U N I V E R S I D A D DE G U A D A L A J A R A

CENTRO DE CIENCIAS BIOLOGICAS Y AGROPECUARIAS

DIVISION DE CIENCIAS BIOLOGICAS Y AMBIENTALES

"SECUENCIAS SIMPLES DE TRES GENOMAS MICROBIANOS: IMPLICACIONES EN EL ORIGEN Y EVOLUCIÓN DE LAS PROTEÍNAS.

T E S I S P R O F E S I O N A L

QUE PARA OBTENER EL TITULO DE :

L I C E N C I A D O E N B I O L O G I A

P R E S E N T A:

MAGALY OLIVO HERNANDEZ

GUADALAJARA, JALISCO., 1999

UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS BIOLOGICAS Y AGROPECUARIAS

COORDINACION DE CARRERA DE LA LICENCIATURA EN BIOLOGIA

COMITÉ DE TITULACION

C. MAGALY OLIVO HERNANDEZ
P R E S E N T E .

Manifestamos a Usted que con esta fecha ha sido aprobado su tema de titulación en la modalidad de TESIS con el título **"SECUENCIAS SIMPLES EN TRES GENOMAS MICROBIANOS : IMPLICACIONES EN EL ORIGEN Y EVOLUCION DE LAS PROTEINAS"**, para obtener la Licenciatura en Biología.

Al mismo tiempo le informamos que ha sido aceptado como Director de dicho trabajo al **M.C. ARTURO CARLOS II BECERRA BRACHO** y como asesor al **DR. ANTONIO LAZCANO-ARAUJO REYES.**

A T E N T A M E N T E
" PIENSA Y TRABAJA "
LAS AGUJAS, ZAPOPAN, JAL., MAYO 28 DE 1999

DR. ARTURO OROZCO BAROCIO
PRESIDENTE DEL COMITE DE TITULACION

COMITE DE TITULACION
LICENCIATURA EN BIOLOGIA

M. EN C. MARTHA GEORGINA OROZCO MEDINA
SECRETARIO DEL COMITE DE TITULACION

c.c.p. M.C. ARTURO CARLOS II BECERRA BRACHO.- Director del Trabajo.
c.c.p. DR. ANTONIO LAZCANO-ARAUJO REYES.- Asesor del Trabajo.
c.c.p. Expediente del alumno

AOB/MGOM/bacg*

Km. 15.5 Carretera Guadalajara - Nogales
Predio "Las Agujas", Nextipac, C.P. 45110 • AP 39-82
Tels (91-3) 682-0248 682-0374 Fax. 6820120

Dr. ARTURO OROZCO BAROCIO
PRESIDENTE DE LA COMISIÓN DE TITULALCIÓN
DIVISION DE CIENCIAS BIOLÓGICAS Y AMBIENTALES

P R E S E N T E.

Por este medio me permito comunicar a usted que habiendo revisado el trabajo de tesis "Secuencias simples de tres genomas microbianos: Implicaciones en el origen y evolución de las proteínas", preparado por la pasante de Biología Magaly Olivo Hernández con código 087556182 y del cual fungí como director, considero que ha sido concluido satisfactoriamente.

Por esta razón, puede proceder a la presentación de la tesis para su revisión en la División a su cargo y el examen profesional correspondiente.

A T E N T A M E N T E
"Por mi raza hablara el espíritu"
México, D.F., a 20 de Mayo de 1999

Director de Tesis
M.en C. Arturo Carlos II Becerra Bracho

Asesor de Tesis
Dr. Antonio E. Lazcano-Araujo Reyes

Sinodales
Dr. Anne Santerre. 28/05/99.
M. en C. Alfonso E. Islas. R. 2/06/99.
Dr. Arturo Orozco Barocio.

AGRADECIMIENTOS

- Quiero agradecer en primera instancia al director de esta tesis el M. en C. Arturo Becerra Bracho por permitirme conocer el valor real de la investigación, por compartir conmigo sus conocimientos, y desde luego por su invaluable enseñanza, amistad, apoyo y paciencia durante el desarrollo de este trabajo sin olvidar el buen sentido del humor que tiene, asimismo su generosa ayuda durante mis estancias en el D.F. Arturo muchísimas gracias.
- Un agradecimiento muy especial al Dr. Antonio Lazcano por haberme aceptado en su laboratorio, por compartir sus conocimientos y estar pendiente de lo que sucede en su laboratorio, en el cual fue muy agradable trabajar por ese gran sentido del humor que lo caracteriza aparte de ser un gran investigador.
- A los compañeros del laboratorio: Luis, Sara, Ervin, Amanda, Héctor, Ana, Ulises, Rosana y Josextu por esa armonía que se siente al estar trabajando con ustedes, además de compartir el gusto de la investigación y la risa.
- A mis sinodales Dra. Anne Santerre L. al Dr. Arturo Orozco B. y al M. en C. Alfonso Islas R. por el interés que demostraron en este tipo de trabajos, por sus atinados comentarios y sugerencias.
- A las mujeres de mi casa mi abuela Micaela, mi madre Yolanda y mi hermana Raquel por ese enorme cariño, respeto y amor han tenido siempre para mí, además de su invaluable apoyo durante mi educación muchas gracias.
- A Luis Bernardo por su cariño, apoyo y su enorme paciencia lo que hizo más fácil y divertido este trabajo..
- A la fundación Becerra-Novoa por compartir su casa conmigo, también por todas sus atenciones durante las estancias que pase en el D.F. muchas gracias.
- Al Ing. Tino Granata L. quien me permitió realizar mi tesis en el centro de computo sin ningún costo, siendo siempre muy accesible.
- De igual manera a Angélica Velázquez por compartir conmigo sus conocimientos de computación y permitirme trabajar en horas poco accesibles aun en sus días de descanso muchas gracias Angy.
- A mis maestros José Luis Navarrete, Patricia Castro, Sergio Guerrero y Georgina Quiroz quienes aclararon siempre mis dudas dentro y fuera de clases, muchas gracias por invertir en mi educación.
- A mis amigas y compañeras de toda la licenciatura Lidia, Martha, Elia y Edith por compartir tantas aventuras.
- A los amigos del D.F. Bernardo, Paty, Arturo, Hugo, Lorena, Alex, Luis y David quienes compartieron la comida, el intercambio de ideas, la diversión y el gusto por la biología.

A Luis Bernardo, compañero
de largos días y noches cortas.

Este trabajo se realizó en el Laboratorio de Microbiología A-107 de la Facultad de Ciencias de la UNAM. Bajo la dirección del M. en C. Arturo Carlos II Becerra Bracho con el apoyo del proyecto PAPIIT IN213598

ÍNDICE

RESUMEN

Las secuencias simples (con baja complejidad composicional) se presentan en regiones codificantes, estas son producidas principalmente por el fenómeno conocido como *slippage-like* (patinaje de la polimerasa), incorporando sus productos en la proteína siendo una fuente de variabilidad genética.

Los análisis de las secuencias simples de los tres genomas microbianos completos de *Haemophilus influenzae*, *Methanococcus jannaschii* y *Saccharomyces cerevisiae* revelan que la baja complejidad tiene una amplia distribución en todos los grupos funcionales propuestos por Riley (1993); la composición, los segmentos cargados, los aminoácidos hidrofóbicos y de transmembrana, los multitipletes, las estructuras repetidas, la periodicidad y la espacialidad que presentaron las secuencias simples revelan que no se tiene una correlación entre las propiedades físico-químicas y la función biológica, manteniendo diferentes estructuras secundarias en las proteínas que participan.

Dentro de las proteínas que presentan una baja complejidad se encontraron las ferrodoxinas, DNA polimerasas ATP-sintetasas, topoisomerasas y ribosomales proteínas muy antiguas; de igual manera las secuencias simples se han visto involucradas en organismos patógenos y en enfermedades genéticas, lo que sugiere que el fenómeno del patinaje de la polimerasa es muy antiguo, estando presente en el último ancestro común a todos los seres vivos (*cenancestro*) y por lo tanto precede a la patogénesis.

La ocurrencia de las secuencias simples en varias funciones, la antigüedad de algunas proteínas y su presencia en los tres dominios, sugiere que el proceso de patinaje de la polimerasa no es solo un importante proceso en la evolución de genomas completos, sino también en el origen y evolución de las proteínas.

1. INTRODUCCION

1.1 Secuencias simples.

Las secuencias simples son regiones de DNA ó proteínas caracterizadas por una escasa variabilidad de sus componentes (nucleótidos y aminoácidos respectivamente), por lo que presentan una baja complejidad en términos de composición. Estas secuencias también son conocidas como: secuencias crípticas o microsatélites (está última sólo en DNA no codificante). Las secuencias simples se encuentran ampliamente distribuidas tanto en especies procariontes como eucariontes, en regiones no codificantes del material genético como en las codificantes, lo que ha sugerido que este tipo de secuencias tiene un papel importante en la evolución del tamaño de genomas y como fuente de variabilidad genética (Tauttz *et al.*, 1986; Hancock, 1995).

El análisis de las bases de datos de proteínas sugiere que el 25% de los residuos de aminoácidos presenta segmentos con sesgos composicionales y cerca del 50% de las proteínas contiene al menos una región con esta característica (Wootton, 1997). En los últimos años el interés por las secuencias simples se ha incrementado, dado su papel en diferentes fenómenos biológicos como: enfermedades genéticas, adaptabilidad de organismos patógenos (como *Neisseria gonorrhoeae* y *Haemophilus influenzae*) y determinación de parentesco entre poblaciones de especies silvestres (Moxon y Wills, 1999). Por otro lado, el estudio de este fenómeno puede ayudar a entender el origen y la evolución de las proteínas, así como el crecimiento y la variabilidad del material genético (Wootton, 1997).

Las secuencias simples son producidas principalmente por procesos tipo *slippage-like* (patinaje o tartamudeo de la polimerasa, Fig. 1), durante el proceso de la replicación del DNA y son reguladas por los mecanismos de reparación del DNA (Tautz y Schlötterer, 1994).

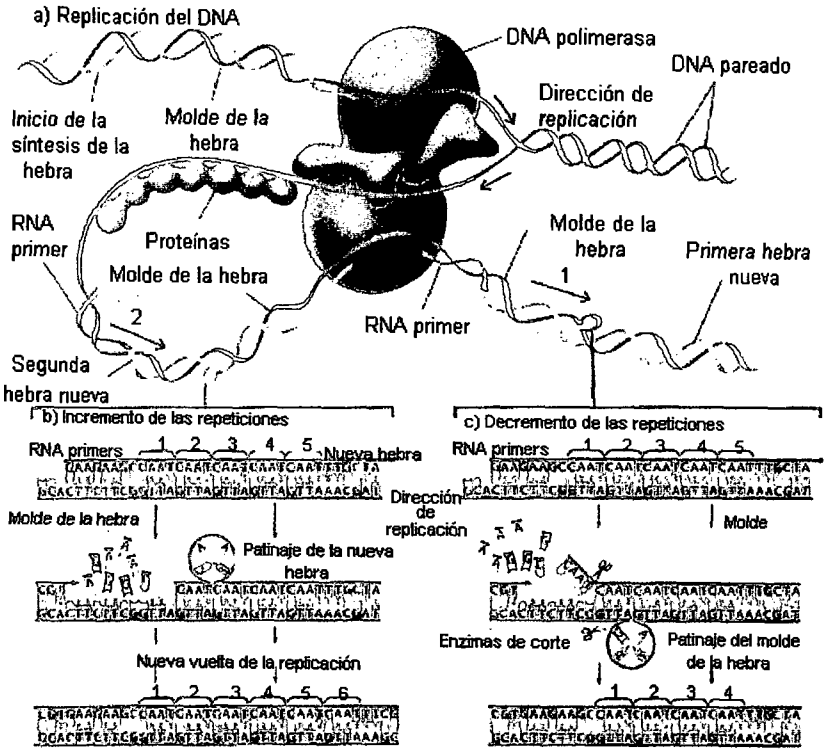


Fig. 1. Proceso tipo *slippage-like* o patinaje de la polimerasa, modificado de Moxon y Wills, (Scientific American, enero de 1999). a) el complejo de la DNA polimerasa separa la doble hélice del DNA copiando ambas hebras. La DNA polimerasa empieza a sintetizar pequeños fragmentos (1) inicia con un RNA *primer*, generando un fragmento pequeño, la DNA polimerasa se patina formando una horquilla, se siguen sintetizando fragmentos (2) cuando la polimerasa termina el fragmento, el RNA *primer* es removido y los dos fragmentos son unidos al DNA. Incrementando el número de repeticiones b) ocurre cuando la nueva hebra se patina provocando repeticiones, al ensamblarse con el molde viejo se ganan repeticiones en la nueva hebra c) sucede cuando las enzimas de reparación cortan las repeticiones.

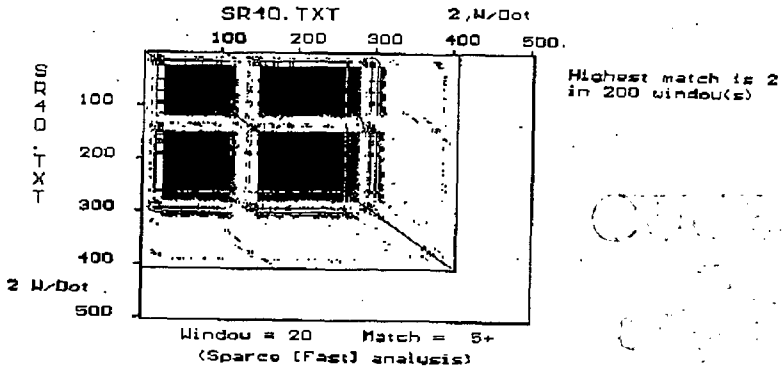
Las secuencias simples pueden identificarse por tres propiedades: a) complejidad: propiedad computable, que resulta de la composición de la secuencia y permite hacer una representación generalizada de los sesgos composicionales; b) patrones: son secuencias sencillas caracterizadas por su contenido, su distribución de residuos y sus k-gramas (un k-grama es una palabra de k-letras); c) periodicidad: es la repetición de residuos o k-gramas a intervalos constantes (Wootton, 1997). (Cuadro 1).

Cuadro 1. Ejemplo de las propiedades de una secuencia simple (G_8A_8)

Complejidad	G A A G G A A A G G G A G A G A	sin patrón ni periodicidad
Periodicidad	G A G A G A G A G A G A G A G A	posee periodicidad y de ello resultan patrones de sus k-gramas
Patrón	G G A G G A A A A G G A A G G A	posee patrones (k-gramas) GGA y AGGA, pero su distribución es irregular y no muestra periodicidad

Estas propiedades pueden ser analizadas y cuantificadas por diferentes algoritmos ya sea por métodos gráficos ó analíticos; la matriz de puntos (*dot-plot*) de una secuencia puede evidenciar la baja complejidad de manera visual por medio de la saturación de puntos en una región de la matriz, lo que hace al análisis visual un instrumento importante en la determinación de secuencias simples. (Fig. 2).

a)



b)

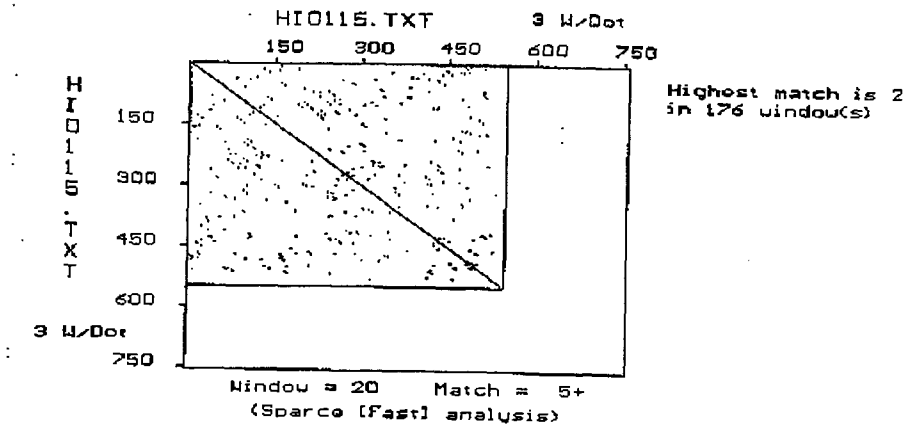


Fig. 2. Representación visual de dos secuencias de aminoácidos, por medio de una matriz de puntos (*dot-plot*), para la determinación de secuencias simples, a) la secuencia SR40 al compararse contra ella misma, presenta segmentos con sesgos composicionales representados por la saturación de puntos y en b) la secuencia HI0115 al compararse contra ella misma, esta no presenta segmentos con sesgos composicionales.

La complejidad de una secuencia fue definida por Wootton y Federhen (1993) como el producto del número de secuencias posibles en un segmento *n*, por el inverso del logaritmo (base 20 ó 4, según el caso), de una ventana de tamaño L (ecuación 1). Esta ecuación puede ser utilizada para cuantificar la complejidad de una secuencia, dando una alternativa de análisis al método gráfico (*dot-plot*).

$$K_1 = \frac{1}{L} \log_{20} \left(\frac{L!}{\prod_{i=1}^{20} n_i!} \right)$$

Ecuación 1. Donde K_1 es el valor complejidad, n_i son los 20 aminoácidos ó los 4 nucleótidos de la secuencia y L es el número de secuencias probables (Wooton y Fedheren, 1993). (Apéndice I).

1.2 Genómica y filogenias universales.

El eminente desarrollo de las técnicas de secuenciación del material genético ha producido un importante acervo de información biológica, junto a este incremento de datos moleculares, el desarrollo de las ciencias de la informática ha permitido un fácil análisis de los mismos. Esto en conjunto, ha permitido plantear nuevas preguntas en el estudio de la Evolución, en especial en el análisis de genomas completos, disciplina que hoy se conoce como genómica, la cual ofrece nuevas perspectivas en análisis y caracterización de los mecanismos de expresión, modificación y evolución del genoma (Clark, 1994).

El análisis filogenético de las secuencias 16/18S rRNA no solamente permitió a Woese y Fox (1977) caracterizar a las arqueobacterias como una rama

monofilética claramente definida, sino que también mostró que todas las especies conocidas forman parte de un mismo árbol filogenético en el cual es posible distinguir otros dos linajes, formados por las eubacterias y eucariontes. Este árbol evolutivo sin raíz, que resultó de la comparación de las secuencias de genes ortólogos de rRNA, se trifurca a partir de un ancestro común. Dicha información llevo a uno de los planteamientos más importante de la biología contemporánea, la filogenia universal de los seres vivos en tres grandes grupos taxonómicos, los dominios: Arquea, bacteria y Eucaria (Woese *et al.*, 1990).

En la actualidad se cuenta con 20 genomas completos secuenciados provenientes de los tres dominios, **a)** Bacteria: *Haemophilus influenzae* (Fleischmann *et. al.*, 1995), *Mycoplasma genitalium* (Fraser *et. al.*, 1995), *Synechocystis* sp. (Kaneko *et. al.*, 1996), *Mycoplasma pneumoniae* (Himmelreich *et. al.*, 1996), *Helicobacter pylori* (Tomb *et. al.*, 1997), *Helicobacter pylori* (Alm *et* al., 1999), *Escherichia coli* (Blattner *et. al.*, 1997), *Bacillus subtilis* (Kunst *et.al.*, 1997), *Borrelia burgdorferi* (Fraser *et al.*, 1997), *Aquifex aeolicus* (Deckert *et al.*, 1998), *Mycobacterium tuberculosis* (Cole *et al.*, 1998), *Treponema pallidum* (Fraser *et al.*, 1998), *Chlamydia trachomatis* (Stephens *et al.*, 1998), *Rickettsia prowazekii* (Andersson *et al.*, 1998); **b)** Arquea: *Methannococcus jannaschi* (Bult et *al.*, 1996), *Methanobacterium thermoautotrophicum* (Smith *et al.*, 1997), *Archaeoglobus fulgidus* (Klenk *et al.*, 1997), *Pyrococcus horikoshii* (Kawarabayasi *et al.*, 1998); y **c)** Eucaria: *Saccharomyces cerevisiae* (Goffeau *et al.*, 1997). y *Caenorhabditis elegans* (www.sanger.ac.uk/Projects/C_elegans/).

1.3 Planteamiento.

Uno de los principales desafios en el estudio de la evolución temprana de la vida, es entender como se origina e incrementa el material genético y como este se vuelve fuente de variación para la evolución de las proteínas. La duplicación génica (paráloga), y la simbiosis son los principales mecanismos propuestos para solucionar estos problemas. Si bien, se ha comprobado el importante papel de estos fenómenos en la elongación del material genético (Horowitz, 1945; Waley, 1969), también es cierto que estos no resuelven completamente el problema.

En este trabajo se pretende caracterizar a las secuencias simples que se encuentran en regiones codificantes de los genomas completos, tratando de entender el papel que estas han jugado en la evolución de las proteínas. Si las secuencias simples están presentes en un alto porcentaje en las regiones codificantes del genoma y como estas son producidas principalmente por el patinaje de la polimerasa (*slippage-process*), es factible pensar que este fenómeno a tenido que ver directamente con el incremento de las proteínas, tanto en su número como en su tamaño y es posible que jugara un papel en la adquisición de nuevas funciones.

1.4 Objetivos

Objetivo general:

-Conocer el papel de las secuencias simples con baja complejidad de tres genomas microbianos completos (*Haemophilus influenzae, Methanococcus jannaschii, Saccharomyces cerevisiae*) en el origen y evolución de las proteínas.

Objetivos particulares:

-Conocer la frecuencia de secuencias simples (utilizando un método gráfico de matriz de puntos) en los genomas microbianos completos de *Haemophilus influenzae, Methanococcus jannaschii, Saccharomyces cerevisiae*.

-Definir la distribución funcional de las secuencias con baja complejidad en los tres genomas completos.

-Realizar un análisis de las secuencias simples encontradas en los tres genomas microbianos, para conocer la composición, distribución de cargas, distribución de otros aminoácidos tipo, estructuras repetidas, multitipletes, periodicidad y espacialidad. Con el fin de caracterizar el fenómeno.

2. MATERIAL Y METODOS

Las secuencias de aminoácidos de los tres genomas microbianos (*Haemophilus influenzae, Methanococcus jannaschii, Saccharomyces cerevisiae)* se obtuvieron por medio del *anonymous* FTP del GenBank (disponible en http://www.ncbi.nlm.nih.gov/). Se realizó una comparación de cada una de las secuencias por medio de una matriz de puntos realizada con el programa Dot plot Coral Software PROGRAM V1.0 1989. El programa se instaló y corrió en una PC-pentium.Con este programa se realizó un análisis gráfico-visual de todas las secuencias de aminoácidos provenientes de los tres genomas microbianos, el programa permite comparar una secuencia contra ella misma mostrando en la pantalla una matriz de puntos con las intersecciones de los aminoácidos, si existe una repetición de aminoácidos en la secuencia, el resultado visual es representado por una saturación de puntos. Los cálculos del programa fueron realizados con una ventana mínima de cinco (por cada cinco aminoácidos relacionados se coloca un punto), representando secciones de baja complejidad. Solamente se tomaron en cuenta las secuencias simples con secciones de baja complejidad mayores al 20% del total de la secuencia.

Para conocer la distribución funcional de las secuencias simples, estas se agruparon de acuerdo a su actividad metabólica en siete grupos funcionales: I) Metabolismo de macromoléculas, II) Metabolismo intermediario, III) Biosíntesis de moléculas pequeñas, IV) Procesos celulares, V) Estructura celular, VI) Otras funciones (Riley, 1993), y VII) Hipotéticas ó de función desconocida, para ello se

realizo un programa en lenguaje shell-unix, el cual identifico a las secuencias, con base a una lista de palabras claves referentes a cada grupo (Apéndice II).

Para analizar las secuencias que presentan menor grado de complejidad, se utilizó el programa SEG (segment sequence (s) by local complexity), el cual analiza la composición de complejidad local dentro de la secuencia, este programa se utilizó con los parámetros: **L**=30, $\mathbf{K_1}$=2.6 y $\mathbf{K_2}$=2.8 (Wootton y Federhen, 1993), lo que permite identificar a las secuencias con una marcada baja complejidad (Apéndice III).

