cc    c     c
```

Fig. 1.1 Presentación diagonal del programa HOMOLOG.

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

Aquí vemos la presentación horizontal (Fig. 1.2) que tiene la misma finalidad que la de la figura 1.1 pero que permite ver con mayor claridad las zonas homólogas. El algoritmo de esta presentación es diferente al anterior y consiste en realizar deslizamientos de una secuencia sobre la otra para determinar si en cada una de las posiciones posibles de las secuencias existe el mismo elemento. Además el programa puede dar una lista de las secuencias similares mas importantes. HOMOLOG se aplicó tanto a secuencias conductuales como a macromoleculares (Fig. 1.2). El programa TDA que traduce la conducta exploratoria a secuencias de símbolos permite la utilización de HOMOLOG.

```
A$=EGFFDRGASIVEDKLVEGLRTRQSMEQRRHRVRGILRIIKPCNHVLSVSFPIKRDDGZWEVIEGYRAQHSHQRTPCKGGIRYSLDVSVDEVK
SO$=SNLPSEPEFEQAYKELAYTLFWSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI
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 Presentación horizontal del programa HOMOLOG.

B) Biología Molecular

En el control del metabolismo existen mecanismos muy diversos para obtener la cantidad óptima de una cierta sustancia, dada la situación energética del organismo, su fase reproductiva, y la cantidad de sustrato o sustratos existentes. Estos mecanismos incluyen la modificación de la eficiencia de síntesis de A.R.N. a partir de A.D.N. (traducción), la modificación de la eficiencia de la síntesis proteica a partir del templado de tripletes en A.R.N. (traducción), la eficiencia de unión enzima/sustrato y la velocidad de su separación. Para determinar con alta precisión a que nivel se lleva a cabo la regulación de la célula es un gran apoyo el estudiar en la zona de interés la estructura primaria del A.D.N., es decir la secuencia de

bases que lo componen. Se requiere determinar la localización de los sitios de unión de proteinas activadoras y represoras de síntesis de A.R.N, los de la A.R.N. polimerasa, y los de ribosomas. Estos sitios son identificados como secuencias definidas o como estructuras secundarias (estructuras tridimensionales formadas por la interacción entre la bases). Por otra parte existen sistemas más complejos que requieren de una regulación a nivel de traducción y transcripción simultaneamente (26). El caso mejor conocido es el de 'atenuación' en genes de enzimas biosintéticas de aminoácidos, en el cual se han identificado dos características importantes a nivel estructural de A.D.N. Estas son: a) la formación de estructuras secundarias alternativas en la región anterior al gene y b) la presencia de codones en tandem para el aminoácido correspondiente en fase con un péptido hipotético también localizado antes del gene estructural. Los programas creados intervienen en la detección de estas posibles estructuras en la zona de regulación del gene glnA de E. coli.

La ingeniería genética es una rama de la biología molecular que reune varios métodos, tales como la clonación molecular y secuenciación de A.D.N. con el fin de entender los sistema metabólicos celulares y de obtener importantes productos biológicos como son el interferón y la insulina (10,23). Con el advenimiento de la ingeniería genética se hizo posible la comprensión de la infomación almacenada en las secuencias de A.D.N. obtenidas con técnicas de laboratorio cada vez mas eficientes (14,15). La computadora acelera el proceso de análisis de estas secuencias considerablemente.

Una herramienta clave de la ingeniería genética son las enzimas de restricción las cuales son proteinas catalizadoras de la ruptura de A.D.N. de doble cadena (25). Los sitios de corte son comunmente secuencias repetidas invertidas denominadas palindrómicas de 4 a 8 pares de bases (Ej. ATGCA--TGCAT). Debido a secuencias mas largas de este tipo y al aparearse internamente las bases de una misma cadena de A.D.N. o A.R.N. se pueden formar gazas en el espacio. La localización de los sitios de reconocimiento para diversas enzimas de restricción reviste importancia cuando el objetivo es una caracterización fina de alguna región del A.D.N. para facilitar su manipulación posterior. Dado el gran numero de diferentes enzimas de restricción descritas a la fecha se hace indispensable el uso de la computadora para realizar la tarea repetitiva de buscar los sitios de corte en las secuencias.

El programa BUSEC puede encontrar secuencias cortas como estas dentro de secuencias largas. Igualmente detecta sitios de unión de proteinas activadoras para la síntesis de proteína (iniciadores) y sitios de unión de proteínas represoras (operadores). Para detectar estructuras secundarias se requiere el programa BUPAL que encuentra

secuencias repetidas invertidas. El programa HOMOLOG sirve para detectar zonas similares con el mismo sistema de regulación. Otros trabajos igualmente adecuados para la computadora son la cuenta de secuencias de alta repetición (de interés actual), la determinación de contenido de A-T/G-C, y la traducción y uso de tripletes (11).

Objetivos específicos.

En el caso del estudio de la conducta el problema principal estriba en traducir los hechos observados a una secuencia de carácteres. La sección II de este trabajo explica como se logró esta traducción en el estudio de la conducta exploratoria de ratones en un medio extraño y los resultados obtenidos de su análisis. La parte III describe algunos de los proyectos de biología molecular en los que se requirieron los programas mencionados y como fueron desarrollados.

En el trabajo se presentan los ocho programas: BUSEC, BUPAL, COMP, CRADNA, HOMOLOG, MMD, TDA y TRAD. Las funciones de algunos estos, relacionados con biomoléculas, se encuentran en la literatura. Se puede afirmar que estos son únicos tanto en método como en presentación ya que se desarrollaron con las funciones como guía y no con los algoritmos en si.

Datos Técnicos:

Se utilizó 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

A) Antecedentes

La conducta está definida por series de eventos motores llamados unidades conductuales que se presentan en forma de secuencias. Existen series que se presentan con regularidad y con cierta variación. Esta idea es paralela a la de melodía musical (6). Al conjunto de reglas que generan a las melodías se les ha considerado equivalentes a reglas gramaticales o de lógica interna (7). Los cambios cualitativos en estas reglas podrían reflejar cambios en los estados fisiológicos de los individuos o diferencias genéticas entre subespecies o grupos. En forma de gráfica de estados se podría representar este concepto de la manera siguiente (Fig 2.1):

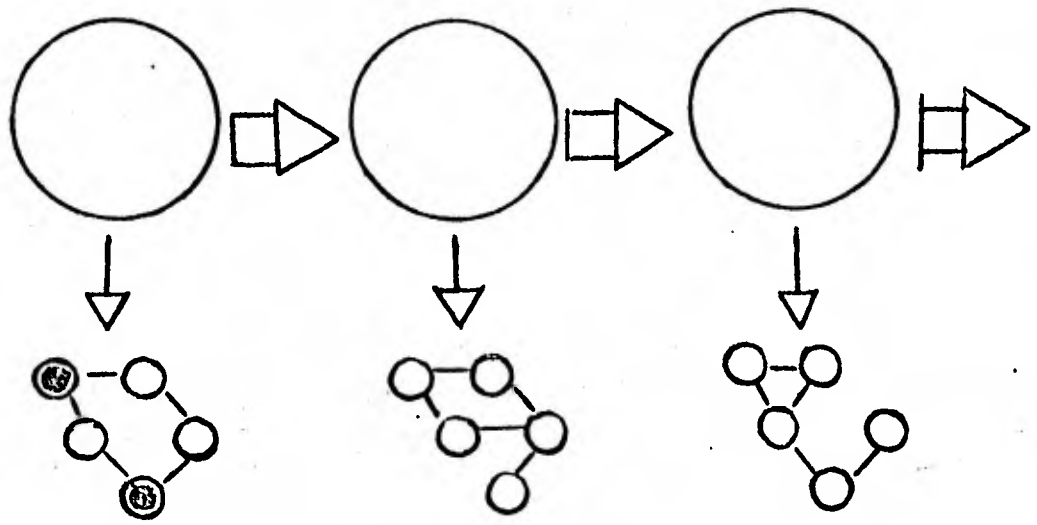


Fig. 2.1 Diagrama de estados conductuales

En este diagrama cada nodo superior corresponde a un estado fisiológico modificado por el origen de la cepa en estudio y/o variables ambientales. Las flechas horizontales en este caso representan la transición de un estado a otro. Cada nodo se puede representar como un subsistema cuyos nodos corresponden a unidades conductuales y las líneas 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, se debería poder determinar el estado fisiológico

correspondiente.

El resultado del análisis 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 (las cuales son representantes de los subsistemas) se desarrollaron los programas TDA (Traduccion De Actividades) y SDA (Secuencias en Diagonales Relacionadas). La función del primero es traducir la actividad del ratón 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 más 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 presentó y en que posiciones dentro de las secuencias.

Al analizar un número considerable de secuencias representativas de los movimientos podríamos encontrar una pauta exploratoria típica. Toda desviación de esta pauta podría tener un significado fisiológico.

B) Diseño experimental

Se estudiaron ratones durante la exploración de un medio novedoso y como control ratones bajo el efecto de el ansiolítico Diazepam (5 mg/kg) . Se utilizaron datos obtenidos de la observación de la conducta exploratoria de ratones albinos (BALB-c) machos durante 15 minutos en un medio clásico de estudio: el 'campo abierto'. Esto es una caja plana de madera sin objetos adicionales, pintada negra de 60 X 60 cm dividida en 36 cuadros de 10 X 10 cm (21) (Fig 2.2) La caja era lavada con amoniaco antes de cada observación. Como prueba adicional después de diez minutos de exploración a todos se les expuso a un ruido aversivo con duración de 2 segundos e intensidad de 100 db.

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 Numeración en campo abierto.

Recordando las secuencias homólogas 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 mediante el programa.

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

a. Actividad realizada
b. Coordenada espacial (localización)
c. Coordenada temporal (duración)

a. Actividad realizada

Se identificaron cinco actividades básicas:

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

Un episodio de aseo incluye lavarse, lamerse, rascarse, y la movilización de la cabeza, las patas, el cuerpo, los genitales, y/o la cola . Interrupciones del aseo por pausas con duración mayor a cuatro segundos determinan una nueva pauta. Los desplazamientos se registran cuando no existen interrupciones para explorar o asearse. La exploración se

determina cuando el ratón husmea una cierta zona. El congelamiento ocurre frecuentemente después del ruido aversivo y su duración mínima es de cinco segundos pero puede mantenerse hasta el final de la observación. Un desplazamiento posterior al estímulo de más de dos cuadros por segundo define un escape.

b. Coordenada espacial

Para clasificar zonas equivalentes dentro de la caja y poder extrapolar la información obtenida de los campos de exploración experimentales a otras condiciones se construyéron las siguientes matrices donde se intenta predecir 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

Fig.2.3 Matriz I - Gradiente de distancia con respecto al origen en campo abierto. Matriz II - Barreras físicas en campo abierto.

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 Matriz de tensión emocional resultante de I y II.

La matriz I es un gradiente discreto con unidades arbitrarias de distancia con respecto al punto de partida del ratón. En la esquina superior izquierda se observan tres ceros correspondiendo a la máxima cercanía al punto de partida. Moviéndose desde esta esquina hacia su opuesta, los valores aumentan de uno a tres. La matriz II clasifica la caja en zonas acotadas por dos, una o ninguna pared. El valor de cero representa zonas de doble pared, el valor de uno corresponde a zonas con una sola barrera y el de cuatro a zonas desprotegidas La suma punto a punto de las dos matrices genera la resultante.

Consideramos que esta matriz refleja la tensión emocional producida por las diferentes regiones de la superficie explorada debido a la lejanía del sitio de origen y a la falta de protección en forma de barreras. En tercera dimensión esta última matriz se vería como la figura 2.4. El ratón bajo observación debe escalar montañas de 'tensión emocional' para desplazarse.

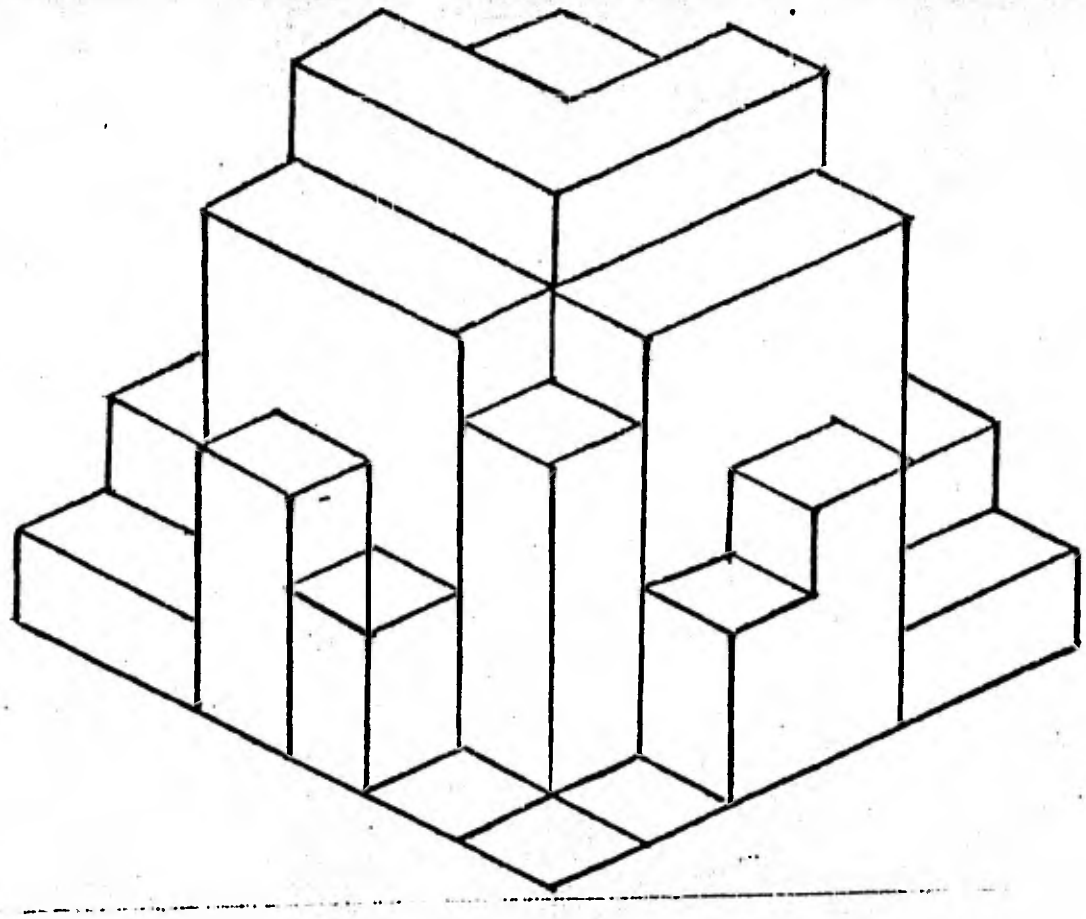


Fig. 2.4 Matriz de 'tensión emocional' en tercera dimensión.