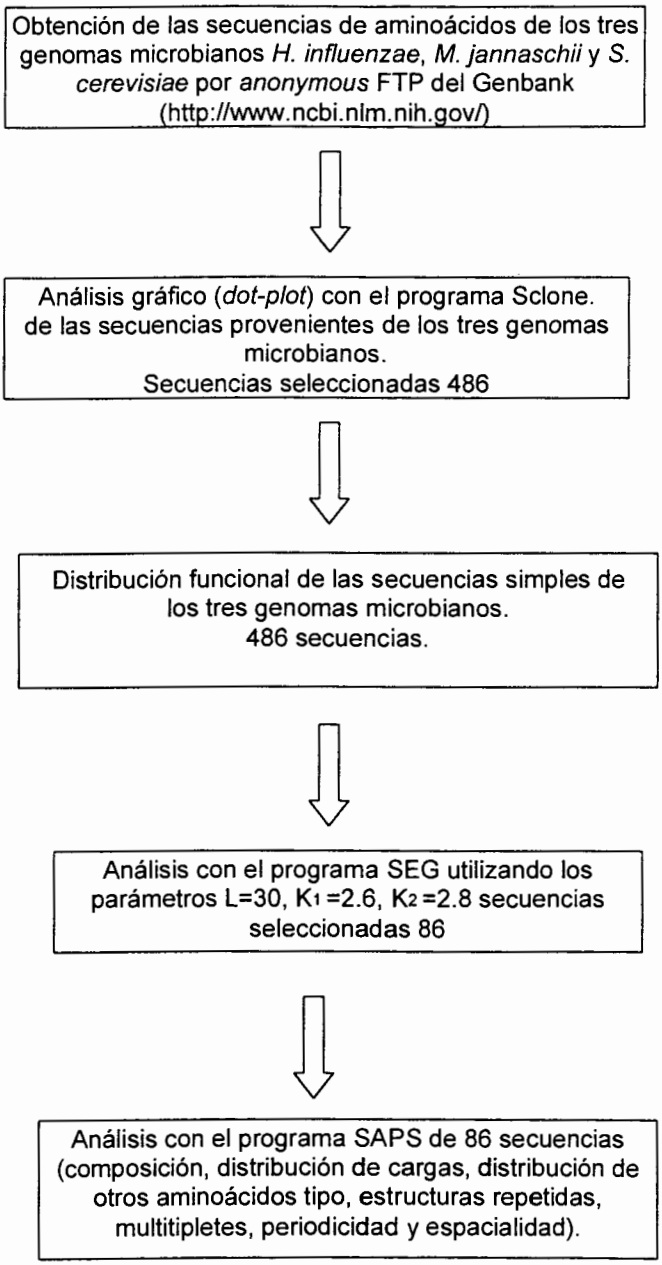
Las secuencias de aminoácidos de los tres genomas que resultaron del SEG, se analizaron con el programa SAPS (Statistical Analysis of Protein Sequences), el cual esta disponible en la página http://ulrec3.unil.ch/software/SAPS_form.html. Este programa analiza **a)** la composición de la secuencia la cual es evaluada al comparar cada aminoácido con el valor porcentual ya establecido que se tiene para cada aminoácido, si el valor se excede se representa con los signos (++) (+), si el valor es muy bajo se usan los signos (--) (-), **b)** la distribución de cargas es evaluada por grupos de 30 a 60 aminoácidos con segmentos cargados ya sean positivos, negativos y mezclados, **c)** la distribución de otros aminoácidos tipo se indica por segmentos de aminoácidos hidrofóbicos y de transmembrana, **d)** las estructuras repetidas son indicadas por dos alfabetos: los 20 aminoácidos y 11 letras del alfabeto las que indican segmentos de aminoácidos hidrofóbicos (LVIF), cargados (KR y ED), pequeños (AG), los aromáticos (YW), el grupo hidroxil (ST) y el grupo amido (NQ), **e)** multitipletes de la secuencia son evaluados por los aminoácidos y las cargas, contando también altitipletes, **f)** la periodicidad es identificada por periodos entre 1

y 10 aminoácidos, por las cargas y la hidrofobicidad, **g)** la espacialidad es evaluada por los 20 aminoácidos en cuanto a la significancia larga o corta, máxima o mínima espacialidad en la distribución de los aminoácidos. Para todos los análisis del programa SAPS, solo se tomaron en cuenta las secuencias en las cuales sus resultados sobrepasan el rango (Apéndice IV).

Los programas utilizados para los anteriores análisis se corrieron en una estación de trabajo SUN-SPARC 20 y SUN-Ultra 5.

Se anexa un diagrama de flujo, mostrando los la depuración que se realizo de las secuencias de aminoácidos provenientes de los tres genomas microbianos.

Diagrama de flujo.



3. RESULTADOS

De las secuencias de aminoácidos analizadas por el método gráfico (*dot-plot)* provenientes de los genomas microbianos de *Haemophilus influenzae, Methanococcus jannaschii* y *Saccharomyces cerevisiae* se seleccionaron 486 que presentaban segmentos con baja complejidad mayores al 20% del tamaño de la secuencia (Cuadro 2), observando el mayor número de secuencias simples en el genoma eucarionte (*S. cerevisiae*), seguido de la arqueobacteria (*M. jannaschii*) y la bacteria (*H. influenzae*).

Cuadro 2. Secuencias simples de los tres genomas microbianos

Genoma	Secuencias codificantes	Secuencias simples	Porciento
Haemophilus influenzae [B]	1,709	57	3.33%
Methanococcus jannaschii [A]	1,738	90	5.17%
Saccharomyces cerevisiae [E]	5,885	339	5.77%

[B] bacteria [A] arquea [E] eucaria

La distribución funcional de las 486 secuencias simples se realizó para cada genoma (Fig.3) con base en los siete grupos funcionales, obteniendo el siguiente acomodo:

Haemophilus influenzae
I Metabolismo de macromoléculas
>HI0067 DNA mismatch repair protein (mutL) {Escherichia coli}
>HI0216 type I restriction enzyme ecoKI specificity protein (hsdS) {Escherichia coli}
>HI0231 ATP-dependent RNA helicase (deaD) {Escherichia coli}
>HI0397 exonuclease VII, large subunit (xseA) {Escherichia coli}
>HI0422 ATP-dependent RNA helicase (srmB) {Escherichia coli}
>HI0641 ribosomal protein L7/L12 (rpL7/L12) {Escherichia coli}
>HI0699 rotamase, peptidyl prolyl cis-trans isomerase (slyD) {Escherichia coli}
>HI0990 Iga1 protease (iga1) {Haemophilus influenzae}
>HI1056 type III restriction-modification ECOP15 enzyme (mod) {Escherichia coli}

>HI0245 queuosine biosynthesis protein (queA) {Escherichia coli}
>HI1283 transcription factor (nusA) {Salmonella typhimurium}
>HI1284 initiation factor IF-2 (infB) {Escherichia coli}
>HI0550 lipooligosaccharide biosynthesis protein {Haemophilus influenzae}

II Metabolismo intermediario
>HI0061 recombination protein (rec2) {Haemophilus influenzae}
>HI0220 aerobic respiration control sensor protein (arcB) {Escherichia coli}
>HI0251 energy transducer (tonB) {Haemophilus influenzae}
>HI0446 fructose-permease IIBC component (fruA) {Escherichia coli}
>HI0448 fructose-permease IIA/FPR component (fruB) {Escherichia coli}
>HI0916 export factor homolog (skp) Pasteurella multocida
>HI0971 biotin carboxyl carrier protein (fabE) {Escherichia coli}
>HI1232 dihydrolipoamide acetyltransferase (aceF) {Escherichia coli}
>HI1545 C4-dicarboxylate transport protein {Rhizobium leguminosarum}

III Biosíntesis de moléculas pequeñas
>HI0264 heme-hemopexin-binding protein (hxuA) {Haemophilus influenzae}
>HI0538 urease protein (ureE) {Helicobacter pylori}

IV Procesos celulares
>HI0049 2-keto-3-deoxy-D-gluconate kinase (kdgK) {Erwinia chrysanthemi}
>HI0664 transport ATP-binding protein (cydC) {Escherichia coli}
>HI0712 transferrin-binding protein 1 (tbp1) {Neisseria meningitidis}
>HI1339 embryonic abundant protein, group 3 {Triticum aestivum}
>HI1706 high-affinity choline transport protein (betT) {Escherichia coli}
>HI1374 cell division protein (mukB) {Escherichia coli}

V Estructura celular
>HI0066 N-acetylmuramoyl-L-alanine amidase (amiB) {Escherichia coli}
>HI0119 adhesin B precursor (fimA) {Streptococcus parasanguis}
>HI0383 outer membrane integrity protein (tolA) {Escherichia coli}
>HI0915 UDP-3-O-(R-3-hydroxymyristoyl)-glucosamine
>HI1579 15 kDa peptidoglycan-associated lipoprotein (lpp) {Haemophilus influenzae}
>HI1685 outer membrane integrity protein (tolA) {Escherichia coli}

VI Otras funciones
>HI0861 virulence associated protein homolog (vacB) {Escherichia coli}
>HI1238 heat shock protein (dnaJ) {Escherichia coli}

VII *hipotéticas y función desconocida*
>HI0147 hypothetical protein (SP:P37675) {Escherichia coli}
>HI0259 hypothetical protein (GB:X73124_13) {Bacillus subtilis}
>HI0271 hypothetical protein (SP:P31680) {Escherichia coli}
>HI0519 hypothetical protein (SP:P33024) {Escherichia coli}
>HI0526
>HI0594
>HI0662
>HI0696 hypothetical protein (GB:U14003_133) {Escherichia coli}
>HI0700 hypothetical protein (GB:U14003_167) {Escherichia coli}
>HI0756 hypothetical protein (SP:P37690)) {Escherichia coli}

>HI1053 hypothetical protein (GB:M66060_5) {Alcaligenes eutrophus}
>HI1058
>HI1404
>HI1495
>HI1514
>HI1517
>HI1566
>HI1601
>HI1718

Methanococcus jannaschii

I Metabolismo de macromoléculas

>MJ0124 type Y restriction enzime
>MJ0262 putative translation factor, FUN 12/biF-2 family o PIR:S48519
>MJ0291 signal recognition particle protein EGAD:6866
>MJ0462 ribosomal protein L29 o SP:P22665
>MJ0487 phenylalanyl-tRNA synthase, subunit alpha o SP:P15625
>MJ0508 ribosomal protein L12 o SP:P10623
>MJ0510 ribosomal protein L1 o SP:P15824
>MJ0564 alanyl-tRNA synthase (ala RS) o SP:P43815
>MJ0591 proteasome, subunit alpha o GI:1002686
>MJ0782 transcription initiation factor IIB PIR:S34116
>MJ0882 ATPase, vanadate-sensive o PIR:A38542
>MJ1042 DNA-dependent RNA polymerase, subunit A' o SP:P41556
>MJ1176 ATP-dependent 26S protease regulatory subunit 4 EGAD:827
>MJ1214 type I restriction enzyme EGAD:28293
>MJ1238 prolyl-tRNA synthetase SP:P07814
>MJ1263 methionyl-tRNA synthetase EGAD:28284
>MJ1422 replication factor C EGAD:237
>MJ1505 putative ATP-dependent RNA helicase, eIF-4A family SP:P40562
>MJ1512 reverse gyrase SP:Q08582

II Metabolismo intermediario

>MJ0152 carbon monoxide dehydrogenase, alpha subunit o SP:P27988
>MJ0156 carbon monoxide dehydrogenase, iron-sulfur subunit o SP:P27273
>MJ0221 ATP synthase, subunit K o SP:P43457
>MJ0222 ATP synthase, subunit Y o SP:P43439
>MJ0514 polyferredoxin o PIR:S24802
>MJ0542 phosphoenolpyruvate synthetase o SP:P42850
>MJ0829 peptide chain release factor
>MJ0832 anaerobic ribonucleoside-triphosphate reductase o SP:P28903
>MJ0860 bifunctional short chain isoprenyl diphosphate synthase o P:S75695_1
>MJ1231 oxaloacetate decarboxylase, alpha subunit SP:Q03030
>MJ1303 polyferredoxin PIR:G30315
>MJ1408 GTP-binding protein, GTP1/OBG-family GP:U43281_15

III Biosíntesis de moléculas pequeñas

>MJ1057 glycosyl transferase GP:U14554_5
>MJ1362 NADH dehydrogenase, subunit 1 GP:L29771_1
>MJ1420 glutamine--fructose-6-phosphate transaminase SP:P17169

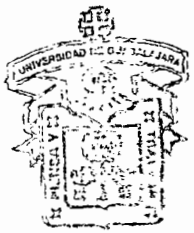
>MJ1269 branched-chain amino acid transport protein livH SP:P08340

IV Procesos celulares
>MJ1156 cell division control protein CDC48 SP:P25694
>MJ1275 NA(+)/H(+) antiporter SP:P26235

V Estructura celular

VI Otras funciones
>MJ0718 chromate resistance protein A o SP:17551
>MJ1166 tungsten formylmethanofuran dehydrogenase, subunit C o GP:X87970_2
>MJ1171 tungsten formylmethanofuran dehydrogenase, subunit C o GP:X87970_6
>MJECL23
>MJECL28

VII hipotéticas y función desconocida
>MJ0223
>MJ0233
>MJ0280
>MJ0345
>MJ0392 hypothetical protein (GP:D64006_95) o GP:D64006_95
>MJ0401
>MJ0449 hypothetical protein (SP:P46348) o SP:P46348
>MJ0634
>MJ0647
>MJ0650
>MJ0652
>MJ0682
>MJ0690
>MJ0694
>MJ0702
>MJ0704
>MJ0748
>MJ0770
>MJ0835
>MJ0875
>MJ0987
>MJ1007
>MJ1017
>MJ1049
>MJ1053
>MJ1134
>MJ1158
>MJ1185
>MJ1189
>MJ1213
>MJ1221
>MJ1225 hypothetical protein (SP:P15889) SP:P15889
>MJ1254
>MJ1280
>MJ1281



>MJ1290
>MJ1293
>MJ1321
>MJ1322
>MJ1396
>MJ1401
>MJ1412
>MJ1519
>MJ1525
>MJ1556
>MJ1564
>MJ1666
>MJ1674

Saccharomyces cerevisiae

I Metabolismo de macromoléculas

>SW ADR6_YEAST P09547 transcription regulatory protein
>GP-780798 (U24143) ribosomal protein S12
>GP-D1003672 (D14080) elongation factor-1 beta
>PIR:S13653 putative atp-dependent rna helicase
>SW-NSR1_YEAST P27476 nuclear localization sequence binding protein
>SW-SPT8_YEAST P38915 transcription factor spt8
>GP-172076 (M88605) contains Cys-4 zinc-finger, resembles erythroid transcription factor GATA-1
>GP-4358 (X57160) RNA1 protein
>SW-T2EA_YEAST P36100 transcription factor a large subunit
>SW-TBP7_YEAST P40340 tat-binding homolog 7
>SW- TIF3_YEAST P34167 translation initiation factor
>GP-967213 (U28158) Protein involved in decay of mRNA containing nonsense codons
>PIR:S48440 pab-dependent poly(a)-specific ribonuclease (EC 3.1.13.4)
>SW-UME6_YEAST P39001 transcriptional regulator ume6
>GP-641352 (A21198) DNA sequence
>GP-D1007731 (D37935) RNA binding protein
>PIR:A34599 DNA-binding protein
>PIR:S55102 high copy DNA polymerase
>SW-SPT5_YEAST P27692 transcription initiation protein
>SW-TIF3_YEAST P34167 translation initiation
>SW-PRCG_YEAST P22141 proteasome component (ec 3.4.99.46)
>GP-172096 (M90688) polyA nuclease
>GP-530998 (L34569) proline rotamase
>PIR:S50615 U6 snRNA-associated protein USS1
>SW-ERF2_YEAST P05453 eukaryotic peptide chain release factor gtp-binding subunit (erf2)
>SW-FKB3_YEAST (peptidyl-prolyl cis-trans isomerase) (ec 5.2.1.8) (proline rotamase)
>SW-YIN0_YEAST transcriptional regulatory protein in fkh1-sth1
>GP-295671 selected as a weak suppressor of a mutant of the subunit AC40 of DNA dependant RNA polymerase I and III
>PIR:S48244 nonsense-mediated mrna decay protein 2 (up-frameshift suppressor 2)

II Metabolismo intermediario

>GP-E11351 (X01474) polyubiquitin precursor fragment

>PIR-A41035 chitinase (EC 3.2.1.14)
>SW-CHIT_YEAST P29029 endochitinase precursor (ec 3.2.1.14
>PIR:A34796 kinesin-like protein kar3 (nuclear fusion protein)
>PIR:S45578 (increasing suppression factor 1)
>GP-706835 (X84162) deubiquinating enzyme
>PIR-D29456 polyubiquitin 5 - yeast
>PIR:S39110 valosin-containing protein homolog AFG2 - yeast
>SW-CG12_YEAST P20438 g1/s-specific cyclin cln2
>SW-HR25_YEAST P29295 casein kinase i homolog hrr25
>SW-KAI1_YEAST P41944protein kinase a interference protein
>SW-KIP1_YEAST P28742 kinesin-like protein kip1
>SW-SFP1_YEAST P32432 zinc finger protein sfp1
>SW-SLT2_YEAST Q00772 mitogen-activated protein kinase slt2/mpk1 (ec 2.7.1.-)
>SW-SR40_YEAST P32583 yeast). supressor protein srp40
>SW-SR54_YEAST P20424 signal recognition particle 54 kd protein
>GP-171945 (M55016) mating factor alpha
>SW-MFA3_YEAST P01149 mating factor alpha-1 precursor
>SW-MID2_YEAST P36027 mating process protein mid2 (serine-rich protein sms1)
>SW-UCRH_YEAST P00127 yeast). ubiquinol-cytochrome c reductase complex
>SW-MDJ1_YEAST P35191 mdj1 protein precursor
>GP-433624 (X60327) casein kinase-1
>SW-PIR1_YEAST Q03178 pir1 protein precursor
>SW-SED1_YEAST Q01589 protein precursor
>S W-PIR3_YEAST Q03180 pir3 protein precursor
>SW-BEM2_YEAST P39960 gtpase activating protein Bem2p: bud-emergence protein
>GP-173134 (M63498) regulatory protein

III Biosíntesis de moléculas pequeñas
>GP-396560 (X74428) serine-rich protein
>GP-854568 (X87611) adenylate cyclase
>SW-NAB2_YEAST P32505 nuclear polyadenylated rna-binding protein nab2
>SW-ABP1_YEAST P15891 actin binding protein.
>PIR-DNBYPA polyadenylate-binding protein - yeast
>SW-PGD1_YEAST P40356 yeast). poly-glutamine domain protein
>SW-PPZ1_YEAST P26570 serine/threonine protein phosphatase (ec 3.1.3.16)
>SW-PPZ2_YEAST P33329 serine/threonine protein phosphatase (ec 3.1.3.16)
>GP-171044 (M28164) alpha-agglutinin
>GP-633630 (Z47746) probable zinc finger protein
>GP-854474 (Z49810) serine/threonine specific protein phosphatase
>GP-927693 (U33007) serine/threonine protein phosphatase
>SW-KFD3_YEAST P43565 serine/threonine-protein kinase
>SW-KKR1_YEAST P36003 serine/threonine-protein kinase
>SW-KNQ1_YEAST P53894 serine/threonine-protein kinase
>PIR-S62057 proline-rich protein LAS17 - yeast
>GP-172586 (L04655) serine/threonine protein kinase
>SW-AGA1_YEAST P32323 a-agglutinin attachment subunit precursor
>PIR-ALBYG glucan 1,4-alpha-glucosidase (EC 3.2.1.3)
>SW-AMYH_YEAST P08640 (glucan 1,4-alpha- glucosidase)
>SW-BCK1_YEAST Q01389 serine/threonine protein kinase
>SW-NRK1_YEAST P38692 serine/threonine-protein kinase nrk1 (ec 2.7.1.-)
>SW-SSN6_YEAST P14922 glucose repression mediator protein

>SW-ST20_YEAST Q03497 serine/threonine-protein kinase ste20 (ec 2.7.1.-)
>STT4_YEAST P37297 phosphatidylinositol 4-kinase stt4 (ec 2.7.1.67)
>PIR-DNBYPA polyadenylate-binding protein

IV Procesos celulares
>SW-USO1_YEAST P25386 intracellular protein transport protein
>SW-WEB1_YEAST P38968 protein transport protein sec31
>PIR:S70292 FUN12 protein transport protein
>GP-4018 myosin heavy chain type II
>GP-4047 (X56084) nitrogen permease reactivator protein
>SW-CC24_YEAST P11433 cell division control protein
>PIR:S22853 datin oligo(a)/oligo(t)-binding
>SW-CKS1_YEAST P20486 (cell division control protein cks1)
>SW-CTR1_YEAST P49573 copper tranport protein ctr1 (copper transporter 1)

V Estructura celular
>GP-3730 (X53424) glycolipid-anchored surface protein
>GP-1230689 (U51033) Weak similarity to Cell surface glycoprotein precursor of Halobacterium Halobium
>SW-CWP1_YEAST P28319 cell wall protein cwp1 precursor
>SW-EM70_YEAST P32802 endosomal p24a protein precursor (70 kd endomembrane protein) (pheromone alpha-factor transporter)
>GP-172790 (M21129) omnipotent suppressor
>GP-2258166 (U20618) regulatory protein of adenylate cyclase
>GP-311109 (L16900) intrastrand crosslink recognition protein
>PIR-S77699 inner cell wall mannoprotein
>SW-IXR1_YEAST P33417 intrastrand crosslink recognition protein (structure-specific recognition protein)
>SW-SLA1_YEAST P32790 cytoskeleton assembly control protein sla1
>SW-PUB1_ YEAST P32588 nuclear and cytoplasmic polyadenylated rna-binding protein pub1
>GP-4076 (Z15036) nuclear pore complex protein
>GP-496731 (Z32672) nucleoporin
>GP-642341 (X83099) GLFG motif nucleoporin
>SW-N145 _YEAST P49687 nucleoporin
>SW-NU49_ YEAST Q02199 nucleoporin
>SW-NU57_ YEAST P48837 nucleoporin
>SW-NOP3_YEAST Q01560 nucleolar protein 3 (mitochondrial targeting supressor 1 protein)
>SW-N159_YEAST P40477 nucleoporin
>PIR: nucleoskeletal-like protein
>PIR:S52700 nuclear pore protein
>PIR:A35622 nuclear pore protein

VI Otras funciones
>GP-871535 (X87806) verprolin
>GP-415259 (L15626) dynein
>PIR:S25194 zuotin – yeast
>GP-D1020704 (AB003521) flocculin
>GP-E304948 (A39780) Flocculation protein
>SW-R167_YEAST P39743 reduced viability upon starvation protein 167.