Es posible que exista una relación 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 en cada punto exista un valor constante producto de esta relación, pero se requerirá de un análisis estadístico 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 especifica el numero de veces que aparece ese estado de manera continua en la secuencia. De este modo y a diferencia de otros estudios, nuestra gramática permite pasar de un estado a si mismo.

Por otra parte la posición dentro e la secuencia corresponde a la aparición cronológica del evento. Al realizar deslizamientos razonables, es decir cronológicamente equivalentes, de una secuencia sobre la otra se esta tomando en cuenta esta coordenada.

C) Descripción de código, algoritmo y experimento.

La siguiente tarea es la de definir los carácteres 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 carácteres correspondientes.

CODIGO

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

Fig. 2.5 Código utilizado para traducir.

El algoritmo en el programa TDA para realizar la traducción consiste en los siguientes pasos.

I. Leer el elemento en turno -->>

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

III. Definir duración 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 leído.

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

VI. Acumular en la cadena en producción y salir de la subrutina si se ha terminado la secuencia.

VII. Regresar al punto I.

Para determinar si se ha cruzado por alguna zona no registrada en los datos introducidos se utilizó la tabla de la figura 2.6. En las esquinas se se encuentran los numeros de los cuadros encontrados en la información sin procesar. Si ocurre un desplazamiento de esquina a esquina se habrá cruzado por las zonas indicadas en la tabla y se concatena con la secuencia en producción. En las esquinas del esquema se encuentran los valores de la numeración del campo abierto. Los números que se encuentran entre las esquinas

corresponden a los valores de "emocionalidad" por los cuales el ratón probablemente pasó para desplazarse de unos a otros.

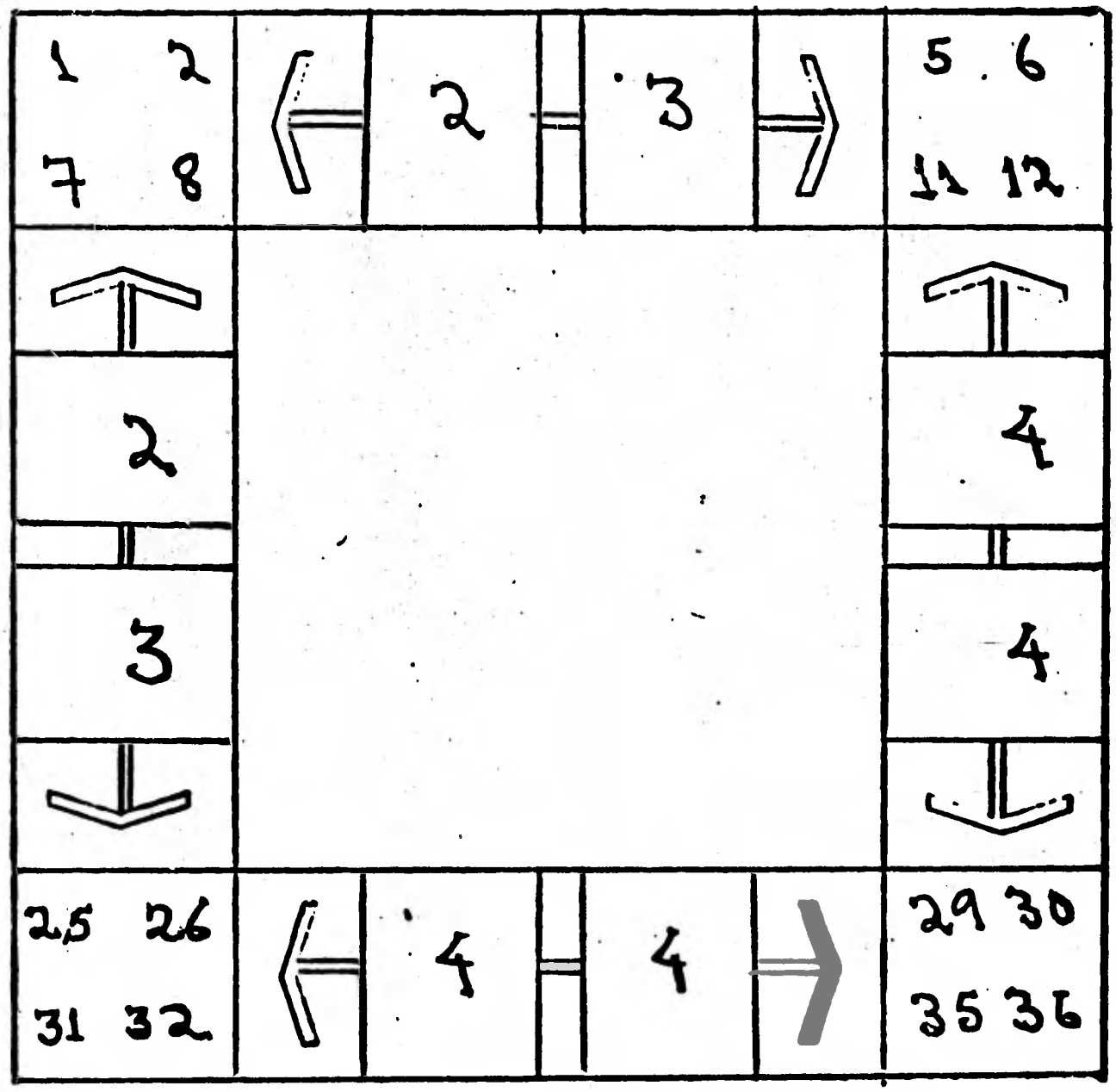


Fig. 2.6 Tabla utilizada para determinar las zonas por las que hubo desplazamiento.

Con el código ya definido se puede hacer una traducción de la información anotada a lápiz a secuencias de caracteres. La figura 2.7 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

C@B@BBCCBGQ

Fig.2.7 Ejemplo de traducción.

La primera expresión es la manera en que se registra la información. 'E' significa exploración. Los siguientes dos números corresponden al tiempo en minutos y segundos cuando comenzó la actividad y el siguiente número indica el cuadro o los cuadros en los que se realizó la actividad. En este caso después de explorar el cuadro uno, el ratón se desplazó por los cuadros 1,4,6,y 8. La expresión inferior es la misma información pero traducida utilizando los códigos con las zonas clasificadas. Esto es el tipo de secuencia que se puede usar en un programa como HOMOLOG. Nuestra gramática hipotética debera contener una aproximación de la siguiente secuencia como "melodía" central:

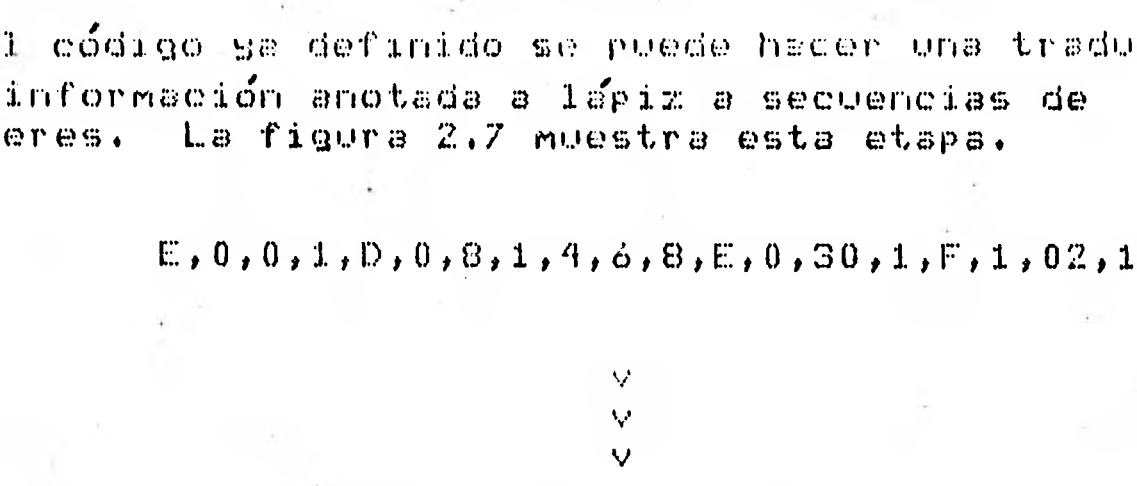


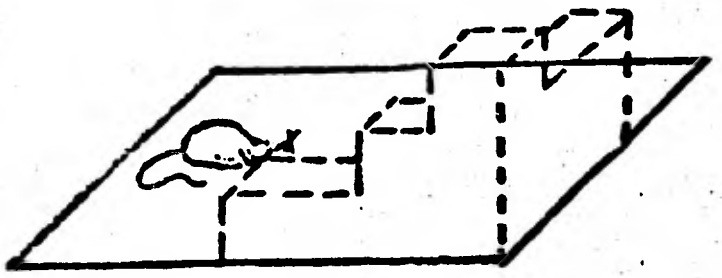
Fig. 2.8 Ciclo hipotético básico de secuencias de pautas.

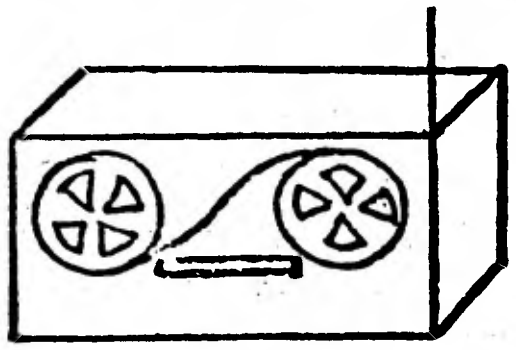
A ésto se le denominó el ciclo típico. En una secuencia promedio se debía esperar que se repitiera tres o cuatro veces con valores cada vez mas altos de ansiedad.

El siguiente diagrama esquematiza el flujo de

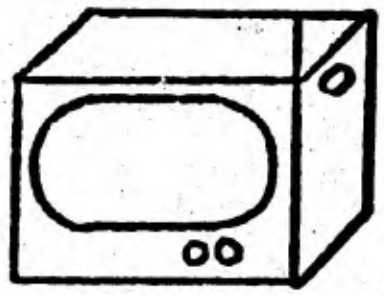
información dentro del sistema montado. Los ultimos tres pasos son los que se corresponden a este trabajo.

OBSERVACION ======>> GRABACION EN VIDEOCINTA ==>>



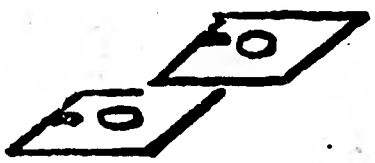


REVISION DE VIDEOCINTA===>>ANOTACION DE PAUTAS GENERALES==>>



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)

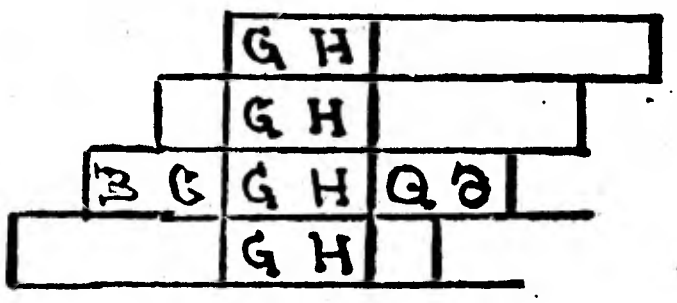


Fig. 2.9 Flujo de información en el estudio de la conducta exploratoria.

D) Discusión

Este proceso se aplicó 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 homólogos explícitos se crearon los archivos de SDR. El deslizamiento permite una flexibilidad temporal en las homologías encontradas. La restricción impuesta más importante fue la de hacer el deslizamiento en ambos sentidos por no más de siete eventos.

La figura 2.10 muestra las secuencias encontradas en dos o más conductas exploratorias. Encontramos que existen más secuencias homólogas entre los ratones bajo la influencia de la droga que entre los no afectados, un resultado similar a otros donde drogas uniformizan las poblaciones (Marcela Santis, comunicación personal).