>SW-HKR1_YEAST P41809 hansenula mrakii killer toxin-resistant protein 1 precursor
>SW-HS82_YEAST P02829 heat shock protein hsp82
>SW-DR48_YEAST P18899 ddr48 stress protein (dna damage-responsive protein 48) (flocculent specific protein)
>GP-E304948 (A39780) Flocculation protein
>SW-FLO1_YEAST P32768 flocculation protein flo1 precursor
>SW-SFL1_YEAST P20134 flocculation suppression protein
>PIR: (multicopy suppression of a budding defect 2)
>SW-TIR1_YEAST P10863 cold shock induced protein tir1 precursor
>SW-TIR2_YEAST P33890 cold shock induced protein tir2 precursor
>GP-171388 (M36110) DDR48 stress protein
>SW-CG12_YEAST P20438 g1/s-specific cyclin cln2
>GP-42684 (S42682) 46-kDa negative regulator of the Ras-cyclic AMP pathway=RPI1
>GP-472525 (L31766) synthetic lethal 39
>SW-ALK1_YEAST P43633 dna damage-responsive protein alk1
>SW-H150_YEAST P32478 150 kd heat shock glycoprotein precursor
>GP-854443 (Z49808) Ddr48p FSP=flocculent specific protein
>SW-DR48_YEAST P18899 ddr48 stress protein (dna damage-responsive protein 48) (ddrp 48) (flocculent specific protein)

VII hipotéticas y función desconocida

>SW-GTS1_YEAST P40956 gts1 protein
>PIR-S33>SW-IFH1_YEAST P39520 ifh1 protein (rrp3 protein)651 protein - yeast
>PIR:S05807 SAN1 protein - yeast
>PIR:S62061 SCD5 protein - yeast
>SW-SED4_YEAST P25365 sed4 protein
>GP-832919 (X71664) But2
>SW-YKU1_YEAST (Z28201) ORF YKL201c MH1
>GP-809098 (Z49260) Bul1p
>GP-927738 (U33050) Tom1p; CAI: 0.16
>GP-984964 (U20237) Sik1p
>GP-171536 (M16717) open reading frame
>PIR:S54016 SOK2 protein - yeast
>GP-2340996 (U21094) Ylr437cp
>GP-E251944 (Z74917) ORF YOR009w
>GP-3413 (X16385) Baf1 protein (AA 1-731) unknown
>GP-3781 (X03245) ded1 unknown
>GP-D1002336 (D11088) Cks1 protein
>GP-D1002337 (D11088) Cks1.2 protein
>GP-D1002338 (D11088) Cks1.3 protein
>SW-SOK1_YEAST P40317 sok1 protein
>SW-SP41_YEAST P38904 yeast spp41 protein
>PIR:S61992 SLG1 protein - yeast
>GP-809086 (Z49260) unknown
>GP-817890 (Z49703) unknown
>GP-825569 (Z49705) unknown
>GP-E221836 (Z69382)) unknown
>GP-E318906 (Z28080) ORF YKL081w
>GP-E183305 (Z49809) unknown
>GP-E242679 (X97751) unknown
>GP-E243293 (Z72598) ORF YGL076c

>GP-E245463 (Z73136) ORF YLL031c
>GP-E245807 (Z73286) ORF YLR114c
>GP-E246768 (Z73616) hypothetical 119.5 kDa protein unknown
>SW-YB59_YEAST P38311 hypothetical 12.1 kd in dur1,2-ngr1 intergenic region
>GP-E252085 (Z75128) ORF YOR220w
>GP-E252143 (Z75217) ORF YOR310c
>GP-E252970 (Z74053) ORF YDL005c unknown
>GP-E253063 (Z74201) ORF YDL153c
>GP-E308403 (Z74105) ORF YDL058w
>GP-E312758 (Z72557) ORF YGL188c
>GP-171485 (L07289) gene product
>GP-896461 (U31331) Tel1p
>GP-E312887 (Z72946) ORF YGR159c
>GP-E322120 (Z73259) ORF YLR086w
>GP-E332821 (Z73562) ORF YPL207w
>SW-GDS1_YEAST P41913 gds1 protein - yeast
>GP-S47695_1 (S47695) YBL03-15B
>SW-SCJ1_YEAST P25303 scj1 protein. - yeast unknown
>SW-SIS1_YEAST P25294 sis1 protein unknown
>SW-YB59_YEAST P38311 hypothetical 12.1 kd in dur1,2-ngr1 intergenic region
>SW-YBC9_YEAST P38201 hypothetical 42.6 kd in aac2-rpl19 intergenic region
>SW-YCC8_YEAST P25367 hypothetical 42.6 kd in bik1-fus1 intergenic region
>SW-YEA7_YEAST P40002 hypothetical 72.5 kd in gcn4-wbp1 intergenic region
>SW-YEN7_YEAST P39955 hypothetical 100.3 kd in mei4-caj1 intergenic region
>SW-YFB0_YEAST P43582 hypothetical 22.7 kd in aua1-cdc4 intergenic region
>SW-YFC3_YEAST P43573 hypothetical 91.4 kd in ste2-frs2 intergenic region
>SW-YG1F_YEAST P53214 hypothetical 57.5 kd in vma7-rps31a intergenic region
>GP-2340996 (U21094) Ylr437
>SW-MTH1_YEAST P35198 MTH1 protein - yeast
>GP-4814 gene product
>GP-E190152 (X89715) gene product
> PIR:A36426 SPA2 protein - yeast
>GP-3712 (Y00829) gene product
>GP-914986 (U32445) gene product
>GP-861110 (X87672) gene product
>GP-567922 (L36903) varl gene product
>GP-4550 gi|4550 (X66278) gene product
>GP-496695 (X79489) CDC27 D-618 protein
>GP-559524 (X78344) CAT8
>SW-YHT1_YEAST P38835 hypothetical 95.1 kd in act5-yck1 intergenic region
>SW-YHV8_YEAST P38853 hypothetical 131.1 kd in rec104-sol3 intergenic region
>SW-YIM9_YEAST P40468 hypothetical 269.9 kd in fkh1-sth1 intergenic region
>SW-YJ9P_YEAST P47179 hypothetical 118.4 kd in rps7b-dal5 intergenic region
>SW-YJU7_YEAST P39526 hypothetical 229.9 kd in nuc1-prp21 intergenic region
>SW-YK22_YEAST P36135 hypothetical 46.9 kd in gap1-nap1 intergenic region
>SW-YK25_YEAST P36138 hypothetical 21.1 kd in gap1-nap1 intergenic region
>SW-YKR2_YEAST P36049 hypothetical 49.7 kd protein in gin2-ste3 intergenic region
>SW-YNF8_YEAST P53947 hypothetical 35.0 kd in nop2-omp2 intergenic region
>SW-YTP2_YEAST P38429 hypothetical 22.9 kd protein in tps3 5'region (orf2)
>GP-1072409 (Z54141) unknown
>GP-1002516 (U27358) Hgh1p

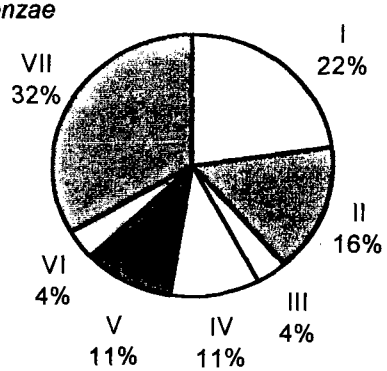
>GP-171374 (M60415) DAL81
>GP-172526 (M16165) S1 protein
>GP-2264354 (U22383) Ylr466wp
>GP-2340034 (U19028) Ylr338wp
>GP-343949 (J01525) var1(40.0)
>GP-544510 (U14913) Ylr194cp
>GP-587539 (Z46373) unknown
>GP-625104 (U19729) unknown
>GP-663232 (Z48148) hypothetical 137.7 kD protein in subtelomeric Y' repeat region
>GP-677198 (L03188)
>GP-78626 (S78624) YCR592
>SW-N100 _YEAST Q02629 unknown
>GP-806319 (M22580) unknown protein
>GP-854468 (Z49809) unknown
>GP-854597 (X87611) ORF YJR83
>GP-927801 (U33057) Ydr534cp
>GP-971261 (U34775) Yer190wp
>GP-984176 (Z54139) unknown
>SW-BMH2_YEAST P34730 unknown
>GP-E190749 (Z50178) unknown
>GP-E236693 (Z71255) unknown
>GP-E236826 (Z71255) unknown
>GP-E245755 (Z73145) ORF YLL040c
>GP-E245920 (Z73256) ORF YLR084c
>GP-E247047 (Z73514) ORF YPL158c
>GP-E247049 (Z73519) ORF YPL163c
>GP-E251848 (Z74772) ORF YOL030w unknown
>GP-E251930 (Z74897) ORF YOL155c
>GP-E252202 (Z75291) ORF YOR383c
>GP-E252294 (Z74847) ORF YOL105c
>GP-E252336 (Z74961) ORF YOR053w
>GP-E252449 (Z75290) ORF YOR382w
>GP-E252988 (Z74083) ORF YDL035c unknown
>GP-E252991 (Z74087) ORF unknown
>GP-E304627 (Z35789) ORF YBL029w
>GP-E304709 (Z35948) ORF YBR078w
>GP-E312776 (Z72946) ORF YGR160w
>GP-E312885 (Z72905) ORF YGR119c
>GP-E312896 (Z72902) ORF YGR116w
>PIR-S54624 ROD1 protein - yeast
>PIR-S61046 ARP1 protein - yeast
>PIR-S61977 RLM1 protein - yeast unknown
>SW-ST50_YEAST P25344 YCL032w
>SW-YAF3_YEAST P39719 hypothetical 87.5 kd in acs1-gcv3 intergenic region
>SW-YBF1_YEAST P34217 hypothetical 73.8 kd in shp1-sec17 intergenic region
>SW-YBF3_YEAST P38190 very hypothetical 13.2 kd in shp1-sec17 intergenic region
>SW-YBV8_YEAST P38266 hypothetical 92.8 kd in ymc2-cmd1 intergenic region
>SW-YFG7_YEAST P43537 hypothetical 16.5 kd protein in thi5 5'region unknown
>SW-YG42_YEAST P53297 hypothetical 78.8 kd protein in erg1-rnr4 intergenic region
>SW-YG46_YEAST P53301 hypothetical 52.8 kd in bub1-hip1 intergenic region
>SW-YG5W_YEAST P53335 hypothetical 31.3 kd in taf145-yor1 intergenic region

>SW-YKK5_YEAST P34250 hypothetical 125.6 kd in aat1-gfa1 intergenic region
>SW-YGG7_YEAST P53164 hypothetical 43.5 kd in rpb9-alg2 intergenic region
>SW-YHU3_YEAST P38844 hypothetical 33.4 kd in rpl44-dcd1 intergenic region
>SW-YIM3_YEAST P40472 hypothetical 48.1 kd protein in kgd1-rpi1 intergenic region
>SW-YIO9_YEAST P40457 hypothetical 195.1 kd in dna43-ubi1 intergenic region
>SW-YIQ9_YEAST P40442 hypothetical 99.7 kd in suc2 5'region precursor unknown
>SW-YIR7_YEAST P40434 hypothetical 197.5 kd protein in suc2 5'region. unknown
>SW-YJH4_YEAST P47037 hypothetical 141.3 kd in scp160-mrpl8 intergenic region
>SW-YJH6_YEAST P47035 hypothetical 128.5 kd in scp160-mrpl8 intergenic region
>SW-YJH8_YEAST P47033 hypothetical 89.2 kd in scp160-mrpl8 intergenic region
>SW-YJQ0_YEAST P46999 hypothetical 19.0 kd in tpk1-far1 intergenic region
>SW-YJZ3_YEAST P47094 hypothetical 15.3 kd in mer2-cpr7 intergenic region
>PIR:S38082 pathogenesis-related protein hypothetical 33.8 kd intergenic region
>SW-YN23_YEAST P53832 hypothetical 52.3 kd in mrpl10-erg24 intergenic region
>SW-YNT0_YEAST P53872 hypothetical 22.0 kd in chs1-srp1 intergenic region
>SW-YNW8_YEAST P53862 hypothetical 28.6 kd in ure2-ssu72 intergenic region
>GP-1165299 (U43834) Ydr544cp
>GP-544503 (U14913) Ylr206wp
>GP-694125 (L28920) cloning and analysis of the yeast
>GP-728653 (Z48613) unknown
>GP-915002 (U32517) Ydr326cp
>GP-927751 (U33050) Ydr474cp; CAI: 0.12
>SW-AS10_YEAST P48361 ask10 protein
>GP-E217728 (X94335) YOR3162c
>GP-E223636 (Z46727) Sec7p
>GP-E239064 (X96876) putative ORF
>GP-E245461 (Z73133) ORF YLL028w
>GP-E245541 (Z73258) ORF YLR086w
>GP-E245804 (Z73278) ORF YLR106c
>GP-E245919 (Z73172) ORF YLL067c
>GP-E247083 (Z73541) ORF YPR204w
>GP-E251834 (Z74749) ORF YOL007c
>GP-E251917 (Z74876) ORF YOL133w
>GP-E252273 (Z74793) ORF YOL051w
>GP-E252431 (Z75217) ORF YOR309c
>GP-E252990 (Z74085) ORF YDL037c unknown
>GP-E253056 (Z74188) ORF YDL140c
>SW-EGT2_YEAST P42835 10/96 ORF YNL327w
>PIR-S37788 PIR3 protein - yeast unknown
>SW-UTR2_YEAST P32623 utr2 protein (unknown transcript 2 protein)
>SW-YAG3_YEAST P39712 hypothetical 138.1 kd in flo9-gdh3 intergenic precursor
>SW-YBI1_YEAST P38180 hypothetical 40.8 kd in rhk1-pet112 intergenic region
>SW-YBM6_YEAST P38216 hypothetical 14.6 kd protein in ttp1-gal7 intergenic region
>SW-YBY0_YEAST P38272 hypothetical 47.4 kd in vma2-cks1 intergenic region
>SW-YCS2_YEAST P25356 hypothetical 251.0 kd in cry1-gns1 intergenic region
>SW-YCX9_YEAST P25653 hypothetical 166.0 kd protein in abp1 3'region
>SW-YEW2_YEAST P32634 hypothetical 195.4 kd in rps26b-glc7 intergenic region
>SW-YEW2_YEAST P32634 hypothetical 195.4 kd pin rps26b-glc7 intergenic region
>SW-YG31_YEAST P53269 hypothetical 27.2 kd protein in clb6-spt6 intergenic region
>SW-YG3R_YEAST P53288 hypothetical 22.2 kd in nsr1-tif4631 intergenic region
>SW-YG4G_YEAST P42939 hypothetical 12.0 kd in ade3-ser2 intergenic region

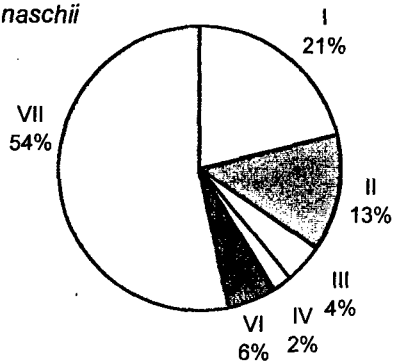
>SW-YG5V_YEAST P53334 hypothetical 40.2 kd in taf145-yor1 intergenic region
>SW-YHC8_YEAST P38739 hypothetical 63.8 kd in gut1-rim1 intergenic region
>SW-YHF0_YEAST P38721 hypothetical 79.0 kd in cbp2 5'region
>SW-YJP9_YEAST P47000 hypothetical 30.8 kd in tpk1-far1 intergenic region
>SW-YK82_YEAST P36170 hypothetical 122.2 kd in sir1 3'region precursor
>SW-YKB4_YEAST P34241 hypothetical 203.3 kd in put3-cce1 intergenic region
>SW-YKF4_YEAST P35732 hypothetical 84.0 kd in nup120-cse4 intergenic region
>SW-YKS7_YEAST P34231 hypothetical 81.0 kd in pat1-mtr2 intergenic region
>SW-YKU2_YEAST P36042 hypothetical 21.2 kd in tor2-pas1 intergenic region
>SW-YN48_YEAST P42846 hypothetical 68.7 kd in stb1-mck1 intergenic region
>SW-YN96_YEAST P53753 hypothetical 121.1 kd in bio3-hxt17 intergenic region
>SW-YNJ1_YEAST P53935 hypothetical 141.5 kd in ypt53-rho2 intergenic region
>SW-YNR6_YEAST P53882 hypothetical 67.4 kd in rps3-psd1 intergenic region
>SW-YSY2_YEAST P24089 hypothetical 137.7 kd in subtelomeric y' repeat region of chr xv
>GP-E223720 (Z68194) unknown
>SW-PC11_YEAST P39081 pcf11 protein unknown
>PIR-S30839 UTR2 protein - yeast
>PIR-S67070 GAC1 protein - yeast
>SW-YHT1_YEAST P38835 hypothetical 95.1 kd in act5-yck1 intergenic region 1
>SW-YIB1_YEAST P40552 hypothetical 26.3 kd in pdr11-faa3 intergenic region
>SW-YNC7_YEAST P53968 hypothetical 76.3 kd zinc finger in hhf2-ume3 intergenic region

En la (Fig. 4) se muestra una comparación porcentual de los grupos funcionales con respecto a los tres genomas microbianos.

Haemophilus influenzae



Methanococcus jannaschii



Saccharomyces cerevisiae

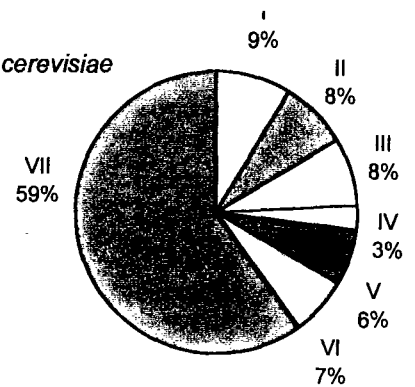


Fig. 3. Distribución porcentual de secuencias simples para cada genoma de acuerdo a su función metabólica: I Metabolismo de macromoléculas, II Metabolismo Intermediario, III Biosíntesis de moléculas pequeñas, IV Procesos celulares, V Estructura celular, VI Otras funciones, VII Hipotéticas y función desconocida.

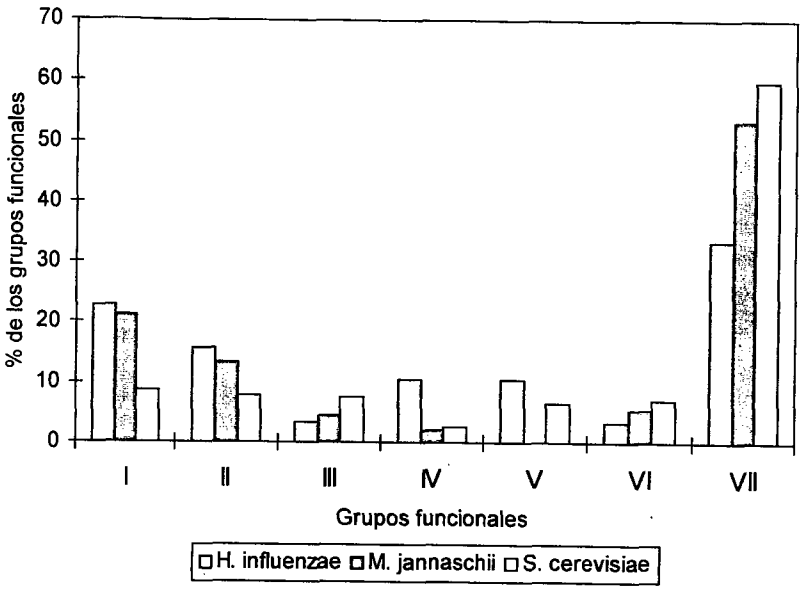
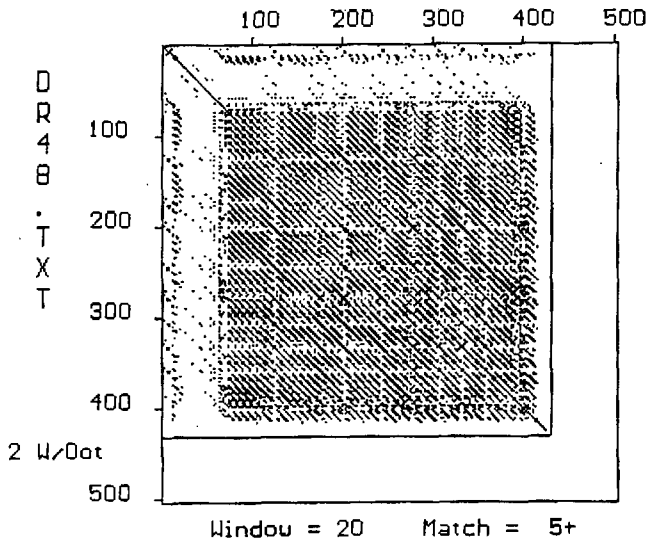


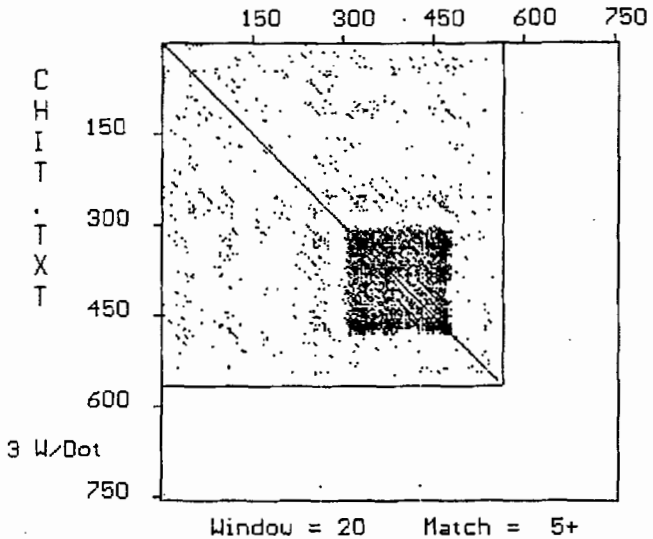
Fig. 4. Agrupamiento de la distribución funcional de los tres genomas microbianos. I. Metabolismo de macromoléculas, II. Metabolismo intermediario, III. Biosíntesis de moléculas pequeñas, IV. Procesos celulares, V. Estructura celular, VI. Otras funciones, VII Hipotéticas y de función desconocida.

Con el fin de obtener solamente secuencias con una marcada baja complejidad para los tres genomas, se realizó una selección posterior de todas aquellas secuencias que tuvieran un valor K de 2.6 a 2.8 y una ventana mayor a 30 con el programa SEG, dando como resultado un total de 86 secuencias simples con baja complejidad; para *H. influenzae* se tienen 13 secuencias, en *M. jannaschii* 12 y en *S. cerevisiae* 61 secuencias simples.

El método cuantitativo de SEG tanto como el análisis gráfico (*dot-plot)* muestran regiones amplias con muy baja complejidad en las 86 secuencias analizadas de los tres genomas microbianos en la (Fig. 5) se muestran solamente algunas de las secuencias con su respectiva matriz de puntos, las restantes están en el (Apéndice V).

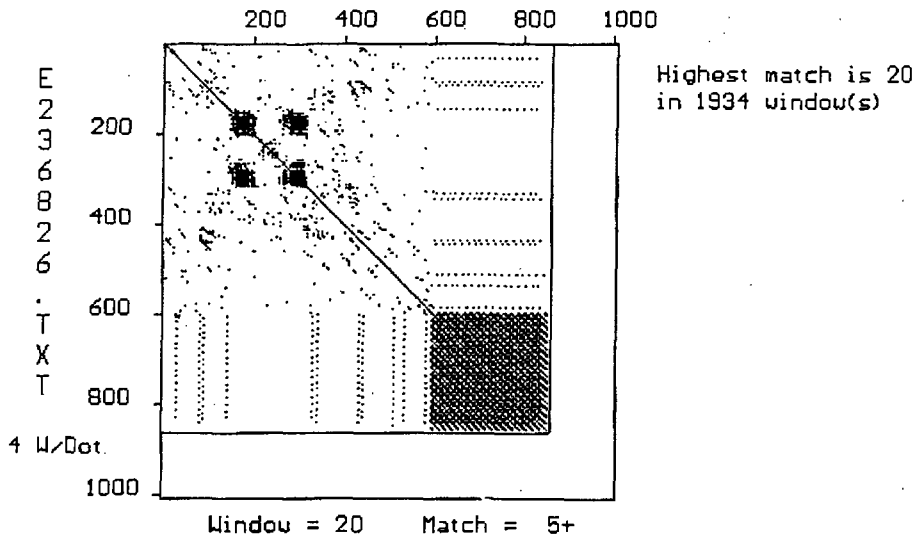


>SW-DR48_YEAST SW:DR48_YEAST P18899 saccharomyces cerevisiae (baker's yeast). ddr48 stress protein (dna damage-responsive protein 48) (ddrp 48) (yp 75) (flocculent specific protein). 10/96
GLFDKVKQFANSNNNNNDSGNNNQGDYVTKAENMIGEDRVNQFKSKIGEDRFDKMESKVR
QQFSNTSINDNDSNNNDSYGSNNNDSYGSNNNDSYGSNNNDSYGSNNNDSYGSNNDDSYG
SSNKKKSSYGSNNDDSYGSSNNNDSYGSNNNDSYGSNNNDSYGSNNDDSYGSSNKNKSSY
GSNNDDSYGSNNDDSYGSSNKKKSSYGSSNNDSYGSNNDDSYGSNNNDSYGSNNDDSYGS
SNKKKSSYGSNNDDSYGSSNNNDSYGSNNDDSYGSSNKNKSSYGSSSNDDSYGSSNNDDS
YGSSNKKKSSYGSNNDDSYGSNNDDSYGSSNKKKSSYGSSNNDSYGSNNDDSYGSSNKKK
SSYGSNNDDSYGSSNNNDSYGSNNDDSYGSSNRNKNSYGSSNYGSSNNDDSYGSSNRGGR
NQYGGDDDY

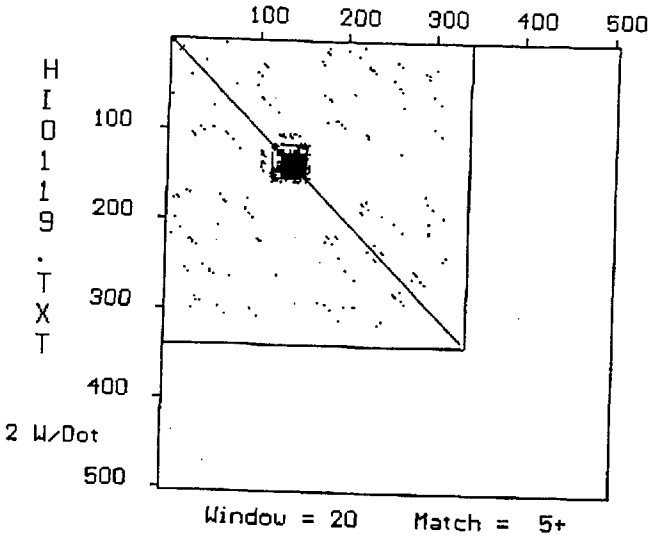


>SW-CHIT_YEAST SW:CHIT_YEAST P29029 saccharomyces cerevisiae (baker's yeast). endochitinase precursor (ec 3.2.1.14). 10/96; gi|596043 (U17243) Endochitinase 2 (Saccharomyces cerevisiae)

MSLLYIILLFTQFLLLPTDAFDRSANTNIAVYWGQNSAGTQESLATYCESSDADIFLLSF
LNQFPTLGLNFANACSDTFSDGLLHCTQIAEDIETCQSLGKKVLLSLGGASGSYLFSDDS
QAETFAQTLWDTFGEGTGASERPFDSAVVDGFDFDIENNNEVGYSALATKLRTLFAEGTK
QYYLSAAPQCPYPDASVGDLLENADIDFAFIQFYNNYCSVSGQFNWDTWLTYAQTVSPNK
NIKLFLGLPGSASAAGSGYISDTSLLESTIADIASSSSFGGIALWDASQAFSNELNGEPY
VEILKNLLTSASQTATTTVATSKTSAASTSSASTSSASTSQKKTTQSTTSTQSKSKVTLS
PTASSAIKTSITQTTKTLTSSTKTKSSLGTTTTESTLNSVAICSMKTTLSSQITSAALVT
PQTTTTSIVSSAPIQTAITSTLSPATKSSSVVSLQTATTSTLSPTTTSTSSGSTSSGSTS
SDSTARTLAKELNAQYAAGKLNGKSTCTEGEIACSADGKFAVCDHSAWVYMECASGTTCY
AYDSGDSVYTQCNFSYLESNYF