La mayoría de las secuencias encontradas no son suficientemente largas o se presentan con frecuencias que no apoyan fuertemente la posibilidad de una secuencia general exploratoria. Existen dos posibles explicaciones para este resultado preliminar: las suposiciones requeridas no reflejan al sistema bajo estudio adecuadamente o la secuencia común general no existe. Se requiere ahora realizar estos análisis con más individuos y en condiciones experimentales variadas. Será necesario afinar los valores y las categorizaciones hechas antes del estudio. Sería interesante revisar las similitudes dentro de la secuencias homólogas ya encontradas usando las mismas técnicas. Teniendo más datos experimentales se puede hacer un análisis estadístico 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 territorial se crea un nuevo sitio de iniciación y posiblemente un nuevo gradiente a partir de ese punto. En estudios futuros será necesario revisar esta posibilidad.

```
T 002 NG@@
T 002 CB@G
T 002 BCA@BDDC
T 002 A@BDDC
T 002 CB@G@@
T 002 CB@G@
T 002 BCA@
T 002 CB@@BCA
T 002 DDBA@CB@
T 002 BACCB@
T 002 ACCB@G
T 002 G@CBCAH
T 002 GNG@
T 002 AACB@
T 002 G@@BCA@
T 002 ABCB@
T 002 CB@C
T 002 CB@BBCAB
T 002 BABCB@
T 002 ACCB@
T 002 AGCB@QBCA
```

```
T 002 @@BCA
T 002 @CB@
T 002 @BAB
T 004 @@N@
T 001 @G@
T 002 @NGNGN
T 002 @CBCA
T 002 GQGQGQ
T 002 GQGQ
T 002 NGQG
T 002 BCA@BDDC@DDBA@CB@
T 002 A@CB@G@
T 002 CB@G@@BCA
T 002 DC@DDBA@CB@
T 002 CB@GN
T 002 B@CB
T 002 G@@G@B
```

```
T 002 CB@N
T 002 B@G@GBCA
T 002 AQCB@
T 002 A@CB@ABCADBDDC
T 002 NGNGN
T 002 NGNG
T 002 GQGQN
T 002 GQNQGN
T 002 QGQN
```

Fig. 2.10 Secuencias encontradas en comun entre diferentes ratones.

III PROGRAMAS PARA EL ANALISIS DE SECUENCIAS DE ACIDOS NUCLEICOS Y PROTEINA, APLICACION A PROBLEMAS ESPECIFICOS

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

En estudios recientes (16,20,22) se ha hecho un análisis de la homología entre los genes involucrados en la fijación de nitrógeno (nif) en diferentes bacterias. Aprovechando esta información es posible extrapolar datos entre las especies para posteriormente intervenir activamente en la regulación del metabolismo y obtener una mejor comprensión de sus mecanismos regulatorios.

Al utilizar una enzima de restricción (25) con A.D.N. como sustrato se generan segmentos con longitudes promedio dependientes del número de bases reconocidas por la enzima. Estos segmentos se pueden visualizar en una placa de electroforesis en gel de agarosa que separe moléculas por tamaño. Con bromuro de etidio se ven las bandas fluorecer en luz ultravioleta. Para detectar homologías se obtienen moléculas híbridas producto de un proceso de desnaturalización y renaturalización con A.D.N. proveniente de otras especies. Específicamente se ha encontrado una gran homología entre el genoma bien conocido de Klebsiella pneumoniae y el de Rhizobium phaseoli (22). Se encontró que el plásmido pAC-30 conteniendo un inserto con los genes estructurales KD y H (Fig. 3.1.1) del operón nif de Klebsiella hibridaba con tres bandas de la digestión total (reacción enzimática completa) por la enzima ECO R1 del genoma de Rhizobium. Esto generó la pregunta si cada una de las bandas corresponde a uno de los genes del plásmido o si un gene de K. pneumoniae estaba hibridando con varios de Rhizobium. La estrategia mas directa para responder a esta pregunta sería obtener secciones totalmente intragénicas de los genes K, D y H, clonarlos en diferentes plásmidos e hibridar cada uno contra el genoma total para observar a que bandas corresponde cada plásmido. Con el propósito de encontrar la combinación mas adecuada de enzimas de restricción que produjera los segmentos necesarios se desarrolló el programa BUSEC.

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

A. Homología absoluta o
B. Similitud porcentual

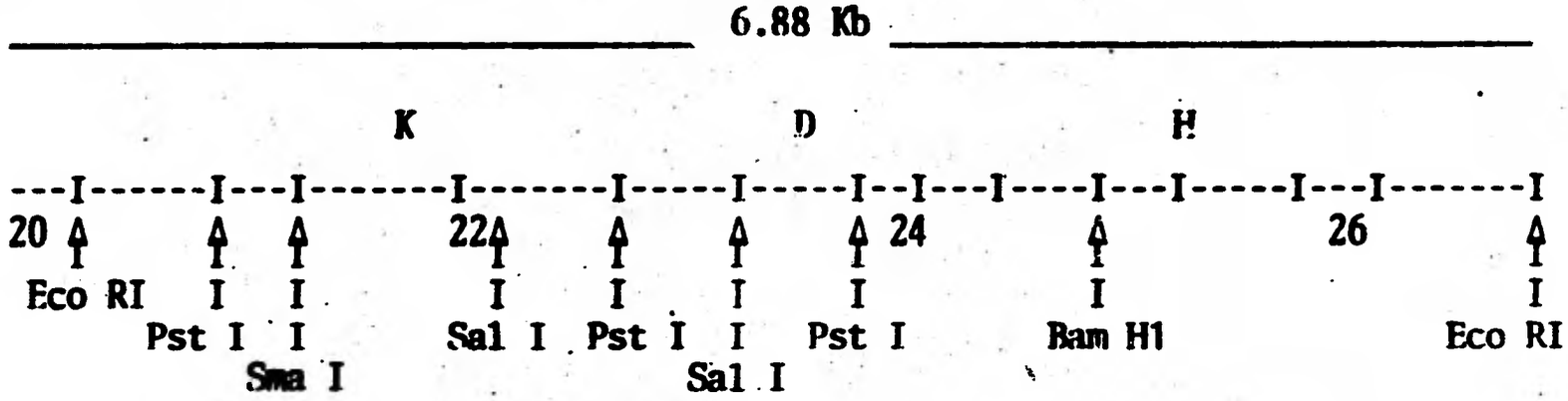


Fig.3.1.1 Inserto de KDH en pAC-25. Las flechas indican sitios de corte por las enzimas de restriccion. Kb=kilobases.

El primer método utiliza tres pasos principales:

1. Comparar la secuencia buscada contra la analizada elemento por elemento .
2. Almacenar la localización de secuencias encontradas en vectores de valores enteros.
3. Incrementar el apuntador de la secuencia en revisión para repetir comparación.

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 con los nombres respectivos de las enzimas que las reconocen.

El segundo algoritmo da flexibilidad a la definición de secuencias similares. La mejoría 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 de comparación de cada elemento, se determina el código ASCII (código estandard internacional para caracteres simbólicos) de los dos elementos a prueba. Se realiza una comparación booleana de estos valores y se considera como "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%.

```
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
```

Fig. 3.1.3 Parte de la secuencia del inserto de KDH que se analizó con BUSEC. La proteína 'D' comienza cerca del final del tercer renglón. La 'H' comienza en la base 900.

Dado que en el lenguaje BASIC se realiza la compilación del programa simultáneamente con la ejecución, el tiempo de procesamiento es relativamente largo. Sería recomendable trancribir este algoritmo a un lenguaje como el PASCAL que aumenta la velocidad de ejecución de siete a diez veces.

Para detectar una secuencia definida en forma más eficiente en ocasiones es preferible dar los elementos más conservados o más importantes como secuencia a buscar y posteriormente filtrar del resultado las secuencias no relevantes repitiendo la búsqueda con los elementos menos determinantes. Por ejemplo, es preferible hacer una búsqueda de 'GTAC' y luego filtrar haciendo una búsqueda completa sobre las secuencias encontradas que hacer una búsqueda del 80% de similitud de 'TGTACA'.

Para almacenar la secuencia ya publicada del inserto KDH (22) se hizo el programa CRADNA que crea un archivo de tipo acceso aleatorio con registros de 24 caracteres cada uno. Como aditamento, el programa facilita la revisión y corrección presentando en pantalla los registrosde tres en tres bases. En este caso el archivo contiene 76 registros. En el momento de ejecución de BUSEC se pregunta que archivo se va a revisar y cuales son los registros primero y último de interés. El resultado parcial, sobre el gene.'D' y parte del 'H', (figura 3.1.3) se encuentra en las siguientes cuatro páginas (figura 3.1.4). El estudio se realiza en tres pasos: 1) con las primeras siete secuencias a detectar, 2) de la 8 a la 16 y 3) de la 17 a la 21. Se pueden ver dos tipos de presentaciones. La primera con título "MATRIZ DE COORDENADAS" da los números de las bases donde se encuentran cada una de las secuencias, luego el número de secuencia optativa, la secuencia, el nombre de la enzima que la reconoce, el nombre del archivo que se revisó, y los registros primero y último. En la segunda presentación se encuentran seis columnas correspondiendo al número de base donde se encontró, en cual de los 76 registros se encontró, 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 detectó en dirección opuesta en la cadena.

Se concluyó posteriormente por otros métodos (18) que existe reiteración de información en Rhizobium.

```
MATRIZ DE COORDENADAS:
[illegible] 1002 1104 1524 1529 1596 1605 NO.SEC: 1   SEC.: AGCT   ENZIMA:
[illegible] 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
ER 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
ER 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
ER 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
ER 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
1ER REG.: 1   ULT. REG.: 76
754 0 0 0 0 0 0 0 0 0 NO.SEC: 7   SEC.: [illegible]   ENZIMA: [illegible] II   ARCHIVO: AR
1ER REG.: 1   ULT. REG.: 76
```

NO. BASE	NO. REG.	SECUENCIA	NO. EN REG.	SEC. ENZ
517	21	AAGTCGGGTCGGTCGACGACCTCG	13	AGCT *
535	22	AAGTCGAACACCTGCTGCAAA	7	AGCT *
673	28	TCGATTTCGTGTTCTATGACGTCC	1	AGCT *
922	30	TGGCCGAAAAGCTCGGTACCCAGA	10	AGCT
1003	41	GCCGCATGACCGTTATCCACT	19	AGCT *
1104	46	AGCTCGAATCCCTGCTGATGCACT	0	AGCT
1524	63	ACCCGAGTCAGCGGCGTCGATAGC	12	AGCT *
1529	63	ACCGGACTCAGCGGCGTCGATAGC	17	AGCT
1596	66	TTCGGCGGCGATAAAAAGCTCAGC	12	AGCT
1605	66	TTCGGCGGCGATAAAAAGCTCAGC	21	AGCT
754	31	AAGCCCACCACATCTACATCGTCT	10	AGATCT

```
MATRIZ DE COORDENADAS:
75 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: 8   SEC.: GAATTC   ENZIMA: ECO RI   ARCHIVO: AR   1ER REG.:
1   ULT. REG.: 76
[illegible] 130 134 348 379 497 503 580 591 772 800 838 858 1025 1180 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 0 0 NO.S
EC: 9   SEC.: GGCC   ENZIMA: HAE III   ARCHIVO: AR   1ER REG.: 1   ULT. REG.: 76
[illegible] 53 54 59 60 61 122 141 201 202 331 392 393 500 569 607 608 615 656 755 797 8[illegible]
[illegible] 754 971 972 973 974 1050 1105 1240 1276 1279 1280 1290 1366 1393 1404 1405 1[illegible]
[illegible] 1568 1697 1670 1723 1765 1812 1813 NO.SEC: 10   SEC.: GGGC   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 NO.SEC: 11   SEC.: AAGCTT   ENZIMA: HIND III   ARCHIVO: AR   1ER RE
G.: 1   ULT. REG.: 76
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 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 NO.SEC: 12   SEC.: CTTAAC   ENZIMA: HPA I   ARCHIVO: AR   1ER REG.
: 1   ULT. REG.: 76
[illegible] 130 134 348 379 497 503 580 591 772 800 838 858 1025 1180 1270 1369 1402 14[illegible]
1441 1460 1472 1509 1704 1735 1744 0 0 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
```

Fig.3.1.4 Resultados del estudio realizado con BUSEC. (secuencias 1 a 13).

```
... 517 1695 1742 ...
... NO.SEC: 14   SEC.: ...   ENZIMA: ...   ARCHIVO: ...
1ER. REG.: 1   ULT. REG.: 76
669 0 0 0 0 ...
0 0 0 0 0 0 0 NO.SEC: 15   SEC.: GGCGC   ENZIMA: PST 1   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 NO.SEC: 16   SEC.: GTCGAC   ENZIMA: SAL 1   ARCHIVO: AR   1ER REG..
1   ULT. REG.: 76
```

NO. BASE	NO. REG.	SECUENCIA	NO. EN REG.	SEC. ENZ
25	1	GAATTCAACGCGTTATCAAGAGAG	1	GAATTC
94	3	TGCGTCGAATAACACACAGGGGGC	22	GGGC
120	5	AGCGCCCATGGCCCCCGGCAGGCGC	10	GGCC
124	5	AGCGCCCATGGCCCCCGGCAGGCGC	14	GGCC *
340	14	TTTACCGTAAAGGCCGTATCCGTA	12	GGCC
399	16	TCGTCGCCCGCGCTGCCGGAGATGG	15	GGCC *
497	20	ACACCATTATGGCAGATGGCCGCGG	17	GGCC
503	24	AATCCGGCGGCCCCGAGCCAG	7	GGCC *
508	24	AATCCGGCGGCCCCGAGCCAG	12	GGCC
591	24	AATCCGGCGGCCCCGAGCCAG	15	GGCC *
772	32	CCTCCGGCGAAATGATGGCCATGT	4	GGCC *
800	33	ACCCGGCCAACAATATCTCCA.	0	GGCC
830	34	CGATCGTTAAATACGGCCAAATCCG	22	GGCC *
850	35	GCAAGGTGCGGCCTCCGGCCGCCTGA	18	GGCC
1025	42	ACCCCGGCCTGTAAACAGCCCAACG	17	GGCC
1108	49	CCCAAGAAAACGCGGCCGAGCAC	12	GGCC
1270	52	ATCCAGCAAGTCGTCCACGTGTTC	2[illegible]	GGCC *
1369	57	CCACCCGGCGTAATGACCGTACGC	1	GGCC *
1402	58	GGCTGCGGCCTACGGCCGCTTCCAAA	1[illegible]	GGCC *
1426	59	GGGCTCGTATTTGGCCCCATTAG	16	GGCC
1441	60	GATATGGCCCATATTTCGGACCGA	[illegible]	GGCC
1460	60	GATATGGCCCATATTTCGGACCGA	26	GGCC *
1472	61	CCGGCTGGCTGCGGCCAGTATTCC	6	GGCC
1509	62	CGCGCCCAACGACCCAACTACTAC	21	GGCC *
1704	71	CCGGTGGCCAACGCCAGCAGCAAG	6	GGCC
1735	72	GCGCTGGATAAACCGGTCATCCCG	7	GGCC *
1744	72	GCGCTGGATAAACCGGTCATCCCG	16	GGCC *
33	1	GAATTCAACGCGTTATCAAGAGAG	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	AGCGCCCATGGCCCCCGGCAGGCGC	2	GCGC
141	5	AGCGCCCATGGCCCCCGGCAGGCGC	21	GCGC
201	8	TACGTCCTGCGCGGCGACAAATAA	9	GCGC
202	8	TACGTCCTGCGCGGCGACAAATAA	10	GCGC *
331	13	CCATGACCATGCGTCAATCCGCTA	19	GCGC
392	16	TCGTCGCCCGCGCTGCCGGAGATGG	8	GCGC *
393	16	TCGTCGCCCGCGCTGCCGGAGATGG	9	GCGC
500	20	ACACCATTATGGCAGATGGCCGCGG	20	GCGC *
569	23	GCTAGCGCGATGTGGCCTGCG	17	GCGC
607	25	TCGGCTGGCGCGGACGCGCGGTGA	7	GCGC
608	25	TCGGCTGGCGCGGACGCGCGGTGA	8	GCGC *

Fig.3.1.4 Resultados del estudio realizado con BUSEC. (secuencias 14 a 16 - primera presentación, 8 a 10 - segunda presentación).

735	51	TGCGCATGCGAATCCCCAAAGCA	12	GCGGC *
747	52	ACGCCTCCAACAATATCTCCA	5	GCGGC *
848	55	GCAAGCTGTGCCTCGACCCCCTGA	8	GCGGC
854	55	TGATCCACTTTGTTCTCCCCACA	16	GCGGC *
921	58	ACATCCTCCAGCCCCCCGACATCC	11	GCGGC
922	58	ACATCGTCCAGCCCCCCGACATCC	12	GCGGC *
923	58	ACATCGTGCAGCCCCCCGACATCC	13	GCGGC
924	58	ACATCGTGCAGCCCCCCGACATCC	14	GCGGC *
1050	63	ACCAATACCCCACCCTCCCCCAGA	16	GCGGC
1105	69	CCCAACAAAACCCCCCCCAGCAC	9	GCGGC *
1240	51	ACCCCCCAACCTAATCTCCCCCTG	16	GCGGC
1270	53	CCCCAAACCCCCCCCAAAACACCCC	6	GCGGC *
1279	53	CCCCAAACCCCCCCCAAAACACCCC	7	GCGGC
1280	53	CCCCAAACCCCCCCCAAAACACCCC	8	GCGGC *
1290	53	CCCCAAACCCCCCCCAAAACACCCC	10	GCGGC
1356	57	CCACCCCCCCCTAATCACCCTACCCC	18	GCGGC
1393	58	CCCTCCCCCTACCCCCCTTCCAAA	1	GCGGC *
1464	61	CCCCCTCCCTCCCCCCAGTATTCC	20	GCGGC
1465	61	CCCCCTCACTCCCCCCAGTATTCC	21	GCGGC *
1547	65	CATTTTCACCACCCCCACATCCTC	7	GCGGC
1548	65	CATTTTCACCACCCCCACATCCTC	8	GCGGC *
1697	70	CCCCTCATCCCTCATCATATCAGC	17	GCGGC
1698	70	CCCCTCATCCCTCATCATATCAGC	18	GCGGC *
1723	71	CCCCTCCCCAACCCCAGCACCAAG	19	GCGGC
1765	73	CTACCCTCCCAACCCTTTCCCCCC	13	GCGGC *
1812	75	ATCCCCAACCACCTCCTCCCCCAC	12	GCGGC
1813	75	ATCCCCAACCACCTCCTCCCCCAC	13	GCGGC *
144	6	AATTCTTCTCTTTCCCACATTTCC	0	CTTAAC *
94	3	TCCCTCCAATAACACACACCCCCC	22	CCCC *
130	5	ACCCCCCATCCCCCCCCCACCCCC	10	CCCC *
134	5	ACCCCCCATCCCCCCCCCACCCCC	14	CCCC
348	14	TTTACCCTAAACCCCCTATCCCTA	12	CCCC *
399	16	TCCTCCCCCCCCCTCCCCCACATCC	15	CCCC
497	20	ACACCATTATCCACACATCCCCCCLC	17	CCCC *
583	24	AATCCCCCCCCCCCCCAGCCCAC	7	CCCC
588	24	AATCCCCCCCCCCCCAGCCCAC	12	CCCC *
591	24	AATCCCCCCCCCCCCAGCCCAC	15	CCCC
772	32	CCTCCCCCCAAATCATCCCCATCT	4	CCCC
800	33	ACCCCCCCCAACAATATCTCCA	8	CCCC *
836	34	CCATCCTTAAATACCCCAAATCCC	22	CCCC
858	35	CCAACCTCCCCCTCCCCCCCTCA	18	CCCC * 24
1025	42	ACCCCCCCTCTAAACACCCCAACC	17	CCCC *
1168	49	CCCAACAAAACCCCCCCCAGCAC	12	CCCC *
1270	52	ATCCACCAACTCCTCCAGGTCTTC	22	CCCC 0*
1269	57	CCACCCCCCCTAATCACCCTACCC	1	CCCC
1402	58	CCCTCCCCCTACCCCCCTTCCAAA	10	CCCC
1426	59	CCCCTCCTATTTCCCCCATTAC	10	CCCC *
1461	60	CATATCCCCCATATTTCCCACCCA	1	CCCC *
1460	60	CATATCCCCCATATTTCCCACCCA	20	CCCC
1472	61	CCCCCTCACTCCCCCCAGTATTCC	8	CCCC *
1507	62	CACCCCCAACCACCCAACTACTAC	21	CCCC
1704	71	CCCCTCCCCAACCCCAGCACCAAG	0	CCCC *
1735	72	CCCCTCCATAAACCCCTCATCCCC	7	CCCC
1734	72	CCCCTCCATAAACCCCTCATCCCC	16	CCCC

Fig.3.1.4 Resultados del estudio realizado con BUSEC (secuencias 10 a 13 - segunda presentación).

```
      MATRIZ DE COORDENADAS:
0 0 0 0 0 0 0 0 0 0 NO.SEC: 17   SEC.: CCCGGG   LONGNA: TCA I   ARCHIVO: AA   11
N REG.: 1   ULT. REG.: 76
517 535 673 922 1003 1104 1524 1529 1596 1605 NO.SEC: 18   SEC.: TCCA   LONGNA:
      ARCHIVO: AA   1ER REG.: 1   ULT. REG.: 76
0 0 0 0 0 0 0 0 0 0 NO.SEC: 19   SEC.: CTCCAG   LONGNA: TRO I   ARCHIVO: AA   11
N REG.: 1   ULT. REG.: 76
999 627 0 0 0 0 0 0 0 0 NO.SEC: 20   SEC.: CCTCGG   LONGNA:  _T II   ARCHIVO: AA
   1ER REG.: 1   ULT. REG.: 76
622 0 0 0 0 0 0 0 0 0 NO.SEC: 21   SEC.: TGATCA   LONGNA: TC. I   ARCHIVO: AA
1ER REG.: 1   ULT. REG.: 76

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

  517         21      GAGTCGGCTCGCTCCAGCAGCCTCG              13          TCCA
  535         22      AAGTCGAAGAGCTCCTCGAAA          7               TCCA
  673         28      TCCATTTCCTCTTCTATCACGTGC               1          TCCA
  922         38      TCCCCGAAAAAGCTCCCTACCCAGA              10         TCCA x
 1003         41      CCCCCATCAACCCTATCCAGT          19              TCCA
 1104         44      ACCTCCAATCCCTCCTCCATCAGT               6          TCCA x
 1524         63      ACCCCAGTCAGCCCCCTCCATAGC               12         TCCA
 1529         63      ACCCCAGTCAGCCCCCTCCATAGC               17         TCCA x
 1596         66      TTCCCCCCCCATAAAAAAGCTCAGC              12         TCCA x
 1605         66      TTCCCCCCCCATAAAAAAGCTCAGC              21         TCCA x


  999         20      ACACCCATTATCCCACATCCTCCCCC              19         CCTCGG
  627         27      AAGAAGCCCCCTACCCAGCACG          9               CCCCCC x

  622         25      TCCCCTCCCCCCCAACCCCCCTCATA.            22         TCATCA


  126          5      ACCCCCCATCCCCCCCCCCCAGCCCC             6          TCTCCC -
  927         38      TCCCCCAAAAAGCTCCCCTACCCACA             15         CCTCCC
 1095         45      TCCCCACCCCCCTCCACCCATCCATC             15         CCTCCC *
 1716         72      CCCCCTCCATAAACCCCTCCATCCCC             18         CCTCCC


  469         27      AAGAAGCCCCCTACCCAGCACG         21              CTCCAG x
```

Fig.3.1.4 Resultados del estudio realizado con BUSEC (secuencias 17 a 21).

B) Búsqueda de zonas de ácidos nucleicos participantes en la regulación de glutamino sintetasa en E.coli.

La glutamino sintetasa es una enzima cuya función es central en el metabolismo nitrogenado de E. coli. (24), y por tanto se encuentra altamente regulada. De ahí la importancia de conocer en detalle la región de control del gene que la codifica. En esta zona se habían encontrado 3 probables sitios de iniciación de transcripción (operadores), dos de los cuales se localizan muy lejos de la iniciación de traducción. Esto hizo pensar en la posible existencia de una región atenuadora en este fragmento de A.D.N. Aunque los metodos genéticos no sugerían esto, se requiere un estudio más detallado para confirmarlo. En la figura 3.2.1 se pueden observar los posibles sitios de reconocimiento para iniciación de transcripción dentro de cajas y el sitio de anclaje ribosomal con puntos sobre las bases. Si existiese una regulación de tipo atenuación como en el operón de triptofano, (26) debería haber alguno o todos los componentes siguientes: 1) codificación para un péptido iniciado después de los posibles promotores más lejanos, 2) dos o más codones para glutamina en tandem y en fase con el péptido hipotético y 3) estructuras secundarias de tipo gaza sobrepuestas en la misma zona.

Para detectar el péptido y los codones en tandem se hizo una traducción de la zona en tres fases con el programa TRAD (Fig. 3.2.2). TRAD utiliza una matriz tridimensional para traducir secuencias de A.D.N. o A.R.N. a secuencias de aminoácidos en las posibles tres fases. Las coordenadas de la matriz corresponden al valor ASCII de los tres componentes del triplete; cada cubo es la abreviatura del aminoácido determinado por el triplete (Fig.3.2.3). Por ejemplo, el codón ATG codifica para el aminoacido metionina. En la matriz tridimensional la abreviatura "MET" se puede localizar en el punto 65,84,71. Para ocupar menos memoria, a cada uno de los índices se le resta 65 ya que éste es el valor ASCII mínimo correspondiente a una letra, la 'A'. Entonces los verdaderos índices son 0,19,6.

Para buscar estructuras secundarias se usó el programa BUPAL. Este programa permite visualizar con facilidad secuencias invertidas repetidas (palindromas) dentro de A.D.N. y A.R.N. y da todas las secuencias mayores a 4 bases con sus respectivas posiciones. El programa consiste en hacer comparaciones de la secuencia de interés contra su secuencia complementaria generada en dirección opuesta.