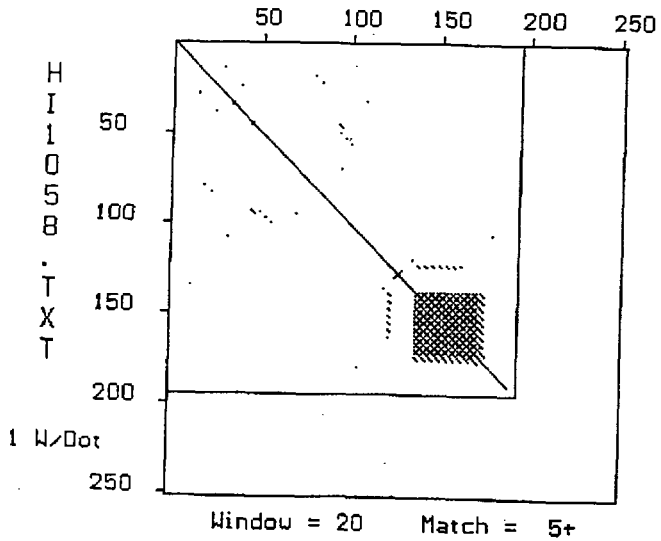


>GP-E236826 gi|1314096|gnl|PID|e236826 (Z71255) unknown [Saccharomyces cerevisiae]; gi|809587 (Z49274) unknown [Saccharomyces cerevisiae]
MHGKELAGRLRKRENDNDLSPNSSSSPAERFRCPHPECNKTFSRQEHLSRHKLNHWPKEI
YVCSYVLPTTNAPCNKTFVRKDLLIRHEKRHSKVKNRLSRPSKDQISSSNKDFSKNAPYN
PSEVPLSTQSGTSTINLIKNSVNPPPSITQESKFRPFLQQAQQPQQVQQSQQPQQIQQLQ
QLQFPQQLRAPLQQPMLQQQMHPQQASPTFPSYDPRIRNNGQNGNQFFNLIFDNRTGVNG
FEVDAANNNGNGNDQNMNINPAVQQQRYQDRNFASSSYQQPLQPLTQDQQQEQYFQQQKL
AQQQQQQQQQQQQQQQLPPQNPFGDPLTSSSSGANLSVMQDLFSTNFLNSDPLQSFMQEL
SEAPQVSIEDTFSDKNTIPPNEKPVQQDEGFQNPPVMFELPQDNIKIPKAQPKENDNPST
SVKDNLSSQKLNINELKRRSSKDSGVGNNSSLNYKEQLRHSMKSVPSFFHPDDPLTKYKIS
KEKCQEMFSFVPELRYVSIESIHKSLKSFWLNFKPQYGLLHKPSFHVDKQPAILNLALIM
TGASFLGSEYREQISDPICGPLRWIIFSHADFQPPSKTYIIQSLLLVEGYEKTSTNRYLH
ERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLH
ERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLH
ERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLH
ERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLHERSFLHHGSTNRYLH
ERSFLHHGSTNRYLHERSF

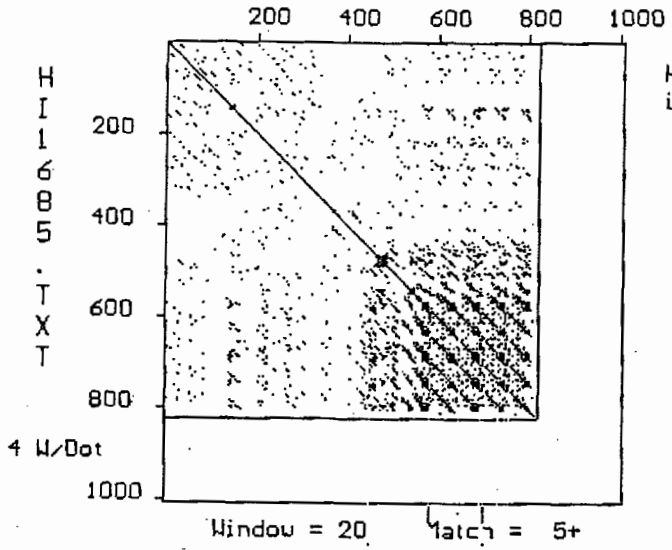


HI0119.

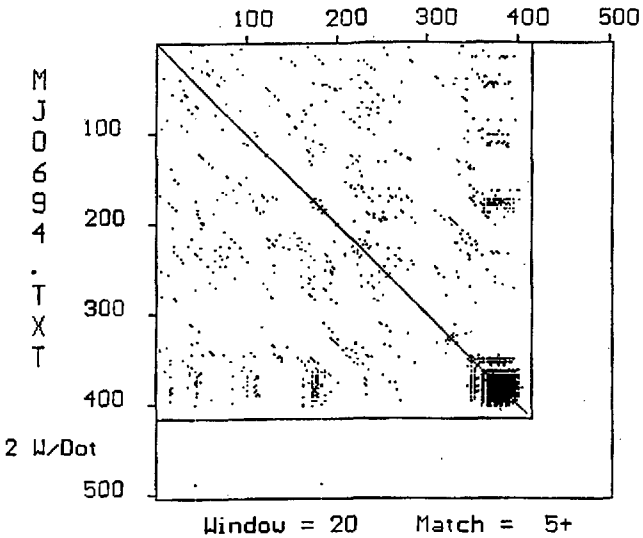
MKKLLKISAI SAALLSAPMM ANADVLASVK PLGFIVSSIA DGVTGTQVLV PAGASPHDYN
LKLSDIQKVK SADLVVWIGE DIDSFLDKPI SQIERKKVIT IADLADVKPL LSKAHHEHFH
EDGDHDHDHK HEHKHDHKHD HDHDHDHKHE HKHDHEHHDH DHHEGLTTNW HVWYSPAISK
IVAQKVADKL TAQFPDKKAL IAQNLSDFNR TLAEQSEKIT AQLANVKDKG FYVFHDAYGY
FNDAYGLKQT GYFTINPLVA PGAKTLAHIK EEIDEHKVNC LFAEPQFTPK VIESLAKNTK
VNVGQLDPIG DKVTLGKNSY ATFLQSTADS YMECLAK



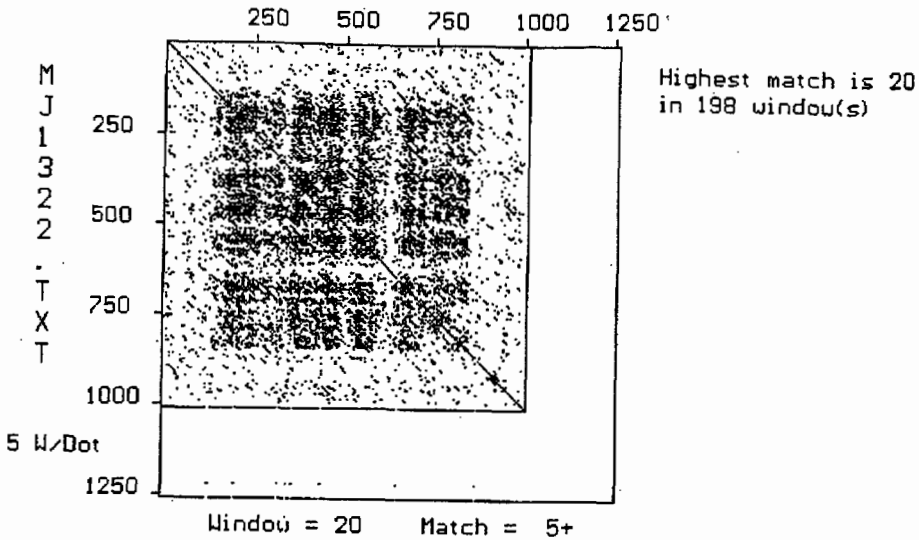
HI1058.
MKTDIQTELT QALLSHEKVW ANEEKTILAK NILLDLVEKT DPTIIGLLLG NDDLKRHFFV
EVNGVLVFKL QDFRFFLDKH SINNSYTKYA NRIGLTDGNR FLKDSSDIVL DFPFKDCVLN
GGQSTEEGEE IYFKRNNSQS VSQSVSQSVS QSVSQSVSQS VSQSVSQSVS QSVSQSVSQS
IIHQINPKKT RNLF



>HI1685 outer membrane integrity protein (tolA) (Escherichia coli)
MADVLSRFNSGKLWDFKGGIHPPEMKSQSNSQPLRHLPLGTDFYIPLKQHLGTTGNLLIK
EGDYVLKGQALTKGDGLRMLPVHAPTSGTIKSIKPYVATHPSGLDEPTIHLQADGLDQWI
ERNPIDDFSTLSSEQLIHKIYQAGIAGLGGAVFPTAAKIQSAEQKVKLLIINGAECEPYI
TCDDRLMRERADEIIKGIRILRYILHPEKVVIAIEDNKPEAISAIRNALQGANDISIRVI
PTKYPSGATKQLIYLLTGIEVPSGERSSSIGVLMQNVGTMFAIKRAIINDEPLIERVVTL
TGNKIAEKGNYWVRLGTPISQILSDAGYQFDKHFPIFAGGPMMGLELPNLNAPVTKLVNC
LLAPDYLEYAEPEAEQACIRCSSCSDACPVNLMPQQLYWFARSEDHKKSEEYALKDCIEC
GICAYVCPSHIPLIQYFRQEKAKIWQIKEKQKKSDEAKIRFEAKQARMEREEQERKARSQ
RAAQARREELAQTKGEDPVKAALERLKAKKANETESTQIKTLTSEKGEVLPDNTDLMAQR
KARRLARQQAASQVENQEQQTQPTNAKKAAVAAALARAKAKKLAQANSTSEAISNSQTAE
NQVEKTKSAVEKTQENSTALDPKKAAVAAAIARAKAKKLAQTNSTSEAISNSQTAENEVE
KTKSAVEKTEENSTALDAKKAAIAAAIARAKAKKLAQANSASEAISNSQTAENEVEKTKS
AVEKTQQNSTALDPKKAAVAAAIARAKAKKLAQANST3EAISNSQTAENEVEKTKSAVEK
TQENSTALDPKKAAVAAAIARAKAKKLAKTQATLENNQE

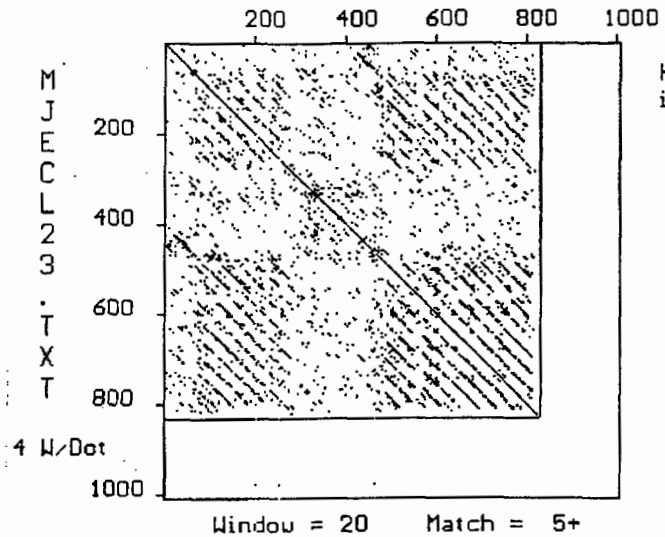


>MJ0694
LIYVTFTPYGAFGVKDNKEVSGLEDIEYKKLFNEEEIPDIMFKLKTQPNKIADELKEEWG
DEIKLETLSTEPFNIGEFLRNNLFKVGKELGYFNNYDEFRKKMHYWSTELTKKVIKSYAQ
QKDKIIIQVAEAISDLDKTLNLLSERLREWYSLYFPELDHLVNKHEVYANLITKLGKRKN
FTKSQLKKILPSKLAGKIAEAAKNSMGGELEDYDLDVIVKFAEEINHLYEKRKELYNYLE
KLMNEEAPNITKLAGVSLGARLIGLAGGLEKLAKMPASTIQVLGAEKALFAHLRMGVEPP
KHGIIYNHPLIQGSPHWQRGKIARALACKLAIAARADYVGDYIADELLEKLNKRVEEIRR
KYPKPPKKKKKEKPKAKKKEKKGKKEKSKKKKDKKKDKKGKKERKVIGKTKSRK

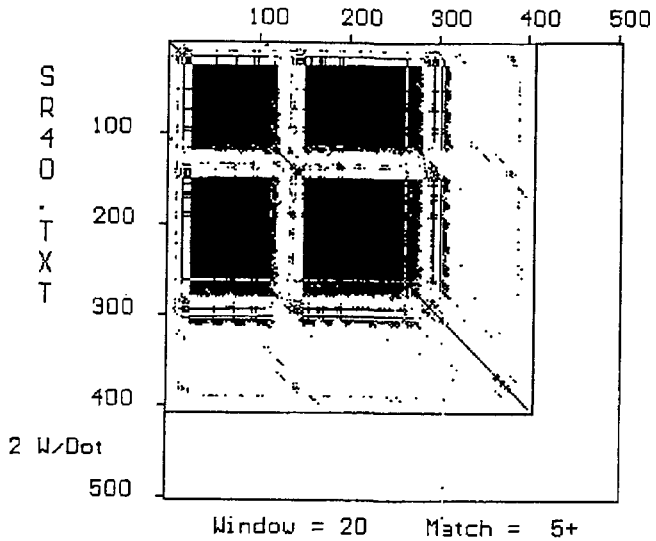


>MJ1322

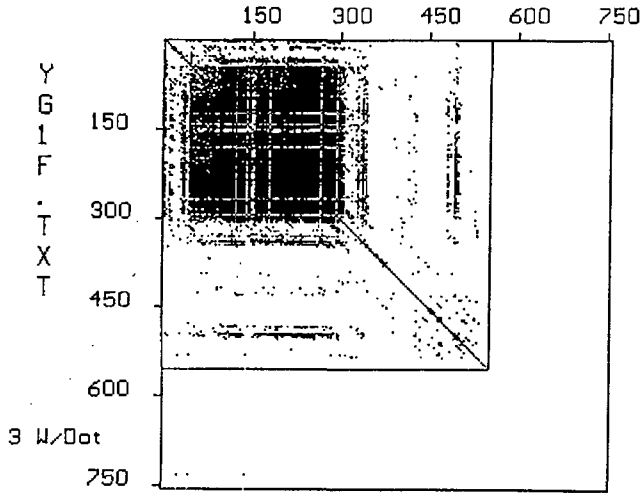
MSMILKEIRMNNFKSHVNSRIKFEKGIVAIIGENGSGKSSIFEAVFFALFGAGSNFNYDT
IITKGKKSVYVELDFEVNGNNYKIIREYDSGRGGAKLYKNGKPYATTISAVNKAVNEILG
VDRNMFLNSIYIKQGEIAKFLSLKPSEKLETVAKLLGIDEFEKCYQKMGEIVKEYEKRLE
RIEGELNYKENYEKELKNKMSQLEEKNKKLMEINDKLNKIKKEFEDIEKLFNEWENKKLL
YEKFINKLEERKRALELKNQELKILEYDLNTVVEARETLNRHKDEYEKYKSLVDEIRKIE
SRLRELKSHYEDYLKLTKQLEIIKGDIEKLKEFINKSKYRDDIDNLDTLLNKIKDEIERV
ETIKDLLEELKNLNEEIEKIEKYKRICEECKEYYEKYLELEEKAVEYNKLTLEYITLLQE
KKSIEKNINDLETRINKLLEETKNIDIESIENSLKEIEEKKKVLENLQKEKIELNKKLGE
INSEIKRLKKILDELKEVEGKCPLCKTPIDENKKMELINQHKTQLNNKYTELEEINKKIR
EIEKDIEKLKKEIDKEENLKTLKTLYLEKQSQIEELELKLKNYKEQLDEINKKISNYVIN
GKPVDEILEDIKSQLNKFKNFYNQYLSAVSYLNSVDEEGIRNRIKEIENIVSGWNKEKCR
EELNKLREDEREINRLKDKLNELKNKEKELIEIENRRSLKFDKYKEYLGLTEKLEELKNI
KDGLEEIYNICNSKILAIDNIKRKYNKEDIEIYLNNKILEVNKEINDIEERISYINQKLD
EINYNEEEHKKIKELYENKRQELDNVREQKTEIETGIEYLKKDVESLKARLKEMSNLEKE
KEKLTKFVEYLDKVRRIFGRNGFQAYLREKYVPLIQKYLNEAFSEFDLPYSFVELTKDFE
VRVHAPNGVLTIDNLSGGEQIAVALSLRLAIANALIGNRVECIILDEPTVYLDENRRAKL
AEIFRKVKSIPQMIIITHHRELEDVADVIINVKKDGNVSKVKING



>MJECL23
LVMRMVFVENKINGVENIRQEIDNLNVSNYKEIYDKFKNVFSKNRDIFNYYVDKLIELMK
NLDDNELSLEIDKFMLCILKDSIIWEFKAGGSVWDLSIKDIKDNIIILGCSNHLFALDIK
TGNKIWEYKVEHNVDSLFIKDNIVMLEYRGGHVCVLDVMTGDKIWESKVGERMWGFSLKD
NIVILGDGDKYIYAIDVRTGGKLWEFEAEWCVWELSIKDDKIILRCEDEYGCEYFYVLDI
KTGEKILEFGGEWHVSDLLISGDVTILADMWGCVYALDTNLYSKIQRVRPQLTNIMKEIV
KIDLTLLKKSLNLNEWDELPIQITNKSLKDITISKISIINEEDILFKDIEPIKIRGRDTK
VINLFINPKVKGKLPIDIVVEFEDEFNIRYKERFTEVLTITKFKGDNVDDMRPEKIDKIL
KQEIDNLNTSNYKEIYSRFKNIFNENRAVENYYMNKLIELVNNSNDELAINVGKFILDIL
GIKRVDELLWEFRAEGGVRLLSIKGDIVILGCVSGHVYAIDIKTGKKLWEFKAEDTVWGL
SIKDDIVVLGCGNIFESIVMLKNGKILEEGYAYALDINTGREIWRSKIKHDVRSLSIKDD
IVVLGCKKGYILALDINAGNMLWEFKAKSGKSIRNLSIKNDILLFGCDNYLYALDIDTGR
ELWRFKAEGEVKSLSIKKDNVLLGCRGGYVYLLDINTGEKMERFKVVGSVLRLSIKDDIV
ILGCNRECVYALDINAGENLWAFKTDGDVNGLSIKNDAVLLGCDNYLYALDINTGEEIWK
FKTESAVLDLSIKDNIVISGCKRGHVYALDFNIIKNYSIIQKIKQVL



>SW-SR40_YEAST SW:SR40_YEAST P32583 saccharomyces cerevisiae (baker's yeast). supressor protein srp40. 6/94; PIR:S38170 SRP40 protein - yeast (Saccharomyces cerevisiae); gi|486581 (Z28317) ORF YKR092c
MASKKIKVDEVPKLSVKEKEIEEKSSSSSSSSSSSSSSSSSSSSSSSSSSGESSSSSSSSS
SSSSSDSSDSSDSESSSSSSSSSSSSSSSSSDSESSSESDSSSSGSSSSSSSSSDESSSES
ESEDETKKRARESDNEDAKETKKAKTEPESSSSSESSSSGSSSSSESESGSESDSDSSSS
SSSSSDSESDSESDSQSSSSSSSSSDSSSDSDSSSSDSSSSDSDSSSSSSSSSSDSDSDSDS
SSDSDSSGSSDSSSSSDSSSDESTSSDSSDSDSDSDSGSSSELETKEATADESKAEETPA
SSNESTPSASSSSSANKLNIPAGTDEIKEGQRKHFSRVDRSKINFEAWELTDNTYKGAAG
TWGEKANEKLGRVRGKDFTKNKNKMKRGSYRGGSITLESGSYKFQD




```
>SW-YG1F_YEAST SW:YG1F_YEAST P53214 saccharomyces cerevisiae (baker's
yeast). hypothetical 57.5 kd protein in vma7-rps31a intergenic region.
10/96; gi|1322994|gnl|PID|e243927 (Z72807) ORF YGR023w [Sac
MASCNPTRKKSSASSLSMWRTILMALTTLPLSVLSQELVPANSTTSSTAPSITSLSAVES
FTSSTDATSSASLSTPSIASVSFTSFPQSSSLLTLSSTLSSELSSSSMQVSSSSTSSSSS
EVTSSSSSSSISPSSSSSTIISSSSSLPTFTVASTSSTVASSTLSTSSSLVISTSSSTFT
FSSESSSSLISSSISTSVSTSSVYVPSSSTSSPPSSSSELTSSSYSSSSSSSTLFSYSSS
FSSSSSSSSSSSSSSSSSSSSSSYFTLSTSSSSSIYSSSSYPSFSSSSSSNPTSSITST
SASSSITPASEYSNLAKTITSIIEGQTILSNYYTTITYSPTASASSGKNSHHSGLSKKNR
NIIIGCVVGIGAPLILILLILIYMFCVQPKKTDFIDSDGKIVTAYRSNIFTKIWYFLLGK
KIGETERFSSDSPIGSNNIQNFGDIDPEDILNNDNPYTPKHTNVEGYDDDDDDDANDENL
SSNFHNRGIDDQYSPTKSASYSMSNSNSQDYNDADEVMHDENIHRVYDDSEASIDENYYT
KPNNGLNITNY
```

Fig. 5. Se muestran 11 secuencias de aminoácidos con su respectiva matriz de puntos, de los tres genomas microbianos analizados con el programa sclone.

De los análisis estadísticos realizados con el programa SAPS para cada secuencia se arrojaron los siguientes resultados (Cuadro 3). Se encontró que el triptofano (W) nunca participa en la baja complejidad en ninguno de los tres genomas; mientras que la cisteína (C) participa solamente en *M: jannaschii* en una proteína la ferrodoxina. Los aminoácidos que se encuentran con más frecuencia en la baja complejidad varían en los tres genomas microbianos (Fig. 6).

Cuadro 3. Análisis con el programa SAPS

a) Composición de aminoácidos

Organismo	aa. con alta composición (++/+)	aa. con baja composisción (--/-)
H. influenzae	SANTDIVEKQAHN	MPFLGR
M. jannaschii	NDIEKFY	AMSTPVQR
S. cerevisiae	EKHGASNTDQ	PMHDVRL

b) Distribución de cargas

organismo	% de secuencias con segmentos positivos	% de secuencias con segmentos negativos	% de secuencias con segmentos mezclados
H. influenzae	7.69	0	15.38
M. jannaschii	8.33	8.33	25
S. cerevisiae	13.11	21.31	37.70

c) Distribución de otros aminoácidos tipo

organismos	% de segmentos con aminoácidos hidrofóbicos	% de segmentos con aa. de transmembrana
H. influenzae	15.38	38.46
M. jannaschii	8.33	25
S. cerevisiae	37.70	55.73

d) Estructuras repetidas

organismo	% de secuencias que presentan patrón	% del promedio de repeticiones en las secuencias
H. infkuenzae	84.61	3.36
M. jannaschii	100	6.16
S. cerevisiae	98.36	9.9

e) Multitipletes

organismo	% de multitip. extensos	% de altitipletes extensos	% de multitip. significativos	% de altitipletes sifnificativos	aminoácidos
H. influenzae	0	15.38	0	0	DH, EP
M. jannaschii	8.33	8.33	0	16.66	K, PC,GC,EW
S. cerevisiae	63.93	8.47	13.11	16.94	TSNQH, GS,GY,FQ,CW,GN,AY,GS ,KC,RN,RY,VP,LY,AP,GQ, PW,FG,GH,SF,ER,EY,NW ,DF,DG,LF,KD,SY,ST,SF

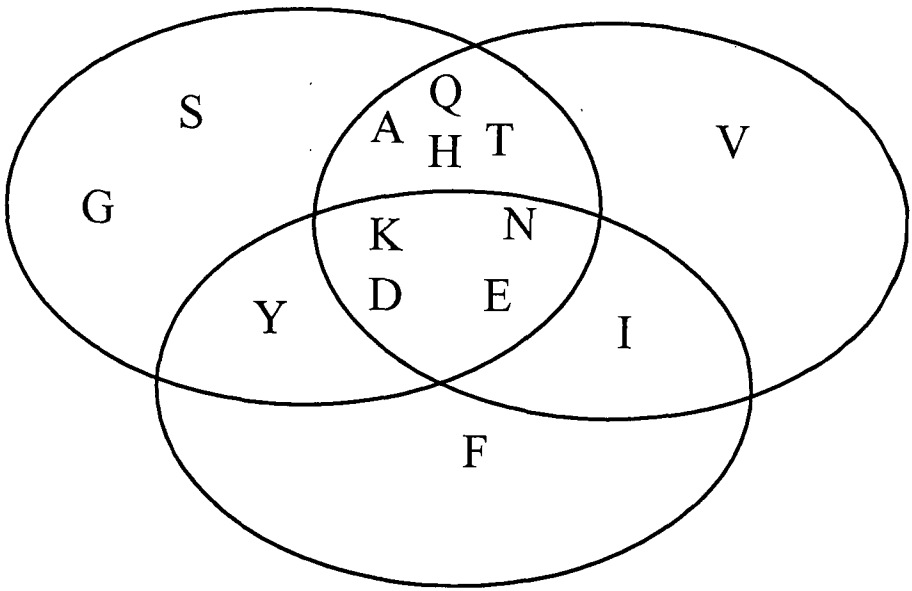
f) Periodicidad

organismo	% de sec. con un alto número de copias	aminoácidos
H. influenzae	58.84	HQKEASV
M. jannaschii	58.33	IKE
S. cerevisiae	77.04	QSNTEGKVDYH

g) Espacialidad

organismo	% de eventos excesivos en la distribución de aminoácidos	aminoácidos
H. influenzae	61.53	SINAK
M. jannaschii	25	KIC
S. cerevisiae	70.49	NQSTHEP

Saccharomyces cerevisiae *Haemophilus influenzae*



Methanococcus jannaschii

Fig. 6. Aminoácidos que se encuentran con más frecuencia en cada genoma Entre los aminoácidos que influyen en la baja complejidad de *H. influenzae* destacan, aspargina (N), lisina (K), y valina (V); para *M. jannaschii*, lisina (K), isoleucina (I) y ácido glutámico (E), y en *S. cerevisiae*, serina (S), treonina (T), glicina (G), glutamina (Q), ácido glutámico (E), aspargina (N) y lisina (K).