BUPAL es una derivación de HOMOLOG que genera una tabla donde se ven todas las posibles combinaciones de apareamientos dentro de una misma cadena. Revisando esta tabla primaria (antes de filtrar las secuencias menos significativas), se puede observar (Fig. 3.2.4) que se generan dos ejes palindrómicos (lineas solidas) en sentido

```
                    20                  40                  60                  80                 100

CCGTCAAATGCGTCAACCACCAGCAGCACTGAGTCTACCATGGACATTACACGTTCAACTTCACCACCGAAGTCGGCGTGCCCGGGGGTATCAACGATGT
GGCAGTTTACGCAGTTGGTGGTCGTCGTGACTCAGATGGTACCTGTAATGTCGAAGTTGAAGTGGTGGCTTCAGCCGCACGGGCCCCCATAGTTGCTACA
                   120                 140                 160                 180                 200

TGATACGGTAATCATTCCATTTGATAGCGGTGTTTTTCGCGAGGATGGTAATCCACGCTCTTTCTCCAAATCGTTGGAGTCCATCACGCGCTCTTGGGTT
ACTATGCCATTAGTAAGGTAAACTATCGCCACAAAAAGCGCTCCTACCATTAGGTGCGAGAAAGAGGTTTAGCAACCTCAGGTAGTGCGCGAGAACCCAA
                   220                 240                 260                 280                 300

TCGGCACGAGAGTCGAACGTACCGGATTGTTGGAGCAGCTTGTCTACCAGGGTGGTTTTACCATGGTCTACGTGCGCGATGATGGCGATATTCAGCAATT
AGCCGTGCTCTCAGCTTGCATGGCCTAACAACCTCGTCGAACAGATGGTCCCACCAAAATGGTACCAGATGCACGCGCTACTACCGCTATAAGTCGTTAA
                   320                 340                 360                 380                 400

TTTCCGATCACAACTTTGCCTCAGGCATTAGAAATAGCGCGTTATTGTAATACGGATTAATCGCACTACAAAACAGGATCACAAACATCCTCCGCAAACA
AAAGGCTAGTGTTGAAACGCAGTCCGTAATCTTTATGCCGCAATAACATTATGCCTAATTAGCGTGATGTTTTGTCCTAGTGTTTGTAGGAGGCGTTTGT
                   420                 440                 460                 480                 500

ATATTGCAGAGTCCCTTTGTGATCGCTTTCACGGAGCATAAAAAGGGTTATCCAAAGGTCATTCGACCAACATGGTGCTTAATGTTTCCGATTGAAGCAC
TATAACGTCTCAGGGAAACACTAGCGAAAGTGCCTCGTATTTTTCCCAATAGGTTTCCAGTAAGCTGGTTGTACCACTAATTACAAAGGCTAACTTCGTG
                   520                 540                 560                 580                 600

TATATTGGTGCAACATTCACATCGTGGTGCAGCCCTTTTGCACGGATGGTTGCGCATGATAACGCCTTTTAGGGGCAATTTAAAAGTTGGCACAGATTTC
ATATAACCACGTTGTAAGTGTAGCACCACGTCGGGAAAACGTGCCTACCAACGCGTACTATTGCGGAAAATCCCCGTTAAATTTTCAACCGTGTCTAAAG
                   620                 640                 660                 680

GCTTTATCTTTTTTACGCGACACGGCCAAGATAATTGCAGATTTCGTTACCACGACGACCATGACCAATCCAGGAGAGTTAAAGT ATG TCC GCT GAA
CGAAATAGAAAAAATGCGCTGTGCCGGTTTTATTAACGTCTAAAGCAATGGTGCTGCTGGTACTGGTTAGGTCCTCTCAATTTCA TAC AGG CGA CTT
                   700                 720                 740                 760

                                                                                 met ser ala glu

his val leu thr met leu asn glu his glu val lys phe val asp leu arg phe thr asp thr lys gly lys glu
CAC GTA CTG ACG ATG CTG AAC GAG CAC GAA GTG AAG TTT GTT GAT TTG CGC TTC ACC GAT ACT AAA GGT AAA GAA
GTG CAT GAC TGA TAC GAC TTG CTC GTG CTT CAC TTC AAA CAA CTA AAC GCG AAG TGG CTA TGA TTT CCA TTT CTT
                   780                 800

gln his val thr ile pro ala his gln val asn ala glu phe
CAG CAC GTC ACT ATC CCT GCT CAT CAG GTG AAT GCT GAA TTC
GTC GTG CAG TGA TAG GGA CGA GTA GTC CAC TTA CGA CTT AAG
```

Fig. 3.2.1 Secuencia anterior al gene glnA indicando sitios de iniciación de transcripción y traducción.

Fig. 3.2.2 Traducción en tres fases de región de control de clrA.

```
1
THR LEU THR LEU LEU ASP TRP SER TRP SER SER TRP FIN ARG ASN LEU GLN LEU PHE TR
PRO CYS ARG VLN LYS LYS ILE LYS ARG ASN LEU CYS GLN LEU LEU ASN CYS PRO FIN LYS
ALA LEU SER CYS ALA THR ILE ARG ALA LYS GLY LEU HIS HIS ASP VLN ASN VLN ALA PRO
ILE FIN CYS PHE ASN ARG LYS HIS FIN ALA PRO CYS TRP SER ASN ASP LEU TRP ILE THR
LEU PHE MET LEU GLU SER ASP HIS LYS GLY THR LEU GLN TYR CYS LEU ARG ARG MET PHE
VLN ILE LEU PHE CYS SER ALA ILE ASN PRO TYR TYR ASN ASN ALA LEU PHE LEU MET PRO
GLU ALA LYS LEU FIN SER

LEU FIN LEU SER TRP ILE GLY HIS GLY ARG ARG GLY ASN GLU ILE CYS ASN TYR PHE GLY
ARG VLN ALA FIN LYS ARG FIN SER GLU ILE CYS ALA ASN PHE FIN ILE ALA PRO LYS ARG
ARG TYR HIS ALA GLN PRO SER VLN GLN LYS GLY CYS THR THR MET FIN MET LEU HIS GLN
TYR SER ALA SER ILE GLY ASN ILE LYS HIS HIS VLN GLY ARG MET THR PHE GLY FIN PRO
PHE LEU CYS SER LYS ALA ILE THR LYS GLY LEU CYS ASN ILE VLN CYS GLY GLY CYS LEU
FIN SER CYS PHE VLN VLN ARG LEU ILE ARG ILE THR ILE THR ARG TYR PHE FIN CYS LEU
ARG GLN SER CYS ASP ARG

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