4. DISCUSION

4.1 Distribución de secuencias simples en los tres genomas microbianos.

Los análisis muestran una amplia distribución de las secuencias simples en los genomas de *H. influenzae*, *M. jannaschii* y *S. cerevisiae,* presentando diferencias en la frecuencia y composición de la baja complejidad, siendo una característica relevante en el genoma eucarionte.

La distribución funcional de las secuencias simples no esta restringida a un solo grupo funcional, sino que se encuentra en todos los grupos que propone Riley (1993). Las secuencias simples se encuentran distribuidas principalmente en el grupo funcional VII (proteínas hipotéticas y de función desconocida), pero esto puede ser provocado por el desconocimiento que aún se tiene de la biología de los genomas completos. Otro grupo funcional con una frecuencia significativa de secuencias simples es el I (metabolismo de macromoléculas). Estos dos grupos pueden ayudar a conocer el papel que han tenido las secuencias simples a lo largo de la evolución de las proteínas debido a que en el grupo de macromoléculas se encuentran proteínas con un origen anterior al último ancestro común, lo que sugiere la antigüedad del fenómeno; por otro lado la presencia de secuencias simples, en proteínas hipotéticas sugiere un papel importante del fenómeno en secuencias que no tienen una alta presión de selección.

El principal fenómeno responsable de la producción de las secuencias simples es el patinaje de la polimerasa (Bebenek y Kunkel, 1990; Richards y Sutherland, 1994; Hancock, 1995; Epplen y Riess, 1997). La creación y elongación de las secuencias simples ocurre durante la replicación del DNA (Schlötterer y

Tautz, 1991). Siendo el patinaje de la polimerasa una característica muy antigua de sistemas de replicación (Li y Nicolaou, 1994; Sievers y von Kiedrowski, 1994) es difícil estimar que tan frecuente se presenta el patinaje de la polimerasa en las especies y que tan antiguo es, pero el encontrar que este fenómeno está presente en proteínas primordiales como: DNA polimerasas, ATP-sintetasas, topoisomeras, ferrodoxinas y proteínas ribosomales, sugiere que el patinaje de la polimerasa precede al ultimo ancestro común de los seres vivos y por lo tanto es más antiguo que los organismos patógenos, hecho que establece que este fenómeno no es una adaptación de la patogénesis (Moxon *et al.*, 1994; High *et al.*, 1996; Saunders *et al.*, 1998).

La ocurrencia de las secuencias simples en varias funciones, la antigüedad de algunas proteínas y su presencia en los tres dominios, sugiere que el proceso de patinaje de la polimerasa no es solo un importante proceso en la evolución de genomas completos, sino también en el origen y evolución de proteínas.

La incorporación de secuencias simples en el genoma ha contribuido al incremento del tamaño del genoma durante la evolución, incorporando los productos del patinaje de la polimerasa (Hancock, 1996). Existe la posibilidad que al incorporarse los productos del patinaje de la polimerasa en la proteína resultante esta tenga una expresión diferente a la original, teniendo una nueva función.

En los experimentos desarrollados *in vitro* (Kornberg *et al.*, 1964), para la síntesis de la DNA polimerasa se observa un alto rango de mutaciones producidas por el patinaje de la polimerasa, pero al ser comparada con los experimentos *in vivo* (Levinson y Gutman, 1987) el número de mutaciones es muy bajo, por lo que

se ha sugerido que en eucariontes no solo actúa el patinaje de la polimerasa, sino que pueden existir algunos otros mecanismos de *turnover*, siendo el patinaje de la polimerasa el mecanismo que mejor explica la evolución del genoma (Levinson y Gutman, 1987b; Hancock, 1996). El que otros mecanismos pudieran participar en el incremento de las proteínas es muy viable, debido a que el genoma de los eucariontes es mucho más grande, lo que le permite una plasticidad de adecuación a varios mecanismos, a diferencia de los procariontes en los cuales pareciera que solo se presenta el patinaje de la polimerasa

4.2 Consecuencias de la incorporación de las secuencias simples en el genoma.

En los últimos años el interés por las secuencias simples ha estado ligado a patogénesis (Moxon *et al.*, 1994, Moxon y Wills, 1999) Los niveles elevados del patinaje de la polimerasa podría ocasionar un crecimiento incontrolado, dando como resultado enfermedades como el cáncer. Varios organismos patógenos como son: *Haemophilus infuenza*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, entre otros presentan pérdida ó ganancia de repeticiones producto del patinaje de la polimerasa, estas puede generar fases de variación y modular alternativas de expresión en los genes (Moxon *et al.*, 1994). Podrían existir genes que actuaran en substratos de *slippage*, aumentando la variación por la pérdida o ganancia de repeticiones (Tautz *et al.*, 1986; Hancock y Dover, 1990; Brian y Meyer, 1992), lo que genera mayores posibilidades de adecuación positiva.

La presencia de baja complejidad en secuencias antiguas indica que este fenómeno precede a la patogénesis, lo cual sugiere que el patinaje de la

polimerasa no solo participa como un mecanismo de incremento de variabilidad en casos extremos (Becerra y Lazcano en preparación).

4.3 Distribución de los aminoácidos en las secuencias simples.

La incidencia de aminoácidos cargados, en las secuencias simples, la presencia de segmentos hidrofóbicos y de transmembrana, las estructuras repetidas y la espacialidad puede sugerir una correlación con la estructura de la proteína y su función en diversos eventos biológicos, como pueden ser la formación de dominios en membranas, activación de la transcripción, control de la división celular, en el ensamblaje de ciertas proteínas, en el incremento de su solubilidad en un medio acuoso, en la estabilidad y conformación de las mismas. Los resultados obtenidos del análisis estadístico de las secuencias simples, sugieren que no existe una correlación entre la baja complejidad con su función biológica o sus características físico-químicas, manteniendo diferentes estructuras secundarias en las proteínas en que participan.

La presencia de determinados aminoácidos en las secuencias simples, podría estar vinculada, a la facilidad de poder intercambiar un nucleótido por otro en sus codones.

5. CONCLUSIONES

- Las secuencias simples se encuentran distribuidas en los tres principales linajes celulares (Arquea, Bacteria y Eucaria), presentando también una alta distribución en las funciones biológicas.
- Los aminoácidos que con mayor frecuencia se presentan en las secuencias simples de los tres genomas microbianos son: (K) lisina, (E) ácido glutámico, (N) aspargina y (D) ácido aspártico.
- El análisis visual es una importante herramienta para la determinación de secuencias simples, observando principalmente la baja complejidad en los extremos amino y carboxilo.
- La baja complejidad podría jugar un importante papel en la evolución del tamaño de las proteínas, y quizás en el incremento del número de las mismas, siendo un recurso de variabilidad y posiblemente en la adquisición de nuevas funciones de las proteínas durante la evolución.
- El fenómeno de la baja complejidad es muy antiguo (antes del último ancestro común) precediendo a la patogénesis.

6. REFERENCIAS

Alm, RA, LS. ling, DT. Moir, BL. King, ED. Brown, PC. Doig, DR. Smith, B. Noonan, BC. Guild, BL. deJorge, G. Carmel, PJ. Tummino, A. Caruso, M. Uria-Nickelsen, DM. Mills, C. Ives, R. Gibson, D. Merberg, SD. Mills, Q. Jiang, DE. Taylor, GF. Vovis, TJ. Trust. 1999. Genomics-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori.* **Nature 397(6715):176-80**

Andersson, S. G, A. Zomorodipour, J. O.Andersson, T. Sicheritz-Ponten, U.C. Alsmark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, H. H. Winkler, y C. G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. **Nature 396(6707):133-40**

Blattner, F.R., G.Plunkett, CA. Bloch, NT. Perna, V. Burland, M. Riley, J. Collado-Vides, JD. Glasner, CK. Rode, GF.Mayhew, J. Gregor, NT. Davis, H.A. Kirkpatrick, MA. Goeden, DJ Rose, B. Mau, Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. **Science 277:1453-1474**

Bebenek, K. and Kundel TA. 1990. Frameshift errors initiated by nucleotide misincorporation. **Proc. Natl. Acad. Sci. USA 87:4946-4950**

Brian, D. R. and Meyer T. F. 1992. Genetic variation in pathogenic bacteria. **Reviews 8(12):422-427**

Bult, C. J., White O, G. J. Olsen, L. Zhou, R.D. Fleischmann, G. G.Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayn, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C.I. Reich, R. Overbee, E.F. Kirkness, K.G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N.S.M. Geoghagen, J. C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschi.* **Science 273:1017-1140.**

Clark, S. P., G. A. Evans, y H. R. Garden. 1997. Informatics and Automation Used in Physical Mapping of the Genome. *en*. Smith, D. W. (ed), 19-49 pp. **Biocomputing Informatics and Genome Projects**. Academic Press, Inc.

Cole, ST, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, SV. Gordon, K. Eiglmeier, S.Gas, CE. 3er Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, BG. Barrell, et al., 1998 Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. **Nature 393(6685):537-44**

Deckert, G, PV. Warren, T. Gaasterland, WG. Young, AL. Lenox, DE. Graham, R. Overbeek, MA. Snead, M. Keller, M. Aujay, R. Huber, RA. Feldman, JM. Short, GJ. Olsen, RV. Swanson. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. **Nature 392(6674):353-8**

Epplen, J. T. y O. Riess. 1997. Repetitive sequences in DNA. *en* Bishop, M. J. Raulings (eds), 185-195 pp. **DNA and protein sequnce analysis: a practical approach.** IRL Press, Oxford

Fleischmann, R D, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J.Scott, R. Shirley, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J.L. Fuhrmann, NS. M.Geoghagen, C. L. Gnehm, L.A. McDonald, K. V. Small, C.M. Freiser, H. O. Smith, J.C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. **Science 269, 496-512.**

Fraser, C. M, J. D. Gocayne, O. White, M. D. Adam, R. A. Clayton, R. D. Fleischmann, C. J.Bult, A. R. Kerlavage, G. Sutt, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C Hu, T.S. Lucier, S. N. Peterson, H.

O. Smith, C. A. Hutchinson III, J. C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium*. **Science 270, 397-403.**

Fraser, C. M, S. Casjens, WM. Huan, GG. Sutton, R. Clayton, R. Lathigra, O. White, KA. Ketchum, R. Dodson, EK. Hickey, M. Gwinn, B. Dougherty, J. F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, AR. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, MD. Adams, J. Gocayne, JC. Venter et al., 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. **Nature 390(6660):580-6**

Fraser, CM, SR. Norris, GM. Weinstock, O. White, GG. Sutton, R. Dodson, M. Gwinn, EK. Hickey, R. Clayton, KA. Ketchum, E. Sodergren, JM. Hardham, MP. McLeod, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, JK. Howell, M. Chidambaram, T. Utterback, L. McDonald, P. Artiach, C. Bowman, MD. Cotton, JC. Venter et al., 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. **Science 281(5375):375-88**

Goffeau et. al.,1997.Complete genome yeast. **Nature 387 (Suppl.) 5-105**

Hancok, J. M. and Dover A. G. 1990. "Compensatory slippage" in the evolution of ribosomal RNA genes. **Nucleic Acids Res 18(20):5949-5954**

Hancok, J. M. 1995. The contribution of slippage-like processes to the genome evolution. **J. Mol. Evol. 41:1038-1047**

Hancock, J. M. 1996. Simple sequence and the expanding genome. **BioEssays 18(5)421-425**

High, N. J, M. P. Jennings and E. R. Moxon. 1996. Tandem repeats of the tetramer 5'-CAAT-3' present in lic2A are required for phase variation but not lipopolysaccharide biosynthesis in *Haemophilus influenzae*. **Molecular Biology 20:165-174**

Himmelreich, R, H. Hilbert, H. Plagens, E. Pirkl, BC. Li, R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. **Nucleic Acids Res 24(22):4420-49**

Horowitz, N. H. 1945.On the evolution of biochemical synthesis. **Proc. Natl. Acad. Sci. USA 31:153-157**

Kaneko, T, S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, S. Tabata. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. **DNA Res. 3(3):109-36**

Kawarabayasi, Y, M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, H. Kikuchi. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3 (supplement). **DNA Res 5(2):147-55**

Klenk, HP, RA. Clayton, JF. Tomb, O. White, KE. Nelson, KA. Ketchum, RJ. Dodson, M. Gwinn, EK. Hickey, JD. Peterson, DL. Richardson, AR. Kerlavage, DE. Graham, NC. Kyrpides, RD. Fleischmann, J. Quackenbush, NH. Lee, GG. Sutton, S. Gill, EF. Kirkness, BA. Dougherty, K. McKenney, MD. Adams, B. Loftus, JC. Venter, et al. , 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. **Nature 390(6658):364-70**

Kornberg A, LL. Bertsch, JF. Jackson, HG. Khorana. 1964. Enzymatic synthesis of deoxyribonucleic acid, XVI. Oligonucleotides as templates and the machanism of their replication. **Proc. Natl. Acad. Sci. USA 51:315-323**

Kunst, F, N. Ogasawara, I. Moszer, AM. Albertini, G. Alloni, V. Azevedo, MG. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, SC. Brignell, S. Bron, S. Brouillet, CV. Bruschi, B. Caldwell, V. Capuano, NM. Carter, SK. Choi, JJ. Codani, IF. Connerton, A. Danchin, et al., 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. **Nature 390(6657):249-56**

Levinson, G. and Gutman GA. 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. **Nucleic Acids Res. 15:5323-39**

Levinson, G. and Gutman GA. 1987b. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. **Mol. Biol. Evol. 4(3):203-21**

Li, T. and Nicolaou K. C. 1994. Chemical self-replication of palindromic duplex DNA. **Nature 369:218-220**

Moxon, E. R. P.B. Rainer. M. A. Nowak, and R. E. Lenski. 1994. Adaptative evolution of highly mutable loci in pathogenic bacteria. **Curr. Biol. 4:24-33**

Moxon, R. and Wills C. 1999. DNA microsatellites: Agents of evolution?. **Scientific American Jan 94-99**

Riley, M. 1993. Functions of the Gene Products of *Escherichia coli* **Microbiological Reviews 57(4):862-952**

Richards, R I. and Sutherland GR. 1994. Simlple repeat DNA is not replicated simply. **Nature Genet 6:114-116**

Saunders, N J, Peden J F, Hood D W, Moxon R. 1998. Simple sequence repeats in the *Helicobacter pylori* genome. **Molecular Microbiology 27:1091-1098**

Schlötterer, C. and Tautz D. 1991. Slippage synthesis of simple sequence DNA. **Nucleic Acids Res. 20:211-215**

Sequencing Consortium. 1999. Genome Sequence of Nematode *C. elegans*: A platform for Investigating Biology. **Science 282:2012-2018**

Sievers, D. and von Kiedrowski. 1994. Self-replication of complementary nucleotide-based oligomers. **Nature, 369:221-224**

Smith, DR, LA. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush, JN. Reeve, et al., 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. **J. Bacteriol 179(22):7135-55**

Stephens, RS, S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, RL. Tatusov, Q. Zhao, EV. Koonin, RW. Davis. 1998. Genome sequence of an obligate intracellular pathogen of humans: *chlamydia trachomatis*. **Science 282(5389):754-9**

Tautz, D., Trick M. and Dover G A. 1986. Cryptic simplicity in DNA is a major source of genetic variation. **Nature 322: 652-656**

Tautz D. and Schötterer C. 1994. Simple sequences. **Current Opinion in Genetic and Development 4:832-837**

Tomb, JF, O. White, AR. Kerlavage, RA .Clayton, GG. Sutton, RD. Fleischmann, KA. Ketchum, HP. Klenk, S. Gill, BA. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, EF. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, HG. Khalak, A. Glodek, K. McKenney, LM. Fitzegerald, N. Lee, MD. Adams, JC. Venter et al. ,1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. **Nature 7;388(6642):539-47**

Waley, S.G. 1969. Some aspects of the evolution of metabolic pathways. **Comp.Biochem. Physiol. 30, 1-7**

Woese, C.R. and Fox, G.E. 1977.The concept of cellular evolution. **J. Mol. Evol. 10: 1-6**

Woese, C. R., O. Kandler and M. L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains Archea, Bacteria, and Eukarya. **Proc. Natl. Acad. Sci. USA. 87:4576**

Wootton, J. and Federhen S. 1993. Statiscs of local complexity in amino acid sequences and sequence databases. **Computers Chem. 17:149-163**

Wootoon, J. 1997. Simple sequences of protein and DNA. *en* Bishop, M. J. Raulings (eds), 185-195 pp. **DNA and protein sequence analysis: a practical approach.** IRL Press, Oxford

7. APÉNDICE

I. Formulario.

Basado en la complejidad del estado del vector de una ventana de la secuencia, la complejidad composicional local K_1 , así como la información necesaria por posición, obtenida de la composición de la ventana de un residuo de la secuencia en particular. De un alfabeto de N-residuos (usualmente N=4 o

$$K_1 = \frac{1}{L} \log_N \left(\frac{L!}{\prod_{i=1}^{N} n_i!} \right)$$

20) y una ventana de extensión L:

Donde n_i es N números en la complejidad del estado del vector. El logaritmo es tomado como base N en el rango de 0 a 1. La complejidad es expresada frecuentemente en las unidades de información, los algoritmos son expresados en base 2, bits, base *e* ó nats.

Otra medida de información de la complejidad composicional local, K_2 , usualmente expresada en bits, algunas veces usada en lugar de K_1.

$$K2 = - \sum_{i=1}^{N} \frac{n_i}{L} \left[\log_2 \frac{n_i}{L} \right]$$

K_2 es una aproximación que converge hacia K_1 en una ventana extensa.

Probabilidad del estado de complejidad.

Asumiendo una probabilidad igual en apariencia de los cuatro nucleótidos o de los 20 aminoácidos, la probabilidad P_0 de la ocurrencia de algún estado de complejidad es:

$$P_0 = \frac{1}{N^L} \left(\frac{L!}{\prod_{i=1}^{N} n_i!} \right) \left(\frac{N!}{\prod_{\kappa=0}^{L} r_\kappa!} \right)$$

Donde r_κ es el total de los números en el tiempo, los números K ocurren en el número de n_i de la complejidad del estado del vector. (Wootton y Federhen, 1993).

II. Programa Shell-Unix

```
grep '[a-z]'  archivo30 > file2
grep -n '.' file2 > file3
grep -v '[\>]' file3 > file4
cut -d":" -f1 file4 > file5
awk -f prog1 file5 > file6
grep -n '.' file3 > compar
egrep -f file6 compar > resultado1.1
cut -d":" -f3-8 resultado1.1 > res1.0
grep '[\>]' res1.0 > resultado
rm file*
rm compar
rm resultado1.1
grep -n '.' res1.0 > res2
grep -v '[\>]' res2 > res3
cut -d":" -f1 res3 > res4
awk -f prog1 res4 > res5
grep -n '.' res2 > compar1
egrep -f res5 compar1 > resultado1.2
cut -d":" -f3-8 resultado1.2 > res2.0
grep '[\>]' res2.0 > resultado1
rm res1.0
rm res2
rm res3
rm res4
rm res5
rm compar1
rm resultado1.2
grep -n '.' res2.0 > res2.2
grep -v '[\>]' res2.2 > res2.3
cut -d":" -f1 res2.3 > res2.4
awk -f prog1 res2.4 > res2.5
grep -n '.' res2.2 > compar2
egrep -f res2.5 compar2 > resultado2.2
cut -d":" -f3-8 resultado2.2 > res3.0
grep '[\>]' res3.0 > resultado2
rm res2.0
rm res2.2
rm res2.3
rm res2.4
rm res2.5
rm compar2
rm resultado2.2
grep -n '.' res3.0 > res3.2
grep -v '[\>]' res3.2 > res3.3
cut -d":" -f1 res3.3 > res3.4
awk -f prog1 res3.4 > res3.5
grep -n '.' res3.2 > compar3
egrep -f res3.5 compar3 > resultado3.2
```

```
cut -d":" -f3-8 resultado3.2 > res4.0
grep '[\>]' res4.0 > resultado3
rm res3.0
rm res3.2
rm res3.3
rm res3.4
rm res3.5
rm compar3
rm resultado3.2
prog1
{       s=$1-1        print ":"s":"}
```

Palabras claves para cada grupo funcional.

I. Metabolismo de macromóleculas.

ribosomal, RNA, tRNA, DNA, transcription, recombination, translocation, modification, helicase, rRNA, polimerase, histone, chromosome, elongation, topoisomerase, factor, RNase, exonuclease, endonuclease, peptidase, protease, restriction, ribonuclease, gyrase, UV, primosomal, ribonucleotide, peptidyl, polynucleotide, DNAk, DNAj, DNAc, phosphotransferase, tolA,VacB, translocation, transkelotase, protein synthesis, ribosome, chromatin, nuclear.

II. Metabolismo Intermediario.

degradation, intermediary, aerobic, anaerobic, fermentation, ATP, proton, broad, regulatory, CoA, energy, cytochrome, NADH, dehydrogenase, pyruvate, hydrolase, carbon, hydrogenase, nitrogen, fumarate, chlorohydrolase, NADH-, ferrodoxin, polyferrodoxine.

III. Biosíntesis de moléculas pequeñas.

amino acid, glutamate, nitrogen assimilation, aspartame, pyruvate, glycine, serine, sulfur, aromatic amino acid, histidine, nucleotides, salvages, sugar, cofactor, purine, pyrimidine, ribonucleotides, biotin, folic acid, lipoate, molybdobterin, pantothenase, pyridoxine, pyridoxal, thiamine, riboflavine, thioredoxin, glutaredoxine, glutaredoxine, glutathione, menaquinone,

ubiquinones, heme, porphyrins, fatty acids, lipids, polyamine, glutatione, argininosuccinate, glucose, phosphoribosyl, aminoimidazole, ADP-heptose, carbamoyl, orotate, GMP, GTP, UMP, CTP, AMP, phosphoribosyltransferase, quinolinate, nicotinate, thiamin, phosphotidylglycerophosphate, hemolysin, mannosyltransferase, translocase, erp, xylanase, cobalamin, geranyl, chorismate, alginate, oxaloacetate, uridylate, cytidylytransferase, glycerol, glucosamine, phosphomannomutase, urease, urea, lipoamide, dehydrolipoamide, aliphatic, glucanase, pyridine, L-alanine, picolinate, dihydropicolinate, formate, glycogen, xylulokinase, fucosyltransferase, uridylatecytidylyltransferase, glycolysis, fructose, arginine, fatty, sterol, glycosylation, glyoxylate, phospholipid.

IV. Procesos celulares.

transport, cell division, chemotaxis, secretion, osmotic, enterobactin, permease, molybdate uptake, septum, antibiotic, chemotactic, suppressor, osmotically, potassium, channel, cation, efflux, binding, maturation, Na, Na(+), signal, chaperon, grpE, division, transposase, regulated, antiporter, cell cycle, meiosis, homeostasis.

V. Estructura celular.

cell wall, cytoskeleton, endomembrane, surface, cytoplasmic, pore, outer membrane, periplasmic, murein, UDP, flegellar, lipoprotein.

VI. Otras funciones.

criptic, phage, prophages, colicin, plasmid, drug, radiation, colicins, azaserine, integration, terminus, heat shock, chaperone, PEP, nodulation, stress, Hsp, coenzyme, cytotoxin, pathogenicity, antigen, spore, tolerance, emergence, flocculation.

VII. Hipotéticas y función desconocida.

hypothetical, unknown, ORF.

III. Programa Seg.

DOCUMENTATION OF SEG (FROM 'MAN' PAGE)

NAME

seg - segment sequence(s) by local complexity

SYNOPSIS

seg sequence [W] [K(1)] [K(2)] [-x] [options]

DESCRIPTION

seg divides sequences into contrasting segments of low-complexity and high-complexity. Low-complexity segments defined by the algorithm represent "simple sequences" or "compositionally-biased regions".

Locally-optimized low-complexity segments are produced at defined levels of stringency, based on formal definitions of local compositional complexity (Wootton & Federhen, 1993). The segment lengths and the number of segments per sequence are determined automatically by the algorithm.

The input is a FASTA-formatted sequence file, or a database file containing many FASTA-formatted sequences. seg is tuned for amino acid sequences. For nucleotide sequences, see EXAMPLES OF PARAMETER SETS below.

The stringency of the search for low-complexity segments is determined by three user-defined parameters, trigger window length [W], trigger complexity [K(1)] and extension complexity [K(2)] (see below under PARAMETERS). The defaults provided are suitable for low-complexity masking of database search query sequences [-x option required, see below].

OUTPUTS AND APPLICATIONS

(1) Readable segmented sequence [Default]. Regions of contrasting complexity are displayed in "tree format". See EXAMPLES.

(2) Low-complexity masking (see Altschul et al, 1994). Produce a masked FASTA-formatted file, ready for input as a query sequence for database search programs such as BLAST or FASTA. The amino acids in low-complexity regions are replaced with "x" characters [-x option]. See EXAMPLES.

(3) Database construction. Produce FASTA-formatted files containing low-complexity segments [-l option], or high-complexity segments [-h option], or both [-a option]. Each segment is a separate sequence entry with an informative header line.

ALGORITHM

The SEG algorithm has two stages. First, identification of approximate raw segments of low- complexity; second local optimization.

At the first stage, the stringency and resolution of the search for low-complexity segments is determined by the W, K(1) and K(2) parameters. All trigger windows are defined, including overlapping windows, of length W and complexity less than or equal to K(1). "Complexity" here is defined by equation (3) of Wootton & Federhen (1993). Each trigger window is then extended into a contig in both directions by merging with extension windows, which are overlapping windows of length W and complexity less than or equal to K(2). Each contig is a raw segment.

At the second stage, each raw segment is reduced to a single optimal low-complexity segment, which may be the entire raw segment but is usually a subsequence. The optimal subsequence has the lowest value of the probability P(0) (equation (5) of Wootton & Federhen, 1993).

PARAMETERS

These three numeric parameters are in obligatory order after the sequence file name.

Trigger window length [W]. An integer greater than zero [Default 12].

Trigger complexity. [K1]. The maximum complexity of a trigger window in units of bits. K1 must be equal to or greater than zero. The maximum value is 4.322 (log[base 2]20) for amino acid sequences [Default 2.2].

Extension complexity [K2]. The maximum complexity of an extension window in units of bits. Only values greater than K1 are effective in extending triggered windows. Range of possible values is as for K1 [Default 2.5].

OPTIONS

The following options may be placed in any order in the command line after the W, K1 and K2 parameters:

-a Output both low-complexity and high-complexity segments in a FASTA-formatted file, as a set of separate entries with header lines.

-c [characters-per-line] Number of sequence characters per line of output [Default 60]. Other characters, such as residue numbers, are additional.

-h Output only the high-complexity segments in a FASTA-formatted file, as a set of separate entries with header lines.

-l Output only the low-complexity segments in a FASTA-formatted file, as a set of separate entries with header lines.

-m [length] Minimum length in residues for a high-complexity segment [default 0]. Shorter segments are merged with adjacent low-complexity segments.

-o Show all overlapping, independently-triggered low-complexity segments [these are merged by default].

-q Produce an output format with the sequence in a numbered block with markings to assist residue counting. The low-complexity and high-complexity segments are in lower- and upper-case characters respectively.

-t [length] "Maximum trim length" parameter [default 100]. This controls the search space (and search time) during the optimization of raw segments (see ALGORITHM above). By default, subsequences 100 or more residues shorter than the raw segment are omitted from the search. This parameter may be increased to give a more extensive search if raw segments are longer than 100 residues.

-x The masking option for amino acid sequences. Each input sequence is represented by a single output sequence in FASTA-format with low-complexity regions replaced by strings of "x" characters.

EXAMPLES OF PARAMETER SETS

Default parameters are given by 'seg sequence' (equivalent to 'seg sequence 12 2.2 2.5'). These parameters are appropriate for low-complexity masking of many amino acid sequences [with -x option].

Database-database comparisons:

More stringent (lower) complexity parameters are suitable when masked sequences are compared with masked sequences. For example, for BLAST or FASTA searches that compare two amino acid sequence databases, the following masking may be applied to both databases:

```
seg database 12 1.8 2.0 -x
```

Homopolymer analysis:

To examine all homopolymeric subsequences of length (for example) 7 or greater:

```
seg sequence 7 0 0
```

Non-globular regions of protein sequences:

Many long non-globular domains may be diagnosed at longer window lengths, typically:

```
seg sequence 45 3.4 3.75
```

For some shorter non-globular domains, the following set is appropriate:

```
seg sequence 25 3.0 3.3
```

Nucleotide sequences:

The maximum value of the complexity parameters is 2 (log[base 2]4). For masking, the following is approximately equivalent in effect to the default parameters for amino acid sequences:

```
seg sequence.na 21 1.4 1.6
```


EXAMPLES
The following is a file named 'prion' in FASTA format:

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQP
HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPV
ILLISFLIFLIVG
```

The command line:

```
seg prion
```

gives the standard output below

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR

                                   1-49   MANLGCWMLVLFVATWSDLGLCKKRPKPGG
                                          WNTGGSRYPGQGSPGGNRY
ppqgggggwgqphgggwgqphgggwgqphgg   50-94
                gwgqphgggwgqggg
                                  95-112  THSQWNKPSKPKTNMKHM
       agaaaagavvgglggymlgsams   113-135
                                 136-187  RPIIHFGSDYEDRYYRENMHRYPNQVYYRP
                                          MDEYSNQNNFVHDCVNITIKQH
                tvtttttkgenftet  188-201
                                 202-236  DVKMMERVVEQMCITQYERESQAYYQRGSS
                                          MVLFS
             sppvillisflifliv    237-252
                                 253-253  G
```

The low-complexity sequences are on the left (lower case) and high-complexity sequences are on the right (upper case). All sequence segments read from left to right and their order in the sequence is from top to bottom, as shown by the central column of residue numbers.

The command line:

```
Seg prion  x
```

gives the following FASTA-formatted file:-

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxTHSQWNKPSKPKTNMKHMxxxxxxxx
xxxxxxxxxxxxxxxRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHxxxxxxxxxxxxxxxDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSxxxx
xxxxxxxxxxxxG
```

SEE ALSO
---

segn, blast, saps, xnu

AUTHORS

John Wootton: wootton@ncbi.nlm.nih.gov
Scott Federhen: federhen@ncbi.nlm.nih.gov

National Center for Biotechnology Information
Building 38A, Room 8N805
National Library of Medicine
National Institutes of Health
Bethesda, Maryland, MD 20894
U.S.A.

PRIMARY REFERENCE

Wootton, J.C., Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry 17: 149-163.

OTHER REFERENCES

Wootton, J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Computers & Chemistry 18: (in press).

Altschul, S.F., Boguski, M., Gish, W., Wootton, J.C. (1994) Issues in searching molecular sequence databases. Nature Genetics 6: 119-129.

Wootton, J.C. (1994) Simple sequences of protein and DNA. In: Nucleic Acid and Protein Sequence Analysis: A Practical Approach. (Second Edition, Chapter 8, Bishop, M.J. and Rawlings, C.R. Eds. IRL Press, Oxford) (In press).

IV. Hojas de resultados de una secuencia, con el programa SAPS y dos disquettes con todos los resultados estadísticos de las 86 secuencias simples para los tres genomas microbianos. Lista de nombres de los archivos.

S. cerevisiae 171388, 172526, 311109, 396560, 4076, 544510, 694125, 871535, 899399, aga1, albyg, amyh, chit, e190152, e239064, e245919, e245920, e252294, e252336, e252990, e312887, hkr1, kfd3, mcm1, msb2, n100, n145, nu57, pir3, ppz2, S77699, S61977, S61046, sed1, yag3, ycc8, ycx9, yg1f, yg3r, yh11, yib1, yk82, ynw8, sr40, ssn6, uso1, yir7, yj9p, ykf4, ynj1, ynr6, 251944, 854443, 927801, D1020, dr48, e236826, e252970, e308403, kai1, nsr1 m62

H. influenzae HI066, HI0251, HI0264, HI0383, HI0915, HI1058, HI1232, HI1685, HI1718, HI0119, HI0990, HI1601, HI1284

M. jannaschii MJ1254, MJ1322, MJ1422, MJ1505, MJECL28, MJ1303, MJECL23, MJ0124, MJ0233,MJ0487, MJ0694, MJ0875

>SW-YK82_YEAST SW:YK82_YEAST P36170 saccharomyces cerevisiae (baker's yeast). hypothetical 122.2 kd protein in sir1 3'region precursor. 6/94; PIR:S38181 flocculation protein FLO1 homolog YKR102w - yea
SWISS-PROT ANNOTATION:
ID Unknown CONVERTED; PRT; 1169 AA.
DE unknown.
CC SEQIO retrieval from FASTA-format entry. 19-Jan-1998

number of residues: 1169; molecular weight: 122.2 kdal

```
   1  MPVAARYIFL TGLFLLSVAN VALGTTEACL PAGEKKNGMT INFYQYSLKD SSTYSNPSYM
  61  AYGYADAEKL GSVSGQTKLS IDYSIPCNGA SDTCACSDDD ATEYSASQVV PVKRGVKLCS
 121  DNTTLSSKTE KRENDDCDQG AAYWSSDLFG FYTTPTNVTV EMTGYFLPPK TGTYTFGFAT
 181  VDDSAILSVG GNVAFECCKQ EQPPITSTDF TINGIKPWNA DAPTDIKGST YMYAGYYYPI
 241  KIVYSNAVSW GTLPVSVVLP DGTEVNDDFE GYVFSPDDNA TQAHCSVPNP AEHARTCVSS
 301  ATSSWSSSEV CTECTETEST SYVTPYVTSS SWSSSEVCTE CTETESTSTS TPYVTSSSSS
 361  SSEVCTECTE TESTSYVTPY VSSSTAAANY TSSFSSSSEV CTECTETEST STSTPYVTSS
 421  SWSSSEVCTE CTETESTSYV TPYVSSSTAA ANYTSSFSSS SEVCTECTET ESTSTSTPYV
 481  TSSSSSSSEV CTECTETEST SYVTPYVSSS TAAANYTSSF SSSSEVCTEC TETESTSTST
 541  PYVTSSSWSS SEVCTECTET ESTSYVTPYV SSSTAAANYT SSFSSSSEVC TECTETESTS
 601  TSTPYATSST GTATSFTAST SNTMTSLVQT DTTVSFSLSS TVSEHTNAPT SSVESNASTF
 661  ISSNKGSVKS YVTSSIHSIT PMYPSNQTVT SSSVVSTPIT SESSESSASV TILPSTITSE
 721  FKPSTMKTKV VSISSSPTNL ITSYDTTSKD STVGSSTSSV SLISSISLPS SYSASSEQIF
 781  HSSIVSSNGQ ALTSFSSTKV SSSESSESHR TSPTTSSESG IKSSGVEIES TSTSSFSFHE
 841  TSTASTSVQI SSQFVTPSSP ISTVAPRSTG LNSQTESTNS SKETMSSENS ASVMPSSSAT
 901  SPKTGKVTSD ETSSGFSRDR TTVYRMTSET PSTNEQTTLI TVSSCESNSC SNTVSSAVVS
 961  TATTTINGIT TEYTTWCPLS ATELTTVSKL ESEEKTTLIT VTSCESGVCS ETASPAIVST
1021  ATATVNDVVT VYSTWSPQAT NKLAVSSDIE NSASKASFVS EAAETKSISR NNNFVPTSGT
1081  TSIETHTTTT SNASENSDNV SASEAVSSKS VTNPVLISVS QQPRGTPASS MIGSSTASLE
1141  MSSYLGIANH LLTNSGISIF IASLLLAIV
```

COMPOSITIONAL ANALYSIS (extremes relative to: swp23s.q)

The composition of the input sequence is evaluated relative to the residue usage quantile table specified with the `-s species' flag. Low usage in the 1% quantile is indicated by the label -- (e.g., Y-- means that the input sequence uses tyrosine as little as the 1% least tyrosine containing proteins in the reference set); low usage in the 5% quantile is indicated by the label `-' (e.g., L-); high usage above the 95% quantile point is indicated by the label `+' (e.g., A+); and high usage above the 99% quantile point is indicated by the label `++' (e.g., LIVFM++). The usage is evaluated for all 20 amino acids, positive (KR) and negative (ED) charge, total charge (KRED), net charge (KR-ED), major hydrophobics (LVIFM), and the groupings ST, AGP (encoded by CCN, GCN, and GGN codons), and FIKMNY (encoded by AAN, AUN, UAN, and UUN codons).

```
A  : 72( 6.2%); C  : 35( 3.0%); D  : 30( 2.6%); E  : 85( 7.3%); F  : 30( 2.6%)
G  : 38( 3.3%); H  :  9( 0.8%); I  : 42( 3.6%); K  : 31( 2.7%); L--: 38( 3.3%)
M  : 13( 1.1%); N  : 45( 3.8%); P  : 49( 4.2%); Q- : 18( 1.5%); R--: 11( 0.9%)
S++:283(24.2%); T++:194(16.6%); V  : 90( 7.7%); W  :  9( 0.8%); Y  : 47( 4.0%)

KR    - :   42 (  3.6%);   ED     :  115 (  9.8%);   AGP     :  159 ( 13.6%);
KRED    :  157 ( 13.4%);   KR-ED - :  -73 ( -6.2%);   FIKMNY  :  208 ( 17.8%);
LVIFM - :  213 ( 18.2%);   ST   ++:  477 ( 40.8%).
```

CHARGE DISTRIBUTIONAL ANALYSIS

The distribution of charges in the protein sequence is evaluated in terms of clusters, high scoring segments, and runs and periodic patterns. Clusters indicate regions of typically 30 to 60 residues exhibiting a relatively high charge concentration. For high scoring charge segments, positive scores are assigned to charge residues of the appropriate type and negative scores to all other residues. A significant cumulative positive score again indicates a region of high charge concentration. The cluster method and the scoring method will generally pick out the same segments (with the scoring method often delimiting the segment to a narrower range), conferring robustness to the results. Short segments of high charge concentration are displayed as runs (with errors). Periodic pat-

terns focus on those with charges every second or third position, with possible relevance to amphipathic secondary structures; other periodic patterns are displayed in the general periodicity analysis section of the output.

```
   1  00000+0000 0000000000 000000-000 000-++0000 00000000+- 0000000000
  61  00000-0-+0 0000000+00 0-00000000 0-00000--- 00-0000000 00++00+000
 121  -000000+0- ++-0--0-00 000000-000 0000000000 -00000000+ 0000000000
 181  0--0000000 00000-00+0 -0000000-0 00000+0000 -000-0+000 0000000000
 241  +000000000 0000000000 -00-00--0- 000000--00 0000000000 0-00+00000
 301  00000000-0 00-00-0-00 0000000000 00000-000- 00-0-00000 0000000000
 361  00-000-00- 0-00000000 0000000000 00000000-0 00-00-0-00 0000000000
 421  00000-000- 00-0-00000 0000000000 0000000000 0-000-00-0 -000000000
 481  00000000-0 00-00-0-00 0000000000 0000000000 0000-000-0 0-0-000000
 541  0000000000 0-000-00-0 -000000000 0000000000 0000000-00 0-00-0-000
 601  0000000000 0000000000 0000000000 -000000000 000-000000 000-000000
 661  0000+000+0 0000000000 0000000000 0000000000 0-00-00000 000000000-
 721  0+0000+0+0 0000000000 0000-000+- 0000000000 0000000000 000000-000
 781  0000000000 00000000+0 000-00-00+ 0000000-00 0+0000-0-0 000000000-
 841  0000000000 0000000000 000000+000 00000-0000 0+-0000-00 0000000000
 901  00+00+000- -000000+-+ 0000+000-0 0000-00000 00000-0000 0000000000
 961  0000000000 0-00000000 00-00000+0 -0--+00000 0000-00000 -000000000
1021  000000-000 0000000000 0+00000-0- 0000+00000 -00-0+000+ 0000000000
1081  000-000000 0000-00-00 000-0000+0 0000000000 000+000000 000000000-
1141  0000000000 0000000000 000000000
```

A. CHARGE CLUSTERS.

Positive, negative, and mixed charge clusters are distinguished. In each case, cmin indicates the minimum number of charges required for a significant charge cluster corresponding to the given window size; e.g., cmin = 9/30 or 12/45 or 15/60 means that significance requires at least 9 charges in a segment of 30 (or fewer) residues, or 12 charges in a segment of length 45, or 15 charges in a segment of length 60. In the case of positive and negative charge clusters, these counts refer to net charge, i.e., charges of the opposite sign within the window are counted as -1. The sizes of the clusters are optimized for display to indicate the segment of highest charge concentration, but a minimum size of 20 residues is required. A mixed charge cluster that begins and ends within 15 residues of the endpoints of a pure charge cluster is not displayed (since its significance rests mostly on the charged residues comprising the displayed pure charge cluster), unless the -v (verbose output) flag is set, in which case both the pure and the mixed charge cluster are displayed. On the other hand, pure charge clusters that are embedded in mixed charge clusters are displayed separately (indicated by a * preceding the specification of location).

For each cluster are given its location in the sequence (From, to), the quartile of the location (1st, 2nd, 3rd, or 4th quarter of the sequence), length, count, and t-value (standard deviations above the mean; to accommodate the multiple tests performed, the t-value significance threshold is set to 4.0 for sequences up to 750 residues, to 4.5 for sequences of length 750-1500 residues, and to 5.0 for longer sequences); also indicated are residues comprising at least 10% of the cluster.

Positive charge clusters: not evaluated (frequency of + < 5%, too low)

Negative charge clusters (cmin = 10/30 or 13/45 or 16/60): none

Mixed charge clusters (cmin = 12/30 or 16/45 or 19/60):

```
1) From   92 to  138:   DTCACSDDDATEYSASQVVPVKRGVKLCSDNTTLSSKTEKRENDDCD
                        -00000---00-000000000++00+000-000000+0-++-0--0-
   quartile: 1; size: 47, +count:  6, -count: 11, 0count: 30; t-value:  4.57 *
   S:  6 (12.8%);  T:  5 (10.6%);  D:  8 (17.0%);  ST: 11 (23.4%);
```

B. HIGH SCORING (UN)CHARGED SEGMENTS.

For each scoring scheme (scores assigned to residues as displayed), SAPS displays segments of the sequence with aggregate score exceeding the particular threshold values M_0.01 (1% significance level, segments labeled with **), M_0.05 (5% significance level, segments labeled *), or otherwise as indicated. A minimal segment length is set as shown. The expected score/letter should be sufficiently large negative, and the average information per letter should be sufficiently large positive in order for the scoring statistics to apply properly (the program prints out when the conditions are not met and skips evaluations).

High scoring positive charge segments:

```
score=   2.00 frequency=   0.036  ( KR )
score=   0.00 frequency=   0.000  ( BZX )
score=  -1.00 frequency=   0.866  ( LAGSVTIPNFQYHMCW )
score=  -2.00 frequency=   0.098  ( ED )

 Expected score/letter:  -0.991;    Average information/letter:   3.234
 Minimal length of displayed segments set to:  20

M_0.01=  6.98  (cv=  4.53, lambda=  1.56007, k=  0.46361, x=  2.46;
                90% confidence interval for segment length:   5 +-   4)
M_0.05=  5.94  (x=  1.41)

# of segments (>=20 residues) exceeding M_0.05: none
```

High scoring negative charge segments:

```
score=   2.00 frequency=   0.098  ( ED )
score=   0.00 frequency=   0.000  ( BZX )
score=  -1.00 frequency=   0.866  ( LAGSVTIPNFQYHMCW )
score=  -2.00 frequency=   0.036  ( KR )

 Expected score/letter:  -0.741;    Average information/letter:   1.341
 Minimal length of displayed segments set to:  20

M_0.01= 11.08  (cv=  7.42, lambda=  0.95216, k=  0.32918, x=  3.66;
                90% confidence interval for segment length:  11 +-  10)
M_0.05=  9.37  (x=  1.95)

# of segments (>=20 residues) exceeding M_0.05: none
```

High scoring mixed charge segments:

```
score=   1.00 frequency=   0.134  ( KEDR )
score=   0.00 frequency=   0.000  ( BZX )
score=  -1.00 frequency=   0.866  ( LAGSVTIPNFQYHMCW )

 Expected score/letter:  -0.731;    Average information/letter:   1.966
 Minimal length of displayed segments set to:  20

M_0.01=  6.00  (cv=  3.79, lambda=  1.86344, k=  0.61793, x=  2.21;
                90% confidence interval for segment length:   8 +-   5)
M_0.05=  5.13  (x=  1.34)

# of segments (>=20 residues) exceeding M_0.05: none
```

High scoring uncharged segments:

```
score=   1.00 frequency=   0.866  ( LAGSVTIPNFQYHMCW )
```

```
score=   0.00 frequency=   0.000  ( BZX )
score=  -8.00 frequency=   0.134  ( KEDR )

 Expected score/letter:  -0.209;    Average information/letter:   0.013
 Minimal length of displayed segments set to:  20

M_0.01= 140.63  (cv= 142.93, lambda=  0.04942, k=  0.00897, x= -2.30;
                 90% confidence interval for segment length: 756 +- 748)
M_0.05= 107.65  (x= -35.28)
! average information < .10; too small !
```

C. CHARGE RUNS AND PATTERNS.

The table below shows the charge runs and patterns searched for (* stands for + or -) and the required minimum number of matches to the pattern allowing for at most 0 (lmin0), 1 (lmin1), or 2 (lmin2) mismatches or insertions/deletions (1% significance level). Occurrences are arranged in the order in which they appear in the sequence. For each run or pattern are displayed its length (number of matches) and a triplet giving the number of mismatches, insertions and deletions. 0-runs are further characterized by their composition (residues comprising more than 10% of the run).

Run count statistics are compiled for runs of lengths at least 2/3 of the minimal significant length (lmin0); given are the number and locations of such runs.

pattern	(+)	(-)	(*)	(0)	(+0)	(-0)	(*0)	(+00)	(-00)	(*00)
lmin0	3	5	6	67	8	10	12	9	13	15
lmin1	5	6	7	81	9	12	14	12	16	18
lmin2	6	8	9	90	11	14	16	13	18	20

There are no charge runs or patterns exceeding the given minimal lengths.

Run count statistics:

```
  +  runs >=   3:   0
  -  runs >=   3:   1, at   98;
  *  runs >=   4:   1, at  130;
  0  runs >=  45:   0
```

DISTRIBUTION OF OTHER AMINO ACID TYPES

Routinely, SAPS indicates high scoring hydrophobic and transmembrane segments. The display is as desribed above for high scoring charge segments. The scores for the hydrophobic segments correspond to a digitized hydropathy scale. The transmembrane scores were derived from target frequencies in putative transmembrane proteins (see the paper referred to above; note, however, that the scores used in the program have been rederived and differ from the ones given in the paper). With the -a command line flag, the user can invoke a similar analysis for other residue types. In view of the special role of cysteines for protein structure, the spacings of the cysteine residues in the sequence are displayed separately, with particular emphasis on close pairs of cysteines and distances between such pairs.

1. HIGH SCORING SEGMENTS.

High scoring hydrophobic segments:

```
   2.00 (LVIFM)    1.00 (AGYCW)    0.00 (BZX)   -2.00 (PH)   -4.00 (STNQ)
  -8.00 (KEDR)

 Expected score/letter:  -2.485;    Average information/letter:   1.229
 Minimal length of displayed segments set to:  15

M_0.01= 17.50  (cv= 11.44, lambda=  0.61765, k=  0.42563, x=  6.06;
```

```
                90% confidence interval for segment length:  13 +-   7)
M_0.05= 14.86  (x=  3.43)

# of segments (>=15 residues) exceeding M_0.05: none


High scoring transmembrane segments:

   5.00 (LVIF)   2.00 (AGM)   0.00 (BZX)  -1.00 (YCW)  -2.00 (ST)
  -6.00 (P)   -8.00 (H) -10.00 (NQ) -16.00 (KR) -17.00 (ED)

 Expected score/letter:  -2.927;    Average information/letter:   0.574
 Minimal length of displayed segments set to:  15

M_0.01= 48.28  (cv= 33.99, lambda=  0.20783, k=  0.19609, x= 14.30;
                90% confidence interval for segment length:  25 +-  18)
M_0.05= 40.44  (x=  6.45);      M_0.30= 31.11  (x= -2.88)

 1) From    8 to   24:  length= 17, score=44.00  *
       8  IFLTGLFLLS VANVALG
    L:  5(29.4%);  A:  2(11.8%);  G:  2(11.8%);  V:  2(11.8%);
    F:  2(11.8%);

 2) From 1156 to 1169:  length= 14, score=47.00  *
    1156  GISIFIASLL LAIV
    L:  3(21.4%);  A:  2(14.3%);  S:  2(14.3%);  I:  4(28.6%);

# of segments (>=15 residues) exceeding M_0.30:  2


2. SPACINGS OF C.


H2N-28-C-57-C-6
CAC  at   94
   -22-C-17-C-59
CC   at  197   (l= 105)
  -86-C-11-C-13
CTEC at  311   (l= 118)
    -23
CTEC at  338   (l=  31)
    -23
CTEC at  365   (l=  31)
    -32
CTEC at  401   (l=  40)
    -23
CTEC at  428   (l=  31)
    -32
CTEC at  464   (l=  40)
    -23
CTEC at  491   (l=  31)
    -32
CTEC at  527   (l=  40)
    -23
CTEC at  554   (l=  31)
    -32
CTEC at  590   (l=  40)
    -351-C-4-C-26-C-26-C-4-C-160-COOH

--------------------------------------------------------------------------------
REPETITIVE STRUCTURES.