```
TRADUCCION DE SECUENCIA: TTTCCGATCACAACTTTGCCTCAGGCATTAGAAATAGCGCGTTATTGTAATACGC
ATTAATCGCAGTACAAAACAGGATCACAAACATCCTCCGCAAACAATATTGCAGAGTCCCTTTGTGATCGCTTTCACGCA
GCATAAAAATGTTATCCAAGGTCATTCGACCAACATCGTGCTTAATGTTTCCGATTGAAGCACTATATTGGTGCAACA
TTCACATCTTGTGCAGCCTTTTGCACGGATGGTTGCCCATCATAACGCCTTTTAGGGCAATTTAAAAGTTGGCACAG
CTTTCGTTTATCTTTTTTACGCGACACGGCCAAAATAATTGCAGATTTCGTTACCACGACGACCATGACCAATCCAGGA
AGTTAAAT
```

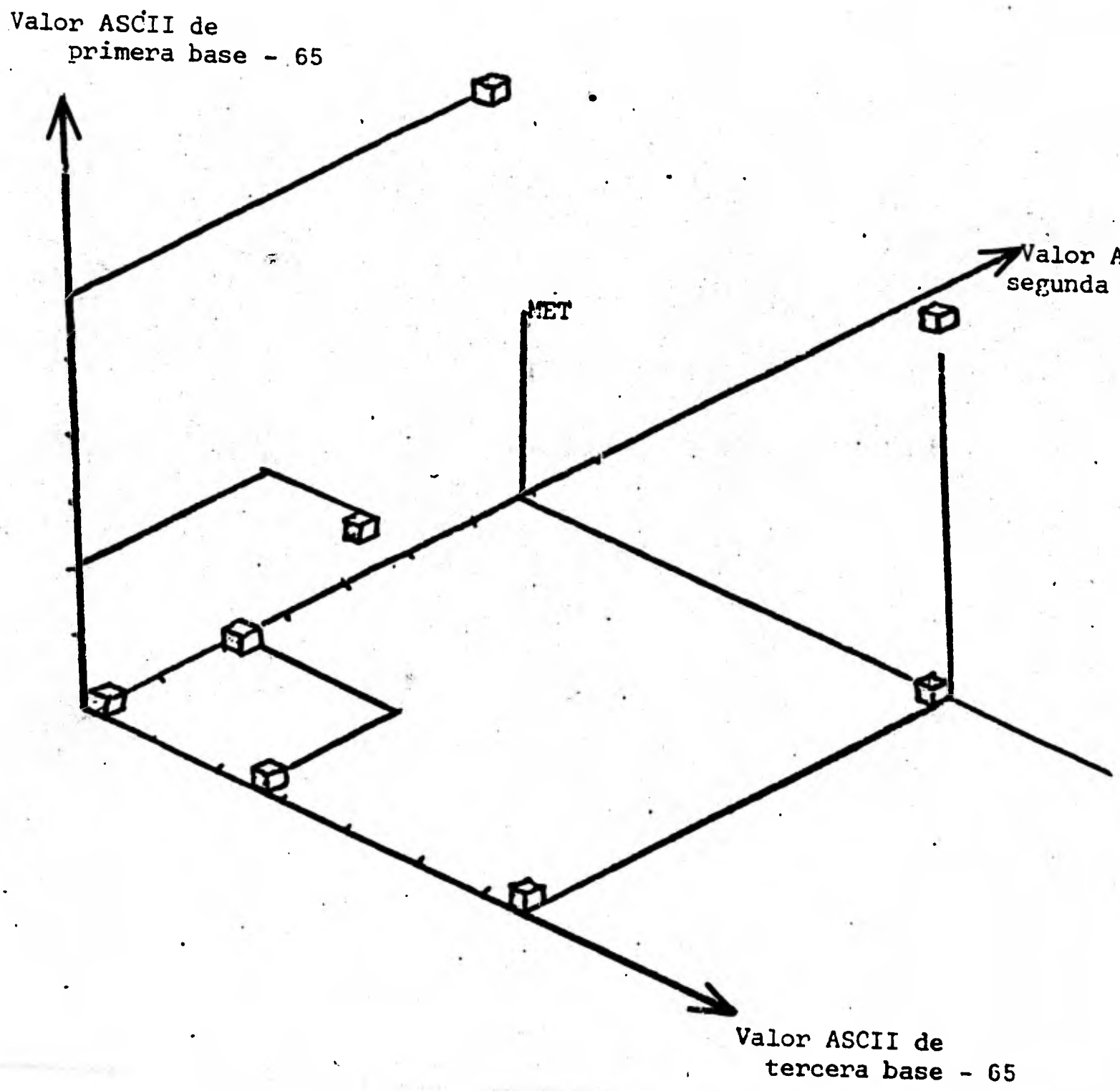


Fig. 3.2.3 Espacio cartesiano representando el método de traducción de tripletes a aminoácidos.

vertical y en forma diagonal de derecha a izquierda. Al principio del proceso de deslizamiento de una secuencia con respecto a la otra y al agregar las últimas bases de una de las secuencias al comienzo de la misma, se pueden observar palindromas cortos en las dos mitades de la matriz separados por la linea segmentada. Conforme se ejecute el programa, los palindromas de la derecha de la secuencia ocupan la mayor parte de la tabla hasta llegar al punto en que se compara toda la secuencia contra toda su complementaria sin agregar bases a los extremos (renglones 63 y 64 de la fig. 3.2.4). En estos renglones se obtienen los palindromas de mayor longitud. Teniendo esto en mente, el análisis de una secuencia mas larga que el número de carácteres por renglón (en este caso 120) 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.

El gran numero de posibles apareamientos internos sugiere como se pueden generar estructuras secundarias de gran complejidad y con funcion regulatoria en la maquinaria de replicación, trancripción, y traducción.

Se buscaron, utilizando BUPAL, posibles estructuras secundarias, la mas importantes se esquematizan en la figura 3.2.7. Esta zona forma dos gazas alternativas. La primera (obtenida a partir de la tabla de la figura 3.2.6) tiene mayor probabilidad de formarse ya que es termodinámicamente más estable.

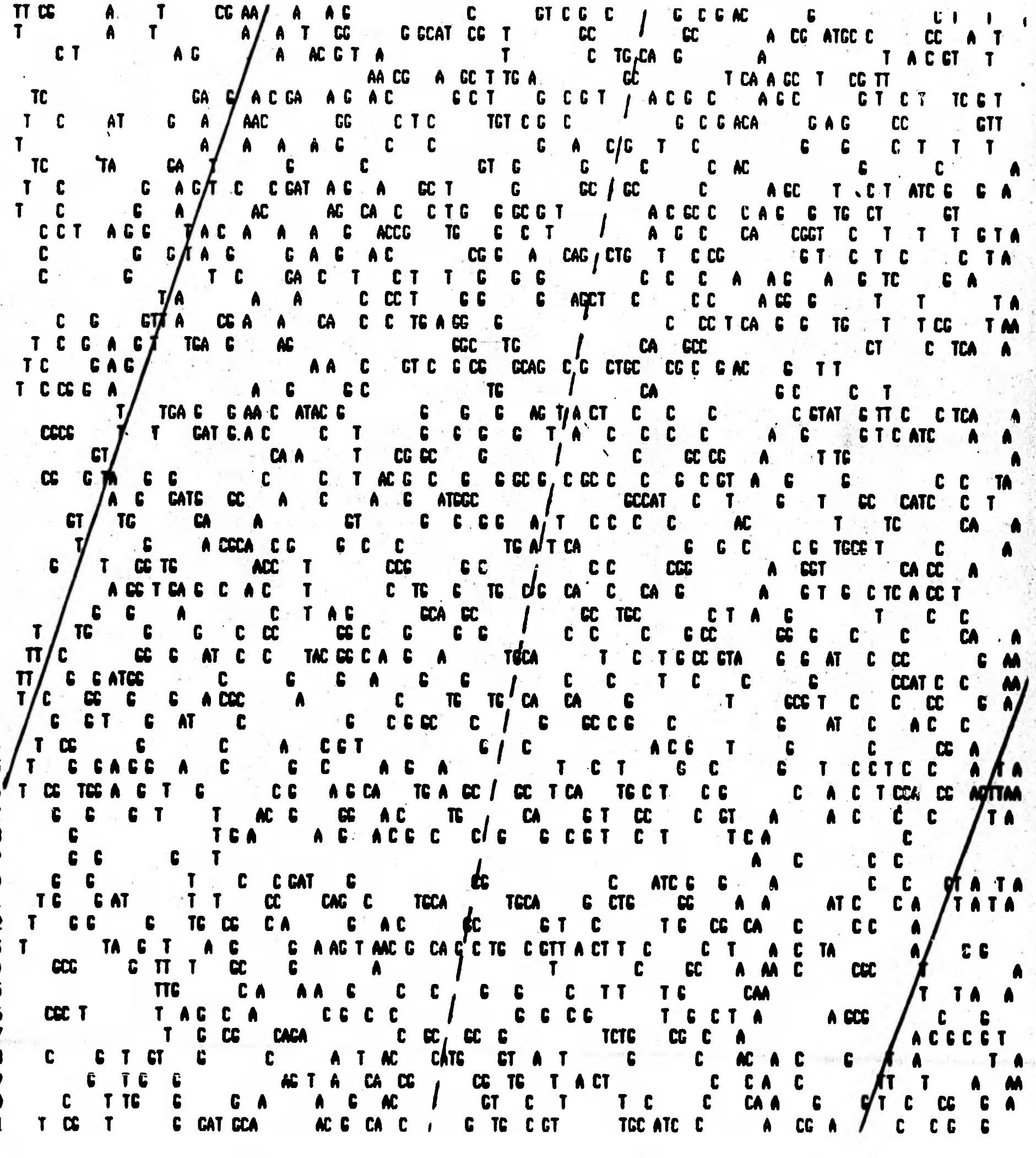


Fig. 3.2.4 Tabla generada por BUPAL con eje palindrómico indicado por línea segmentada. Cada renglón corresponde a un deslizamiento.

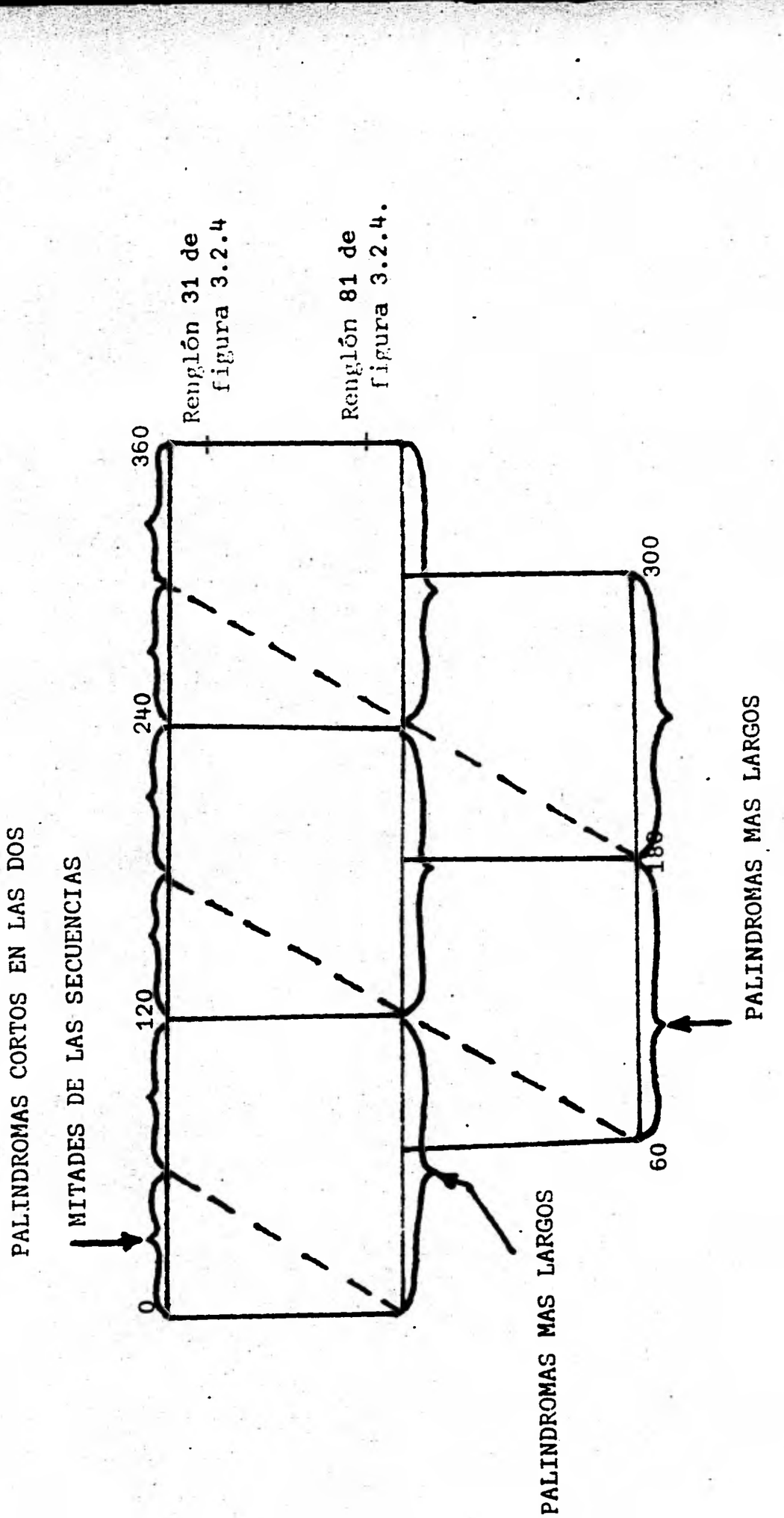


Fig. 3.2.5 Optima estructura para la búsqueda de secuencias repetidas invertidas en un segmento de más de 120 pares de bases. Cada uno de los cuadros corresponde a una tabla completa generada por BUPAL. Los números superiores e inferiores son los números de base. Ejes de simetría para deteccion dé palindromas en matriz generada por BUPAL

DIAGONALES

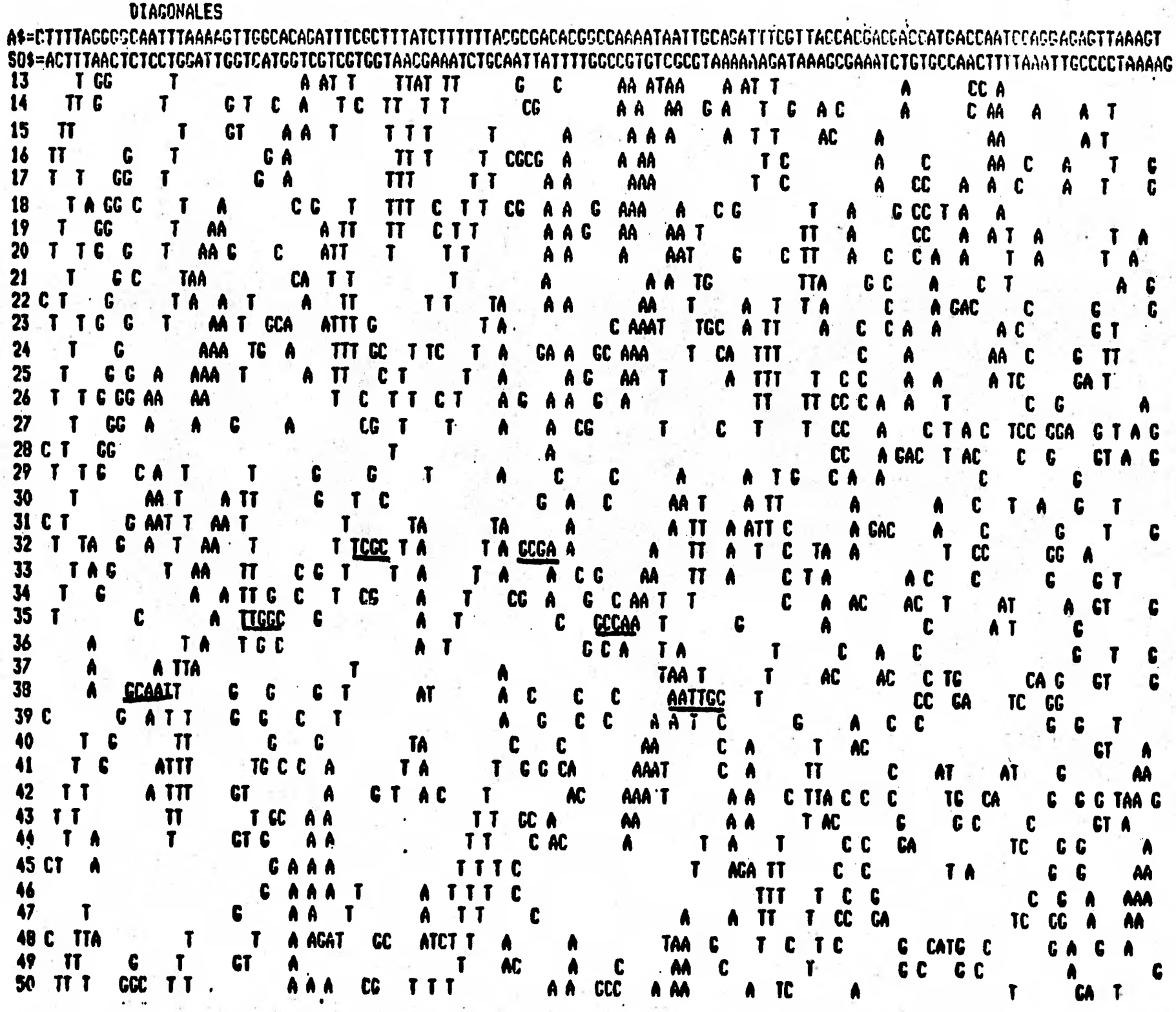
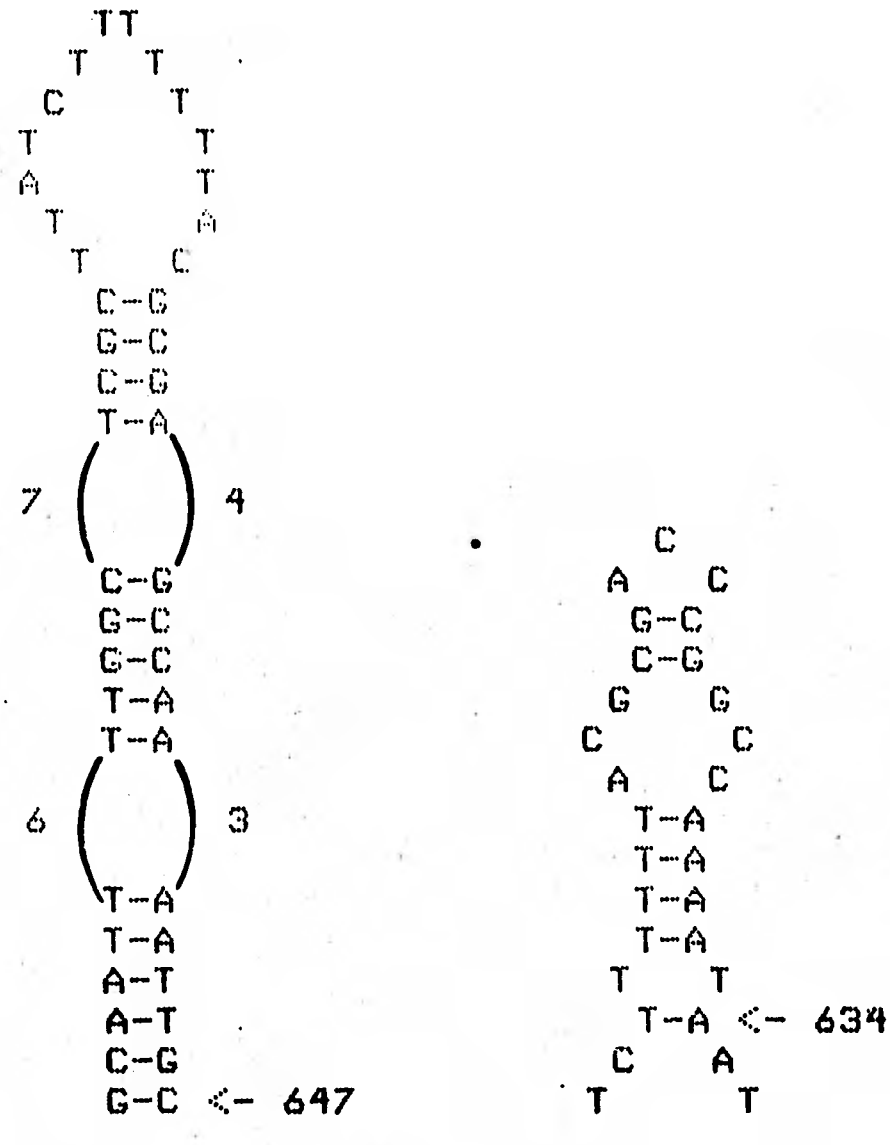


Fig. 3.2.6 Comparación de secuencias palindrómicas dentro de la región de control del gira de E.coli. (bases 565 a 695).



3.2.6 Dos posibles estructuras secundarias encontradas por BUPAL en la región de control de glnA. Los números superiores indican el número de bases entre los grupos de apareamientos. Los inferiores son la posición dentro del segmento secuenciado de la figura 3.2.1.

Al revisar la traducción de la zona se encontró que los péptidos 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 las conformaciones clásicas del mecanismo de atenuación. Podrían, sin embargo, funcionar como sitios de reconocimiento para proteínas operadoras o activadoras.

Este estudio se realizó usando el programa COMP sobre la secuencia complementaria a la introducida por el investigador. COMP consulta una tabla en la cual se encuentra la letra correspondiente a la base complementaria. El índice de la tabla esta dado por el código ASCII de las las letras 'A', 'C', 'G', Y 'T', o sea: 65, 67, 71, 84. Conforme se consulta el índice, se genera la secuencia complementaria.

C) Comparación entre cuatro secuencias de glutamato deshidrogenasa Nadph dependientes y la determinación de distancias mínimas mutacionales

Haciendo estudios sobre las secuencias de proteínas equivalentes en diferentes especies se ha visto una correlación con las clasificaciones hechas por taxónomos clásicos (8,9). En especial se han analizado las estructuras del citocromo C y de la histona H1. En este proyecto se 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 aminoacidos de la proteina de Neurospora, 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).

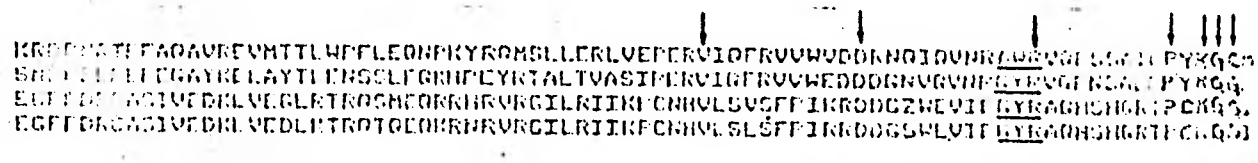


Fig. 3.3.1 Coincidencias dentro de 80 elementos de GDH nadph dependientes de E.coli, N.crassa, pollo y bovino.

Para detectar homologías entre las proteínas se hizo una clasificación de las posibles similitudes entre secuencias de la manera siguiente:

A) Homología absoluta
B) Cambio de una base
C) Estructura o polaridad similar

Una comparación de tipo absoluto implica que tiene que haber el mismo amino ácido en ambas secuencias. Por ejemplo tenemos que entre las secuencias de Escherichia coli y de Neurospora crassa se encontró, como indica la figura 3.3.2 una similitud cercana al 50% y de alrededor de 17% entre bovino y E. coli.