 Repeats are indicated for two alphabets: the 20-letter amino acid alpha-
 bet, and a reduced 11-letter alphabet in which the major hydrophobics
 LVIF, the charged residues KR and ED, the small residues AG, the hydroxyl
 group residues ST, the amid group residues NQ, and the aromatics YW are
 treated as combined letters. For each alphabet, three classes of repeats
 are distinguished: separated repeats, simple tandem repeats, and periodic
 repeats. The separated repeats are largely non-overlapping. They are
 displayed in groups of matching blocks (exceeding a given core block
```

length of contiguous exact matches) and intervening spacer distances (which may be negative, signifying a partial overlap). The core block length in case of the amino acid alphabet is set to 4 for sequences up to 500 residues, to 5 for sequences between 500 and 2000 residues, and to 6 for longer sequences (same values increased by 4 for the reduced alphabet). Simple tandem repeats are displayed in similar layout, but separately. Sequence segments that are highly repetitive with relatively short repeats are displayed as periodic repeats.

A. SEPARATED, TANDEM, AND PERIODIC REPEATS: amino acid alphabet.
Repeat core block length: 5

Aligned matching blocks:

```
[ 211- 215]   TINGI
[ 965- 969]   TINGI

_____________________________

[ 299- 303]   SSATS
[ 897- 901]   SSATS

_____________________________

[ 303- 322]   SSWSSSEVCTECTETESTSY
[ 330- 349]   SSWSSSEVCTECTETESTST
[ 357- 376]   SSSSSSEVCTECTETESTSY
[ 393- 412]   SFSSSSEVCTECTETESTST
[ 420- 439]   SSWSSSEVCTECTETESTSY
[ 456- 475]   SFSSSSEVCTECTETESTST
[ 483- 502]   SSSSSSEVCTECTETESTSY
[ 519- 538]   SFSSSSEVCTECTETESTST
[ 546- 565]   SSWSSSEVCTECTETESTSY
[ 582- 601]   SFSSSSEVCTECTETESTST