```
BUSQUEDA DE HOMOLOGIA

ABSOLUTO

 A$=KRDPNQTEFAQAVREVHTTLWPFLEQNPKYRQHSLLERLVEPERVIQFRVVWVDDPNQIQVNRAWRVQFSSAIGPYKGGHPFHPSVN
SO$=EGFFDRGASIVEDKLVEGLRTRQSMEQRRHRVRGILRIIKPCNHVLSVSFPIKRDDGZWEVIEGYRAQHSHQRTPCKGGIRYSLDVSVDEVK
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$=KRDPNQTEFAQAVREVHTTLWPFLEQNPKYRQHSLLERLVEPERVIQFRVVWVDDPNQIQVNRAWRVQFSSAIGPYKGGHPFHPSVN
SO$=SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI
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 continuación se buscaron coincidencias de aminoácidos que diferían en una sola base de su codón de traducción. Para ello se construyó una matriz en la cual se viera esta relación. 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 última base. Los codones en un mismo renglón difieren en la base central y todos los codones que se encuentran en la misma posición dentro de los diferentes cuadros pero en la misma columna difieren solo en la primera base.

	U		C		A		G	
U	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
C	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
A	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
G	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 aminoácidos usada para determinar diferencias de una base.

La matriz que contiene la relación entre los diferentes aminoácidos se muestra en la figura 3.3.4. (12).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	0	1	0	0	0	1	1	0	0	1	1	0	1	1	1	0	1	1	0
R	0	0	1	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1
N	0	0	1	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	1
D	1	0	1	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	1
C	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	1
Q	0	1	0	0	0	0	1	1	0	0	1	0	1	1	0	0	1	0	0
E	1	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
H	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1
I	0	1	1	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0
L	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	1	0	1	1
K	0	1	0	0	0	0	1	0	0	0	1	1	1	0	1	1	1	1	0
M	0	1	1	0	0	0	1	1	0	0	0	0	0	1	1	1	1	0	0
F	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0
P	1	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	1	1
S	1	1	1	0	0	1	0	0	0	1	0	1	1	0	0	1	1	1	1
T	1	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	0
W	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	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 muestra los aminoácidos que difieren en solo una base.

Tomando en cuenta esta matriz, el porcentaje de homología entre E.coli y N. crassa aumenta al 74% como se ve en la figura 3.3.5.

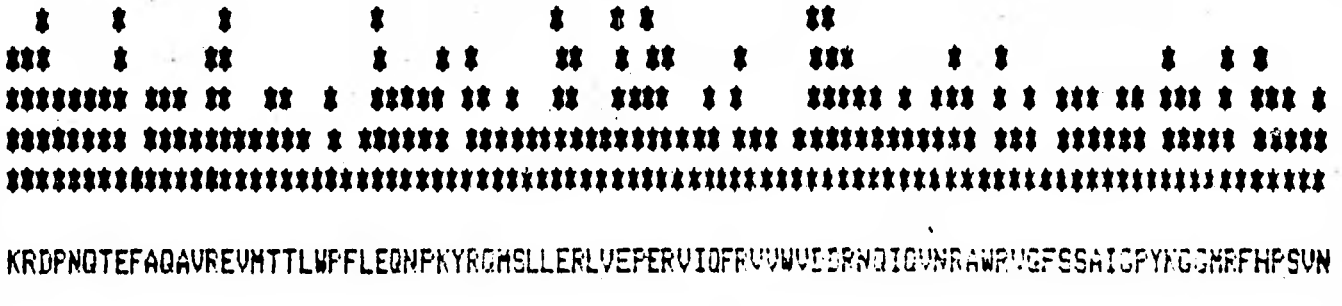
```
A$=KRDPNQTEFAQAVREVMTTLWPFLEQNPKYRQMSLLERLVEPERVIQFRVVWVDDRNQIQVNRAWRVQFSSAIGPYKGGMRFHPSVN
SO$=SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI
1)      PNQT  AQ VRE M    PF   NP  RQ      L EP    QF  V VD RN     RA RV     IGP    M  M S
2)   DP      Q V EVMT  W  L Q PKY 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   WPFLEQNPKYR  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
```

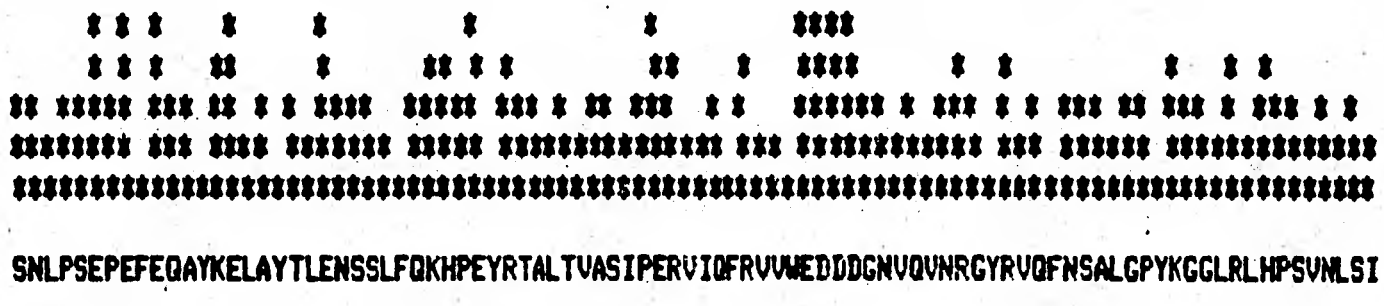
3.3.5 Comparación entre N.crasa y E.coli considerando el cambio de una base.

E COLI



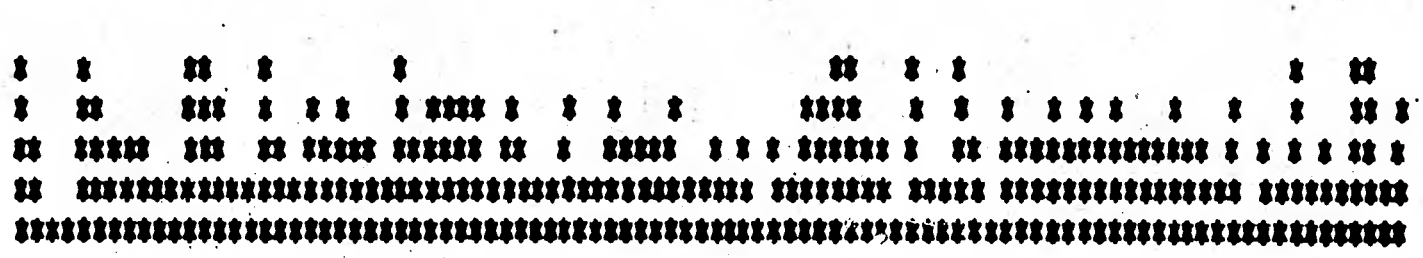
KRDPNQTEFAQAVREVMTTLWFFLEQNPKYRQMSLLERLVEPERVIQFRVVWVDDRNQIQVNRAWRVQFSSAIGPYKGGMRFHPSVN

N.CRASSA



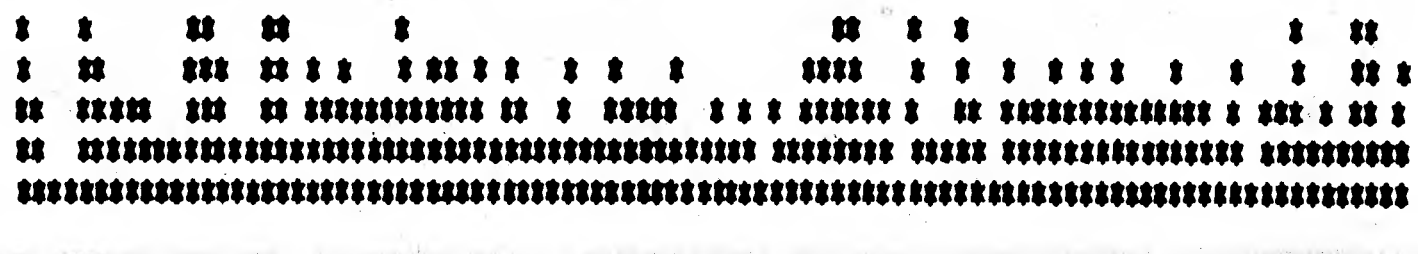
SNLPSEPEFEQAYKELAYTLENSSLFQKHPEYRTALTVASIPERVIQFRVVWEDDDGNVQVNRGYRVQFNSALGPYKGGLRLHPSVNLSI

POLLO



EGFFDRGASIVEDKLVEGLRTRQSMEQRRHRVRGILRIIKPCNHVLSVSFPIKRDDGZWEVIEGYRAQHSHQRTPCKGGIRYSLDVSVDEVK

BOVINO



EGFFDRGASIVEDKLVEDLKTRQTQEQKRNRVRGILRIIKPCNHVLSLSFPIRRDDGSWEVIEGYRAQHSHQRTPCKGGIRYSTDVSVDEVK

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 posición 53 se observa un pico de acidez.

Para detectar características químicas seleccionadas por presiones funcionales a nivel de estructura primaria GRAFPOL presenta un histograma donde la longitud de las barras corresponde a la hidrofobicidad de cada uno de los aminoácidos de la secuencia peptídica.

La clasificación es la siguiente:

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

En la figura 3.3.6 se pueden apreciar los resultados obtenidos del análisis de las secuencias de Glutamato deshidrogenasa de las cuatro ramas filogenéticas. Se detectó un pico de región acida seguido por otro de hidrofobicidad cíclica para finalizar con otra ácida cerca de la posición 53.

Con el criterio de hidrofobicidad/hidrofilicidad (valores de uno a cinco) encontramos una homología del 61% entre 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 AVR    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 Comparación entre E.coli y N.crassa tomando en cuenta la polaridad de los aminoácidos.

Es importante considerar ambos tipos de similitudes ya que una homología puede implicar tanto evolución convergente como divergente dependiendo de la presión funcional para una característica definida.

Para encontrar una relación filogenética mas clara se utilizó un sistema similar al que se usó con citocromo C. Se requiere de la matriz que se muestra en la figura 3.3.8. La matriz es similar a la publicada por Fitch y Margoliash (19) y da la distancia mínima mutacional entre los diferentes aminoácidos. Esta se define como el número de bases que requieren ser sustituídas para que una secuencia se transforme en otra. Por ejemplo tenemos que para que un codón que codifica para alanina como GCG pase a ser uno que codifique para glicina (GGG) la modificación mínima seria en la segunda base y por tanto la distancia mínima 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 que indica distancias mínimas mutacionales entre los diferentes aminoácidos.

Se determinaron las distancias mínimas mutacionales entre cada una de las cuatro secuencias de la figura 3.3.9. Para ello se llevaron a cabo los siguientes pasos:

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

La necesidad de efectuar esto implica que ocurrieron inserciones y deleciones (omisiones de parte de la secuencia) a lo largo de la evolución.

II) Usar la tabla de la figura anterior.

En la siguiente figura se presentan las distancias mínimas de mutación obtenidas.

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

Fig. 3.3.9 Distancias mínimas 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 el eucariote primitivo está mas cercano filogenéticamente a los procariotes que a los eucariotes modernos.

Con estos datos se generó un árbol filogenético muy parecido al clásico..

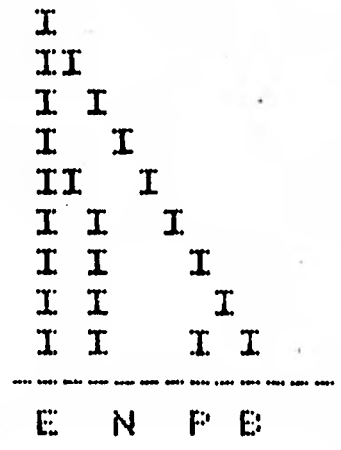


Fig. 3.3.10 Dendrograma obtenido apartir de de las distancias mutacionales.

En el dendrograma resultante (fig. 3.3.10) las distancias mínimas mutacionales corresponden a la suma de las longitudes de las ramas. Así 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 representación se debe tomar con mucha cautela por los siguientes motivos:

a) Se está analizando una sola proteína.

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

c) No podemos definir la historia verdadera de la divergencia debido a la degeneración del código genético por lo que se está tomando la distancia mínima evolutiva y no la real.

d) Ciertas partes de la estructura tienden a conservarse más que otras. (Entre Neurospora, pollo, y bovino existen 21 glicinas en común implicando posiblemente una estructura de tipo alfa-hélice).

e) La velocidad de divergencia varía en diferentes ramas. (Aumenta con la diminución en el tiempo de generación).

Algunas de estas objeciones son válidas también en el modelo de Fitch y Margoliash como ellos mismos lo mencionan (9).

COMENTARIOS Y PRESPECTIVAS:

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

B. Sería interesante aplicar el programa MMD, que representa relaciones arbóreas, a las secuencias conductuales. Con él posiblemente se detectarían categorias dentro de los diferentes tipos de ratones análogas a una relación filogenética. Para aplicarlo, el programa se usaría básicamente 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 cúmulos —usando uno de los siguientes criterios: precedencia, gasto energético, 'gasto emocional', o morfología de cada una de las unidades conductules.

C. La Programación Dinámica es una de las técnicas utilizadas en el reconocimiento de formas (12) que fue desarrollada en el campo de la investigación de operaciones. Se basa en comparar todos los puntos de una gráfica contra todos los relevantes de una segunda forma generando un valor que debe ser minimizado usando probabilidad y estadística. La figura 4.1 indica que haciendo una comparación punto a punto entre dos fonogramas de las palabras 'MASSES' y 'MASHES' la computadora rechaza la tesis de que son homólogas. En cambio al usar la técnica, el procesador puede declarar a ambos patrones como correspondiendo a la misma palabra pero enunciada a velocidades diferentes.

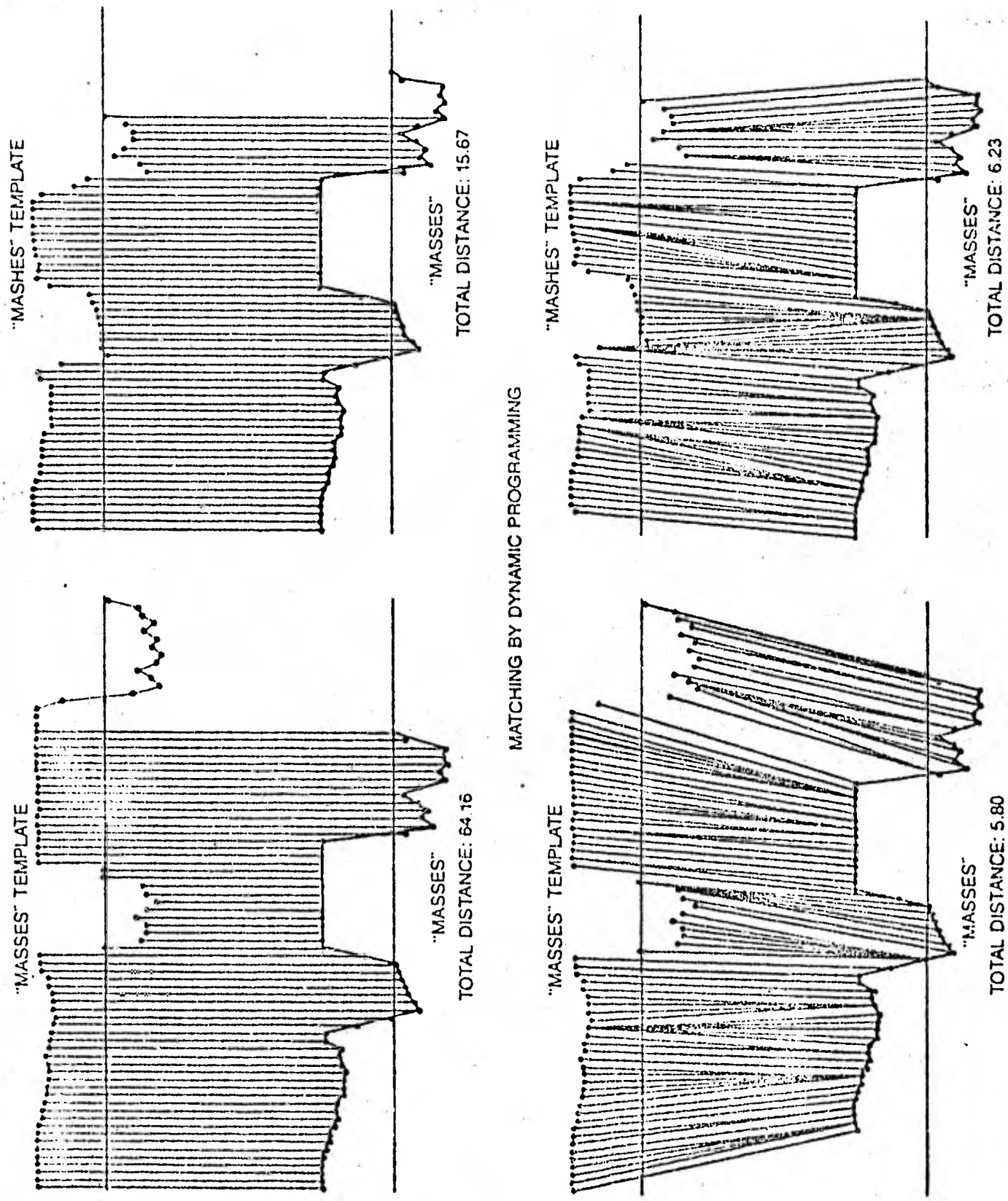
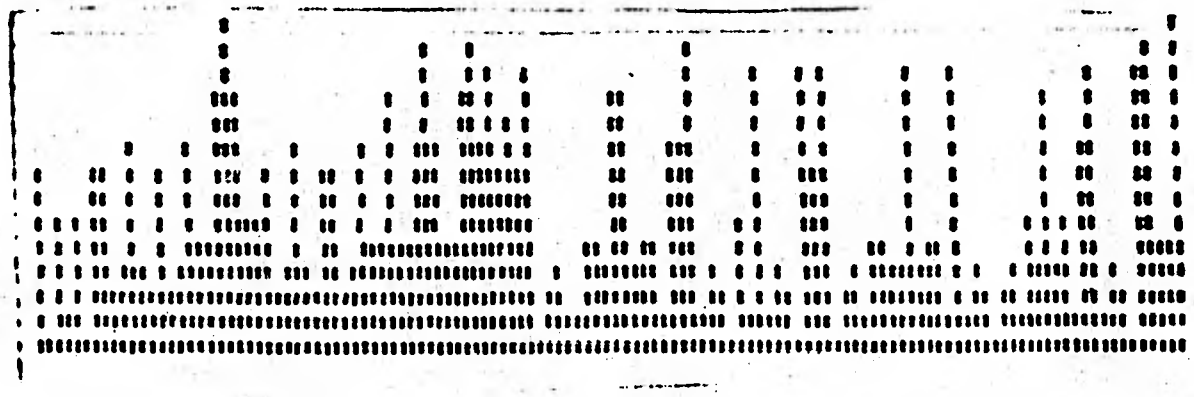
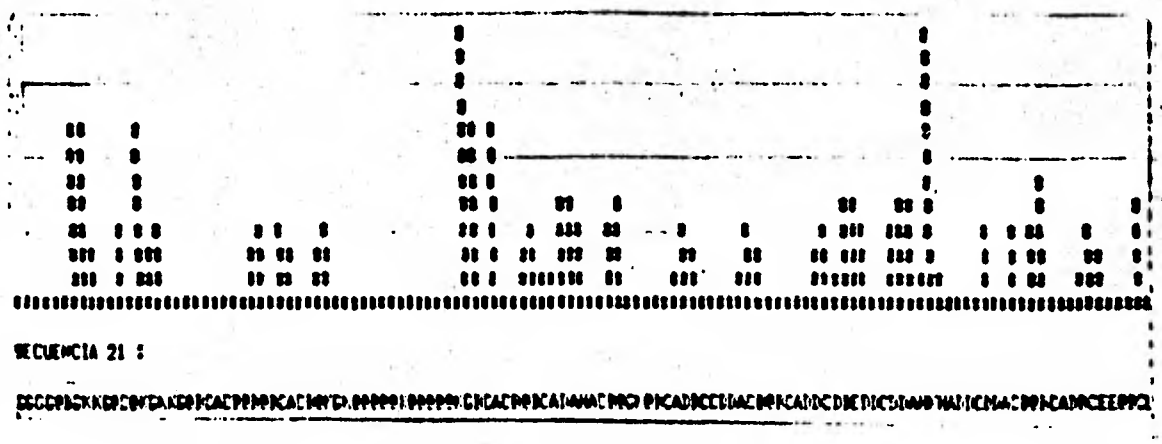


Fig. 4.1 Utilización de la programación dinámica para el reconocimiento de formas. En las dos figuras superiores se demuestra como comparando dos fonogramas, punto a punto no se detecta homología. Usando la técnica se determina que son fonogramas equivalentes.

Para revisar homologías visualmente se pensó representar las secuencias como histogramas usando GRAFPOL y posteriormente buscar la similitudes utilizando alguna técnica de reconocimiento de formas como es la de programación dinámica. Consideramos que es adecuada ya que permite una flexibilidad temporal en el reconocimiento. La figura 4.2 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 podría declarar que son dos estados fisiológicos diferentes los que definieron a las secuencias de eventos. Sin embargo esta hipótesis requiere de una comprobación formal.





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

D. Existen analogías adicionales al hecho de que las secuencias conductuales y macromoleculares se pueden tratar en forma similar, por ejemplo:

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

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

En el primer caso, principalmente el factor genético mas el medio interno determinan la manera en que se producirá una proteína. Esta proteína tendra una función bioquímica. En el segundo sistema, también el A.D.N. más el medio externo determinaran que conducta se presentará y ésta tendrá una función biológica. Las relaciones entre los primeros dos elementos del primer sistema y la regulación de ellas son mucho mejor conocidas actualmente que las correspondientes del segundo sistema. Con la creación de nuevas técnicas de adquisición de conocimiento, este aspecto tendrá mayor comprensión.

El trabajar con símbolos y no con su significado tiene la ventaja de generalizar la función de los programas. Lenguajes computacionales como LISP (List Procesor) que han sido utilizados para trabajar con cálculo proposicional (sistema matemático simbólico) deben ser igualmente efectivos en el tratamiento de los problemas aqui expuestos.

INDICE DE FIGURAS

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

3.2.6 Compaginación de secuencias invertidas repetidas.

3.2.7 Dos posibles estructuras secundarias encontradas por DUFAL en la región de control de glnA.

3.3.1 Coïncidencias dentro de 90 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 aminoácidos usada para determinar diferencias de una base.

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

3.3.5 Comparación entre N.crassa y E.coli tomando en cuenta el cambio de una base.

3.3.6 Histogramas de hidrofobicidad/hidrofilicidad de las enzimas GDH nadph dependientes de E.coli, N.crassa, pollo, y bovino. Cerca de la posición 53 se observa un pico de acidez.

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

3.3.8 Matriz indicando distancias mínimas mutacionales entre los diferentes aminoácidos.

3.3.9 Distancias mínimas mutacionales entre las cuatro secuencias.

3.3.10 Dendrograma obtenido apartir de de las distancias mutacionales.

4.1 Reconcimiento de patrones usando la técnica de programación dinámica (Scientific American).

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

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, [illegible] and Fear in Cyclostomatous Rats. Interdisc. Sci. Rev., 1:43-62.(1976).

3. Diamond, D.M., Fleshler M., Alkon D.L. Differential Reactions and Exploratory Behavior in Rats. J. Comp. Physiol. Psychol. 87: 1129-1133. (1974).

4. Diamantstein K. H., Moon K., Smith E., Nicotinamide Adenine Dinucleotide Phosphate-specific Glutamate Dehydrogenase of Neurospora. J. Biol. Chem. 256:3644-3651. (1981).

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

6. Díaz, J. L., Análisis Estructural de la Conducta . En: La conducta como evento pautado.Mexico:UNAM (en prensa).

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

8. Dickerson, R. E., The Stucture and History of an Ancient Protein.Sci. Am. 226:37-54.(1972).

9. Fitch W. M.Margoliash E., Construction of Phylogenetic Trees. Science 155:279-282. (1967).

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

11. Gouy M., Gautier C., Codon Usage in Bacteria: Correlation with Gene Expressivity.Nucleic Acid Research. 22:7055.7068.(1982).

12. Levinson S. E., Liberman M., Speech Recognition by Computer.Sci. Am. 240:56-68.(1981).

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

14. Maxam A.M., Gilbert W. New method for sequencing in DNA PNAS 74:560-564.(1977).

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

16. Puhler A., Klipp W., Fine Structure analisis of the Gene Region for Nitrogen Fixation (nif) of Klebsiella pneumoniae En: "Biology of Inorganic Nitrogen and Sulfur" Bothe H., Trebst A. , Springer-Verlag Berlin Heidelberg (1981).

18. Quintó C., De La Vega, H., Flores M., Fernández L., Sánchez T., Soberón C., Palacios R., Reiteration of nitrogen fixation gene sequences in Rhizobium phaseoli.Nature. 299, 724-726.
(1982).

19. Rosenberg Martin.,Court Donald. Genetic Regulation.Ann. Rev.Genet. 13:319-353.(1979).

20. Ruvkun M. B., Ausbel F. M., Interspecies Homology of nitrogenase genes. PNAS 77:191-195. (1980).

21. Santis M., Díaz J. L. Location Response to a Startling Noise on the Preferred Grooming Site in Mice. Physiol and Behavior (en prensa).

22. Scott K.F. et al, Klebsiella nif Structural Gene Sequences, J.Mol.Appl.Genet. 1:72-81. (1981).

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

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

25. Werner A., DNA Modification and Restriction, Prog. Nuc. A. Res. Mol. Biol.14:1-26.(1974)

26. Yanofsky C., Attenuation in the Control of Expression of the Bacterial Operon, Nature 289:751-759.(1981).