_____________________________

[ 542- 546]   YVTSS
[ 671- 675]   YVTSS

_____________________________

[ 543- 547]   VTSSS
[ 689- 693]   VTSSS

_____________________________

[ 580- 584]   TSSFS
[ 833- 837]   TSSFS

_____________________________

[ 614- 621]   TSFTASTS
[ 841- 847]   TS_TASTS

_____________________________

[ 701- 706]   SESSES
[ 803- 808]   SESSES

_____________________________

[ 886- 892]   SS__ENSAS
[1046-1054]   SSDIENSAS

_____________________________

[ 937- 951]-(   7)-[ 959- 963]
[ 996-1010]-(   7)-[1018-1022]
```

[937- 951] TTLITVSSCESNSCS
[996-1010] TTLITVTSCESGVCS

[959- 963] VSTAT
[1018-1022] VSTAT

Simple tandem repeat:

[328- 390] see sequence above
[391- 453] see sequence above
[454- 516] see sequence above
[517- 579] see sequence above
[580- 609] see sequence above

B. SEPARATED AND TANDEM REPEATS: 11-letter reduced alphabet.
(i= LVIF; += KR; -= ED; s= AG; o= ST; n= NQ; a= YW; p= P; h= H; m= M; c= C)
Repeat core block length: 9

Aligned matching blocks:

[302- 322] oooaooo-ico-co-o-oooa
[329- 349] oooaooo-ico-co-o-oooo
[356- 376] ooooooo-ico-co-o-oooa
[392- 412] ooioooo-ico-co-o-oooo
[419- 439] oooaooo-ico-co-o-oooa
[455- 475] ooioooo-ico-co-o-oooo
[482- 502] ooooooo-ico-co-o-oooa
[518- 538] ooioooo-ico-co-o-oooo
[545- 565] oooaooo-ico-co-o-oooa
[581- 601] ooioooo-ico-co-o-oooo

Simple tandem repeat:

[333- 395] see sequence above
[396- 458] see sequence above
[459- 521] see sequence above
[522- 584] see sequence above
[585- 610] see sequence above

MULTIPLETS.

Multiplets refer to homooligopeptides of any length (e.g., A2, Q7, etc.); altplets refer to reiterations of two different residues (e.g., RG, EAEAEA, etc.). The multiplet composition of the protein sequence is evaluated for both the amino acid and the charge alphabet. (High) Aggregate altplet counts are evalued only for the charge alphabet. The multiplet sequence is displayed whenever the total multiplet count of the sequence falls outside the expected range (i.e., beyond 3 standard deviations of the mean). Printed are also the histogram of the spacings between consecutive multiplets (differences between starting positions) as well as clusters of multiplets (multiplet clusters are determined in the same way as charge clusters are determined; the binomial test is applied to a compressed sequence over the alphabet {M,S}, where M signifies a multiplet and S signifies a singlet; i.e., the amino acid sequence AADFFFGHRRT... is translated as MSMSSMS..., and the binomial cluster test is applied to the latter sequence). Multiplets and altplets of specific residue content that individually show an unusually high count are indicated, and the positions of all multiplets exceeding a minimum length of 5 residues are shown.

A. AMINO ACID ALPHABET.

1. Total number of amino acid multiplets: 106 (Expected range: 78--139)

2. Histogram of spacings between consecutive amino acid multiplets:
(1-5) 42 (6-10) 31 (11-20) 21 (>=21) 13

3. Clusters of amino acid multiplets (cmin = 14/30 or 19/45 or 22/60): none

4. Significant specific amino acid altplet counts:

Letters	Observed (Critical number)				
VY	19 (18)				
at	243 (l= 2)	272 (l= 2)	322 (l= 2)	326 (l= 2)	353 (l= 2)
	376 (l= 2)	380 (l= 2)	416 (l= 2)	439 (l= 2)	443 (l= 2)
	479 (l= 2)	502 (l= 2)	506 (l= 2)	542 (l= 2)	565 (l= 2)
	569 (l= 2)	671 (l= 2)	923 (l= 2)	1031 (l= 2)	

5. Long amino acid multiplets (>= 5; Letter/Length/Position):
S/7/356 S/7/482

B. CHARGE ALPHABET.

1. Total number of charge multiplets: 10 (Expected range: 1-- 22)
3 +plets (f+: 3.6%), 7 -plets (f-: 9.8%)
Total number of charge altplets: 9 (Critical number: 19)

2. Histogram of spacings between consecutive charge multiplets:
(1-5) 1 (6-10) 1 (11-20) 2 (>=21) 7

PERIODICITY ANALYSIS.

The program identifies periodic elements of periods between 1 and 10 for the amino acid alphabet, for the charge alphabet, and for a hydrophobicity alphabet. Each periodic element consists of an error-free core pattern (of length at least 4 for the amino acid alphabet, 5 for the charge alphabet, and 6 for the hydrophobicity alphabet) which is extended allowing for errors. The numbers of errors are given for each position in the consensus of a periodic pattern involving more than one letter. The displayed periodic patterns would generally not be statistically significant but are listed for the sake of a general interactive appraisal of the sequence. Periodicities of exceptionally high copy number are indicated with a !-mark.

A. AMINO ACID ALPHABET (core: 4; !-core: 5)

Location	Period	Element	Copies	Core	Errors
329- 336	2	S.	4	4	0
356- 362	1	S	7	7 !	0
392- 398	1	S	6	4	1
419- 426	2	S.	4	4	0
455- 461	1	S	6	4	1
482- 488	1	S	7	7 !	0
518- 524	1	S	6	4	1
545- 552	2	S.	4	4	0
581- 587	1	S	6	4	1
596- 623	7	T.....	4	4	0
614- 625	3	T..	4	4	0
691- 710	5	S.....	4	4	0
742- 761	5	TS..S	4	4	/0/1/././1/
755- 790	9	SS.S.....	4	4	/0/1/./1/./././././
808- 839	8	S.....	4	4	0
862- 903	6	S.....	6	4	1
900- 915	4	TS..	4	4	/0/1/././
943- 958	4	S...	4	4	0
1077-1092	4	T...	4	4	0
1130-1145	4	S...	4	4	0

B. CHARGE ALPHABET ({+= KR; -= ED; 0}; core: 5; !-core: 6)

and HYDROPHOBICITY ALPHABET ({*= KRED; i= LVIF; 0}; core: 6; !-core: 8)

Location	Period	Element	Copies	Core	Errors

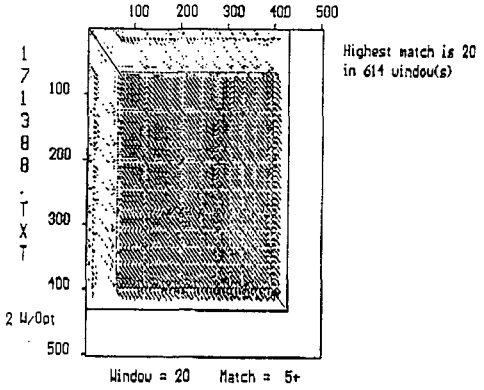
There are no periodicities of the prescribed length.

SPACING ANALYSIS.

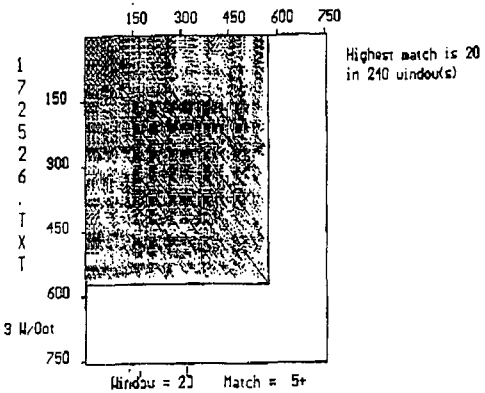
The spacings between consecutive residues of the same type (all 20 amino acids, + and - charge, and combined charge *) are evaluated for significantly large or small maximal and minimal spacings. The output is ordered by the beginning point of the significant spacing. Entries are identified by the residue type, spacing (number of amino acids between the identified positions), rank of the displayed spacing (e.g., 50 alanines in the sequence induce 51 spacings, ranked by decreasing length from 1 to 51), and p-value (probability of exceeding the displayed spacing). A maximal spacing with p-value 0.01 or less is considered significantly large; a maximal spacing with p-value 0.99 or larger is considered significantly small. Similarly, a minimal spacing with p-value 0.99 or larger is considered significantly small, and a minimal spacing with p-value 0.01 or less is considered significantly large (excluding doublets). If the first maximal spacing (rank 1) of a residue is significantly large or small, then also the second maximal spacing (rank 2) is evaluated. Large maximal and small minimal spacings indicate clustering effects, whereas small maximal and large minimal spacings indicate excessive evenness in the distribution of the residues.

Location	(Quartile)	Spacing	Rank	P-value	Interpretation
17- 47	(1.)	S(30)S	2 of 284	0.0014	large 2. maximal spacing
146- 184	(1.)	S(38)S	1 of 284	0.0062	large 1. maximal spacing
241- 665	(2.)	K(424)K	1 of 32	0.0000	large maximal spacing
242- 661	(2.)	I(419)I	1 of 43	0.0000	large maximal spacing
259- 627	(2.)	L(368)L	1 of 39	0.0000	large maximal spacing
271- 611	(2.)	G(340)G	1 of 39	0.0001	large maximal spacing
278- 631	(2.)	D(353)D	1 of 31	0.0005	large 1. maximal spacing
295- 665	(2.)	+(370)+	1 of 43	0.0000	large 1. maximal spacing
593- 945	(3.)	C(352)C	1 of 36	0.0001	large 1. maximal spacing
604- 649	(3.)	P(45)P	2 of 50	1.0000	small 2. maximal spacing
750- 910	(3.)	D(160)D	2 of 31	0.0265	large 2. maximal spacing
925- 989	(4.)	+(64)+	2 of 43	0.9794	small 2. maximal spacing
931- 978	(4.)	P(47)P	1 of 50	1.0000	large 1. maximal spacing
1009-1170	(4.)	C(161)C	2 of 36	0.0065	large 2. maximal spacing

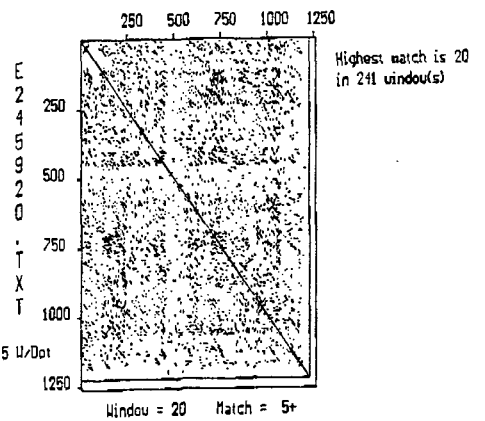
V. Matriz de puntos de las 86 secuencias de aminoácidos de los tres genomas microbianos.
Saccharomyces cerevisiae



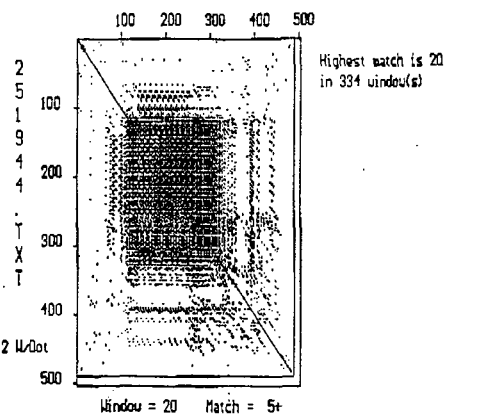
>GP-171388 gi|171388 (M36110) DDR48 stress protein [Saccharomyc cerevisiae]



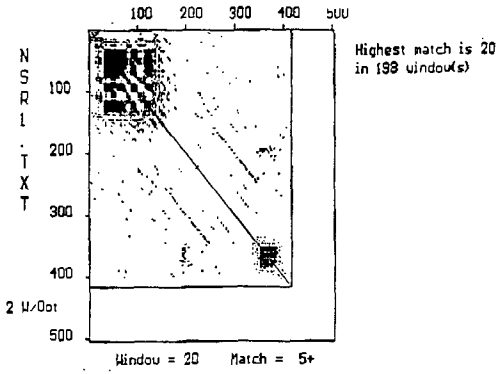
>GP-172526 gi|172526 (M16165) S1 protein [Saccharomyces cerevisiae]



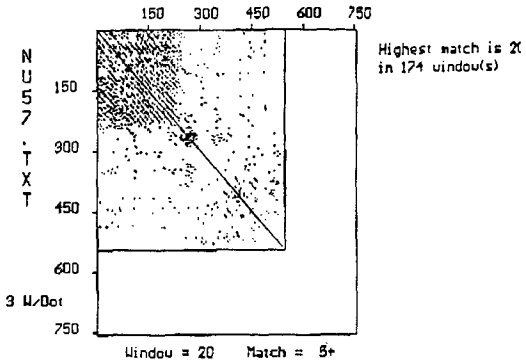
>GP-2245920 gi|1360451|gnl|PID|c245920 (Z73256) ORF YLR084c [Saccharomyces cerevisiae]; gi|1256886 (U53880) YlrO84cp [Saccharomyces



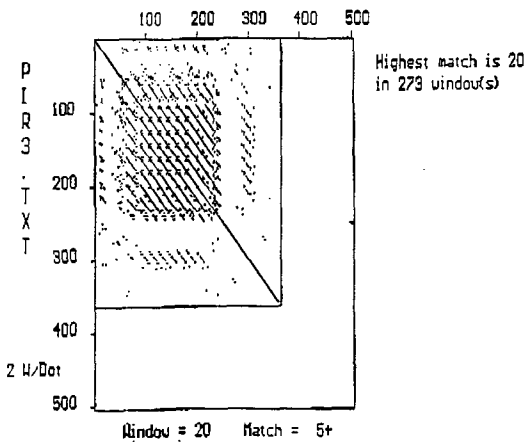
>GP-2251944 gi|1420104|gnl|PID|e251944 (Z74917) ORF YOR009w [Saccharomyces cerevisiae]; gi|1351005 (U43491) hypothetical protein



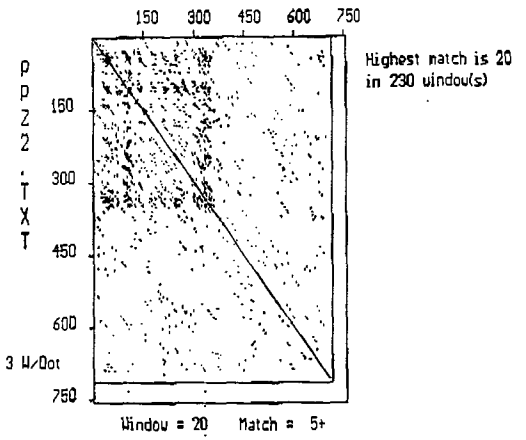
>SW-NSR1_YEAST SW:NSR1_YEAST P27476 saccharomyces cerevisiae (baker's yeast). nuclear localization sequence binding protein 1p67). 10/96;



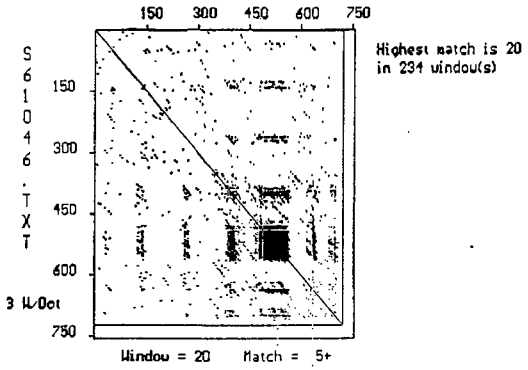
>SW-NUS7_YEAST SW:NUS7_YEAST P48837 saccharomyces cerevisiae (baker's yeast). nucleoporin nup57 (nuclear pore protein nup57). 10/96;



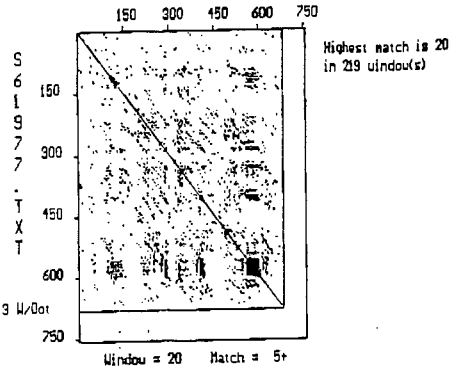
>SW-PIR3_YEAST SW:PIR3_YEAST Q03180 saccharomyces cerevisiae (baker's yeast). pir3 protein precursor. 11/95; gi|404129|gnl|PID|d1003392



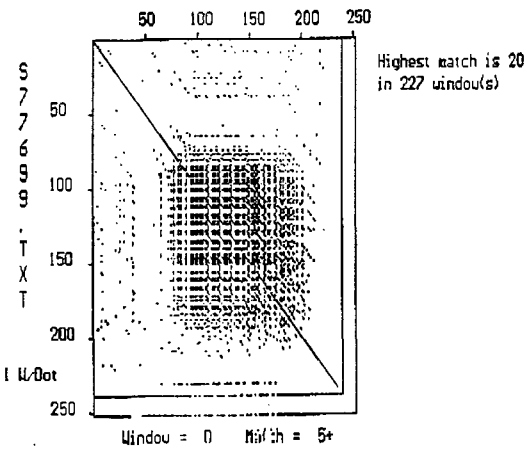
>SW-PP22_YEAST SW:PP22_YEAST P33329 saccharomyces cerevisiae (baker's yeast). serine/threonine protein phosphatase pp-z2 (ec 3.1.3.16). 10/96



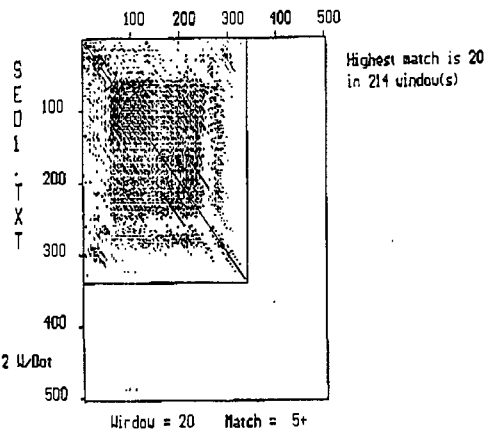
>PIR-S61977 SIR:S61977 RIM1 protein - yeast (Saccharomyces cerevisiae);
gi|1030721|gnl|PID|d1010303 (D63340) Rim1 (Saccharomyces cerevisiae);



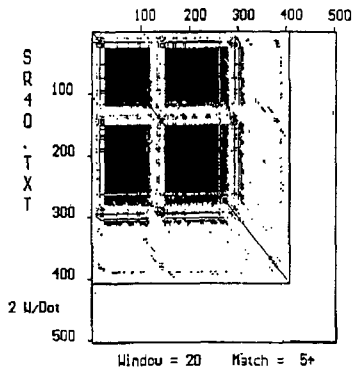
>PIR-S61046 PIR:S61046 ARF1 protein - yeast (Saccharomyces cerevisiae);
gi|1431266|gnl|PID|e253076 (S74215) ORF YDL167c (Saccharomyces



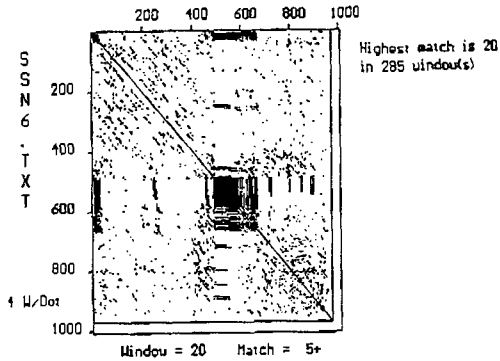
>PIR-S77699 PIR:S77699 inner cell wall mannoprotein ICWP - yeast
(Saccharomyces cerevisiae)



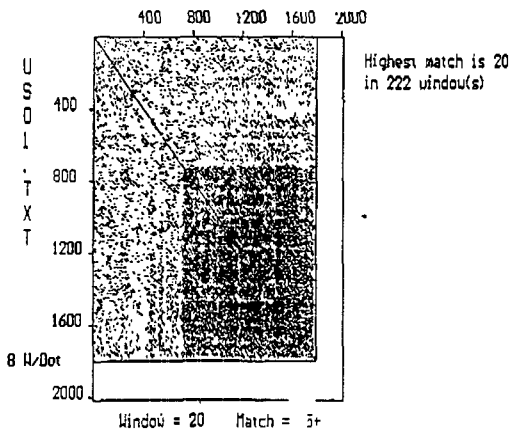
>SW-SED1_YEAST SW:SED1_YEAST Q01589 saccharomyces cerevisiae (baker's
yeast). sed1 protein precursor. 10/96; gi|431543|gnl|PID|e253329



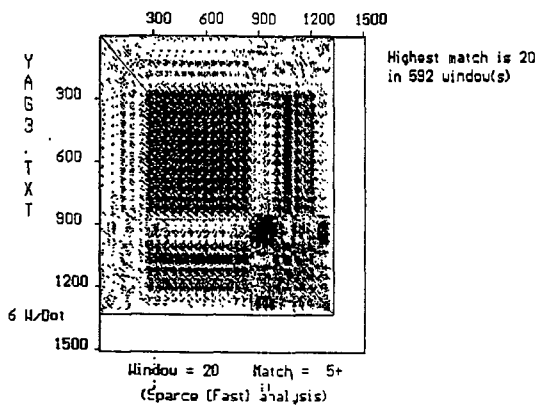
>SW-SR40_YEAST SW:SR40_YEAST P32583 saccharomyces cerevisiae (baker's yeast). Suppressor protein srp40. 6/94; PIR:S38170 SR40 protein - yeast (Saccharomyces cerevisiae); gi|486581 (EZ83171 ORF YKR092c



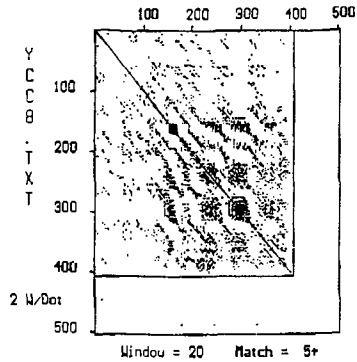
>SW-SSN6_YEAST SW:SSN6_YEAST P14922 saccharomyces cerevisiae (baker's yeast). Glucose repression mediator protein. 2/95; gi|172726 (M17826) SSN6 protein (Saccharomyces cerevisiae); gi|171350 (M23440)



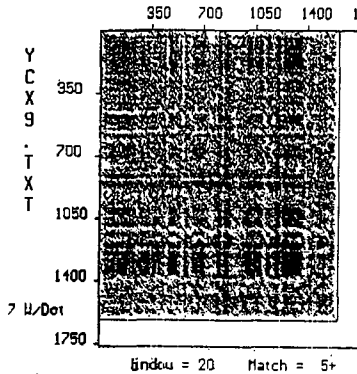
>SW-USO1_YEAST SW:USO1_YEAST P25386 saccharomyces cerevisiae (baker's yeast). Intracellular protein transport protein uso1. 10/96; gi|4778 (X54378) Uso1 protein (Saccharomyces cerevisiae)



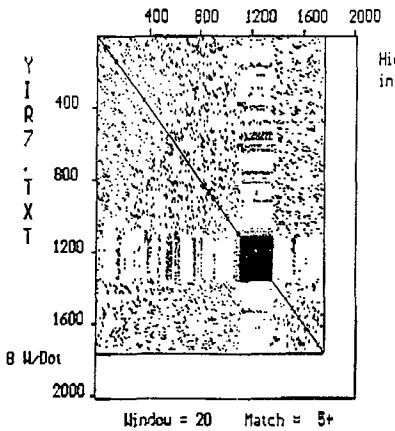
>SW-YAG3_YEAST SW:YAG3_YEAST P39712 saccharomyces cerevisiae (baker's yeast). hypothetical 138.1 kd protein in flo9-gdh3 intergenic precursor. 2/96; gi|758410 (U12980) FLO1 homolog (Saccharomyces cere



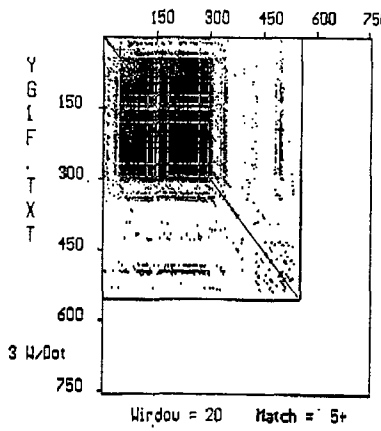
>SW-YCC8_YEAST SW:YCC8_YEAST P25367 saccharomyces cerevisiae (baker's yeast). hypothetical 42.6 kd protein in bik1-fus1 intergenic region.



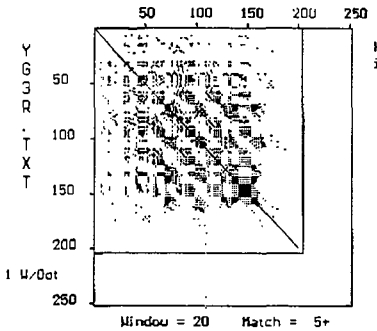
>SW-YCX9_YEAST SW:YCX9_YEAST P23653 saccharomyces cerevisiae (baker's yeast). hypothetical 166.0 kd protein in abp1 3'region. 10/96;



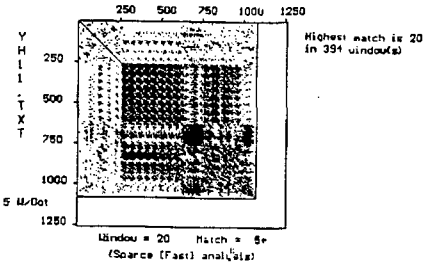
>SW-YIR7_YEAST SW:YIR7_YEAST P40434 saccharomyces cerevisiae (baker's yeast). hypothetical 197.5 kd protein in suc2 5'region. 11/95; gi|603898 (Z47047) unknown [Saccharomyces cerevisiae]; gi|600799 (Z



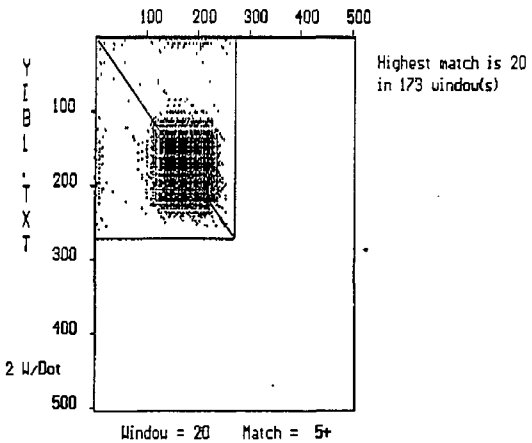
>SW-YG1F_YEAST SW:YG1F_YEAST P53214 saccharomyces cerevisiae (baker's yeast). hypothetical 57.5 kd protein in vma7-rps31a intergenic region.



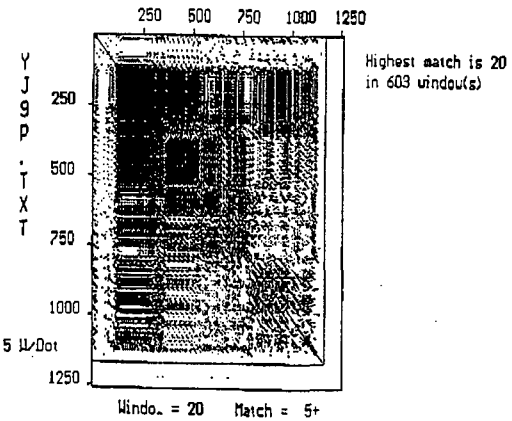
>SW-YG3R_YEAST SW:YG3R_YEAST P53288 saccharomyces cerevisiae (baker's yeast). hypothetical 22.2 kd protein in nsr1-cif4631 intergenic region.



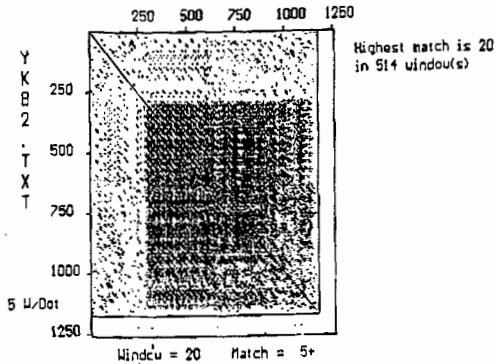
>SW-YH11_YEAST SW:YH11_YEAST P38894 saccharomyces cerevisiae (baker's yeast). hypothetical 112.0 kd protein in twt1-pho12 intergenic region precursor. 3/95; gi|456919 (D00029) YHr211wp [Saccharomyces



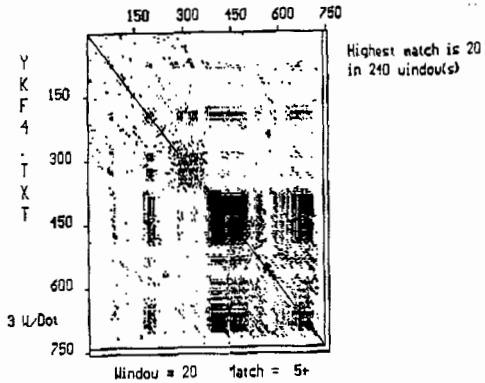
>SW-YTB1_YEAST SW:YTB1_YEAST P40552 saccharomyces cerevisiae (baker's yeast). hypothetical 26.3 kd protein in pdr11-fsa3 intergenic region.



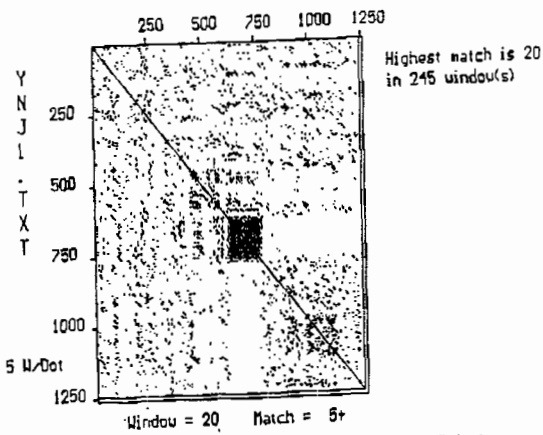
>SW-YJ9P_YEAST SW:YJ9P_YEAST P47179 saccharomyces cerevisiae (baker's yeast). hypothetical 118.4 kd protein in rps7b-dal5 intergenic region precursor. 10/96; gi|1015903 (Z49651) ORF YJR151c [Saccharom



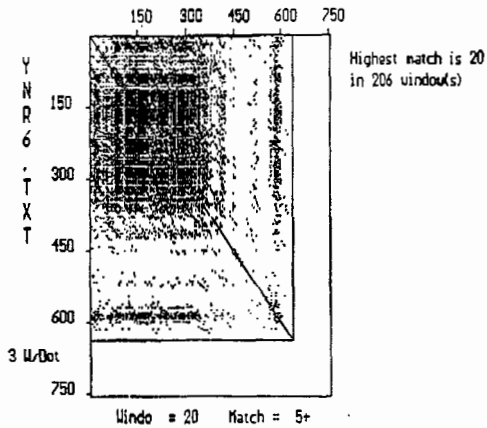
>SW-YKB2_YEAST SW:YKB2_YEAST P34178 saccharomyces cerevisiae (baker's yeast). hypothetical 133.3 kd protein in alr1 3'region precursor. 6/96; PIR:S38181 flocculation protein FLO1 homolog YKR102w - yea



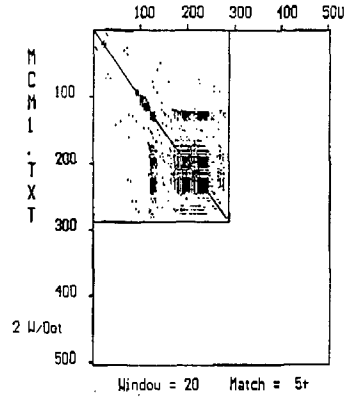
>SW-YKF4_YEAST SW:YKF4_YEAST P15732 saccharomyces cerevisiae (baker's yeast). hypothetical 84.0 kd protein in cup120-cse4 intergenic region. 2/96; PIR:S37876 glutamine-rich protein YKL053c - yeast



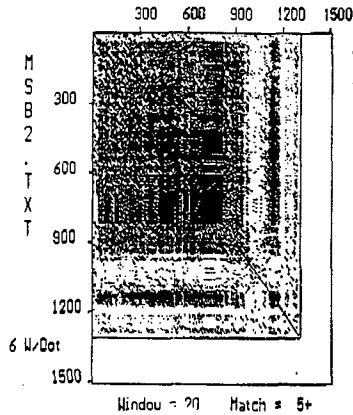
>SW-YNJ1_YEAST SW:YNJ1_YEAST P53935 saccharomyces cerevisiae (baker's yeast). Hypothetical 141.5 kd protein in ypt53-rho2 intergenic region.



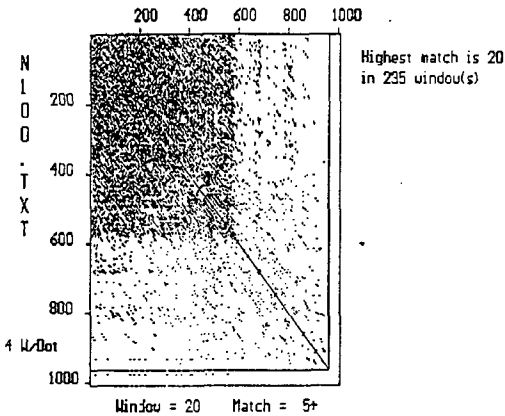
>SW-YNR6_YEAST SW:YNR6_YEAST P53882 saccharomyces cerevisiae (baker's yeast). Hypothetical 67.4 kd protein in rps3-psd1 intergenic region.



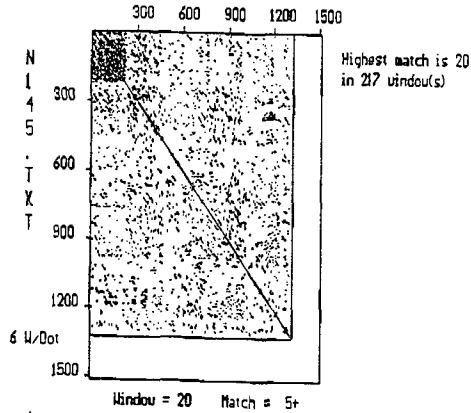
>SW:MCM1_YEAST SW:MCM1_YEAST P11746 saccharomyces cerevisiae (baker's yeast). pheromone response transcription factor (grm/prtf protein).



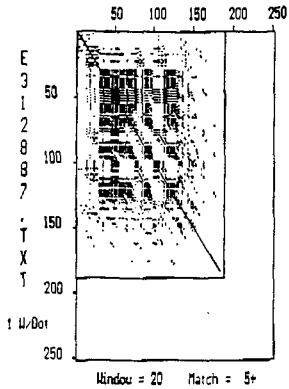
>SW:MSB2_YEAST SW:MSB2_YEAST P32334 saccharomyces cerevisiae (baker's yeast). msb2 protein (multicopy suppression of a budding defect 21.



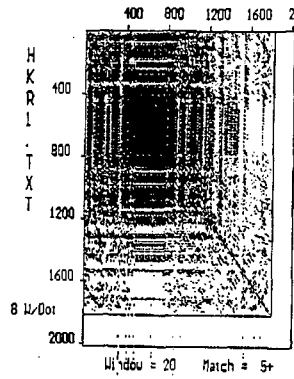
>SW:N100_YEAST SW:N100_YEAST Q02629 saccharomyces cerevisiae (baker's yeast). nucleoporin nup100/nsp100 (nuclear pore protein nup100/nsp100).



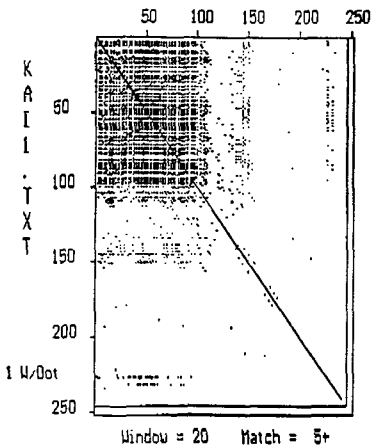
>SW:N145_YEAST SW:N145_YEAST P49687 saccharomyces cerevisiae (baker's yeast). nucleoporin nup145 (nuclear pore protein nup145). 10/95:



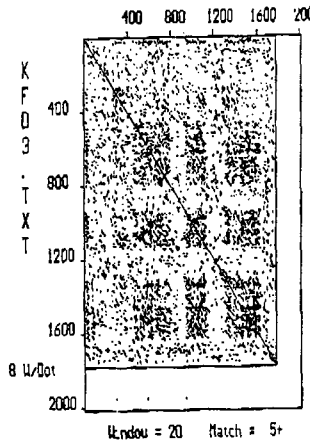
>G7-E312887 gi|1945352|gnl|PID|e312887 (D729469) ORF YGR159c [Saccharomyces cerevisiae]



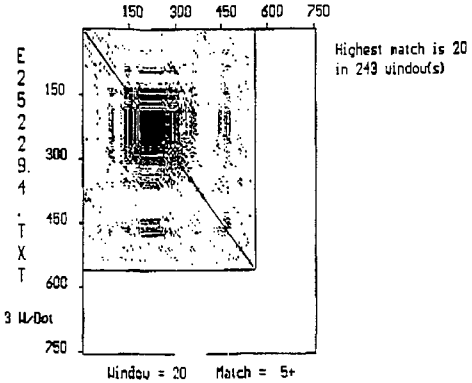
>SW-HKR1_YEAST SW:HKR1_YEAST P41809 saccharomyces cerevisiae (baker's yeast). kanamycin maxkl killer toxin-resistant protein 1 precursor. 10/96; gi|1545660|bbs|144411 (D69101) Hkr1p [Saccharomyces cer



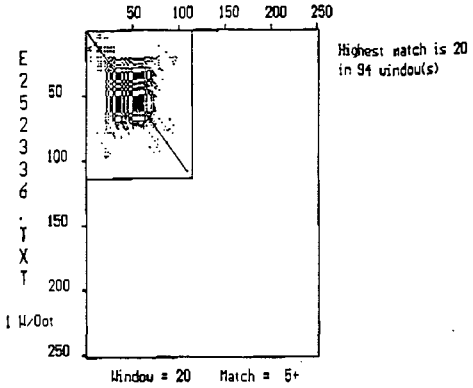
>SW-KAI1_YEAST SW:KAI1_YEAST P41944 saccharomyces cerevisiae (baker's yeast). protein kinase a interference protein (kai1 protein). 10/96



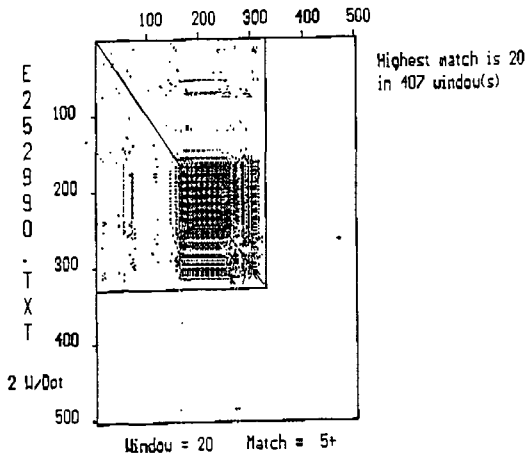
>SW-KFD3_YEAST SW:KFD3_YEAST P43565 saccharomyces cerevisiae (baker's yeast). probable serine/threonine-protein kinase yfl033c (ec 2.7.1.-).



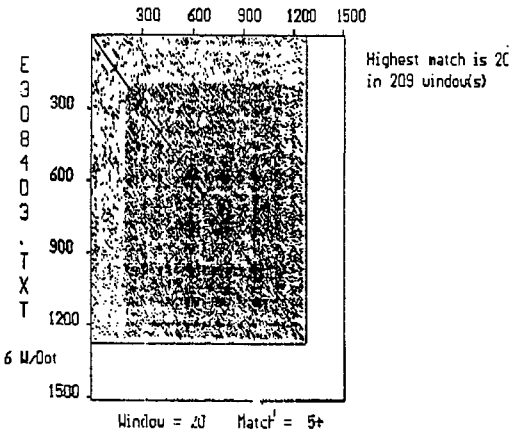
>GP-E252294 gi|1419967|gnl|PID|e252294 (Z74847) ORF YOL105c
[Saccharomyces cerevisiae]; gi|663247 (Z40149) similarity with H.
polymorpha hypothetical protein in LAU2 region [Saccharomyces cerevisiae]



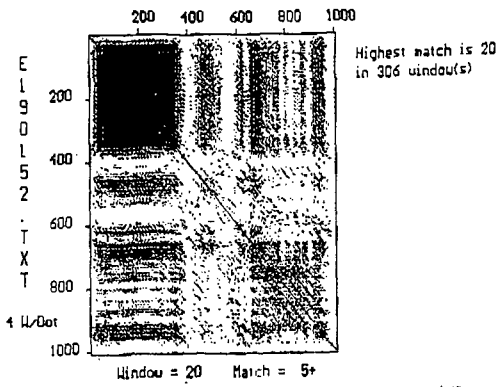
>GP-E252336 gi|1420189|gnl|PID|e252336 (Z74961) ORF YOR053w
[Saccharomyces cerevisiae]; gi|2104865|gnl|PID|e234099 (Z70678) YOR29-04



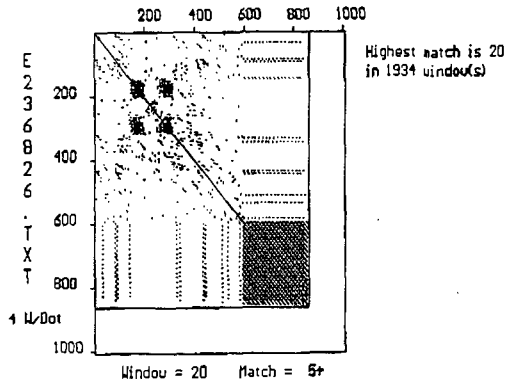
GP-E252990 gi|1431020|gnl|PID|e252990 (Z74085) ORF YDL037c [Saccharomyces
cerevisiae]; gi|1279679|gnl|PID|e237295 (Z71781) unknown [Saccharomyces



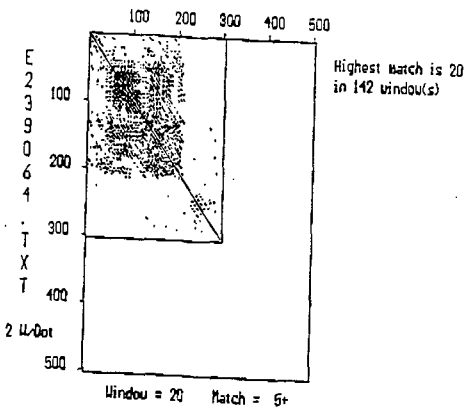
>GP-E308403 gi|1903280|gnl|PID|e308403 (Z74105) ORF YDL058w
[Saccharomyces cerevisiae]



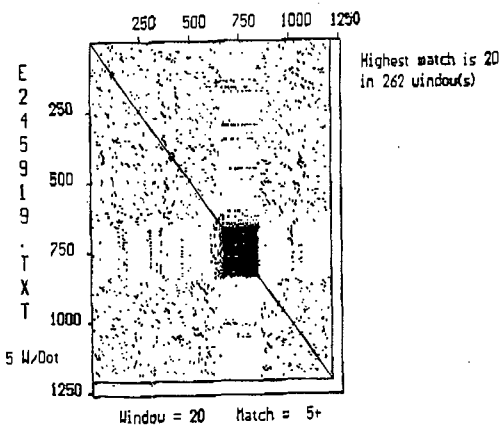
>GP-E190152 gi|1177622|gnl|PID|e190152 (X89715) AQR1001 gene product [Saccharomyces cerevisiae]



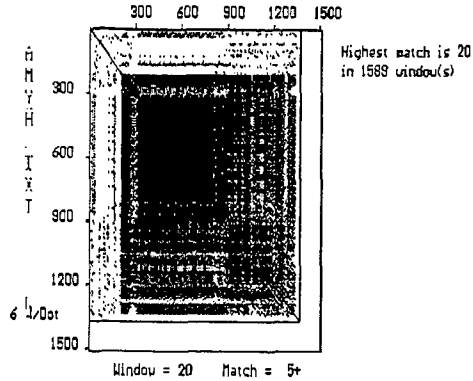
>GP-E236826 gi|1314096|gnl|PID|e236826 (Z71255) unknown [Saccharomyces cerevisiae]; gi|1809587 (U49274) unknown [Saccharomyces cerevisiae]



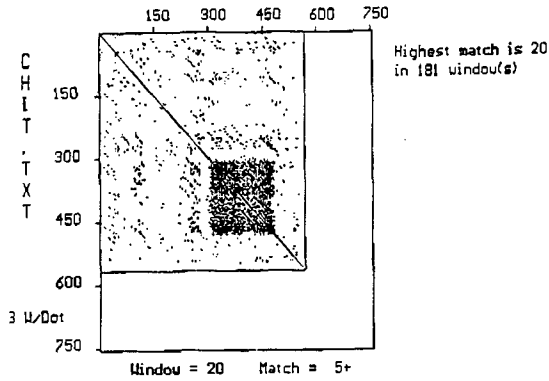
>GP-E239064 gi|1413944|gnl|PID|e1413944 (D64476) putative ORF [Saccharomyces cerevisiae]



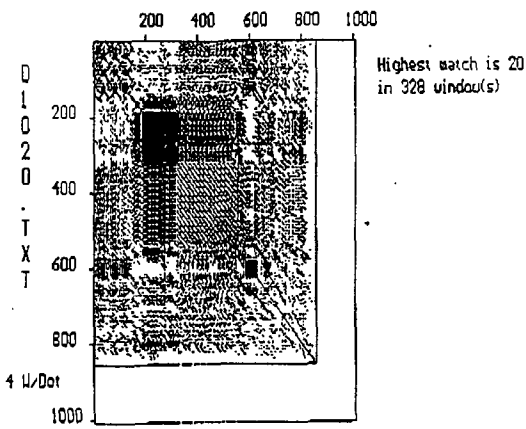
>GP-E245919 gi|1360284|gnl|PID|e245919 (Z73172) ORF YLL057c [Saccharomyces cerevisiae]



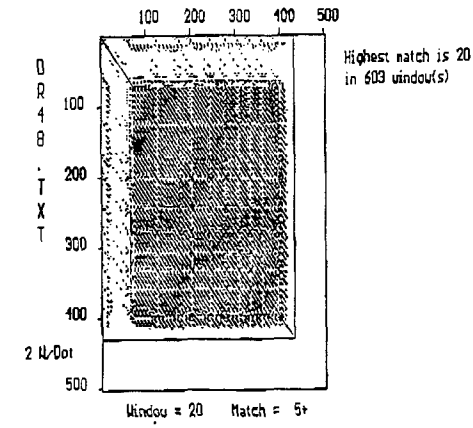
>SW:AMYH_YEAST SW:AMYH_YEAST P08640 saccharomyces cerevisiae (baker's yeast). glucoamylase s1/s2 precursor (ec 3.2.1.3) (glucan 1,4-alpha-glucosidase) (1,4-alpha-d-glucan glucohydrolase). 2/95: PIR:J



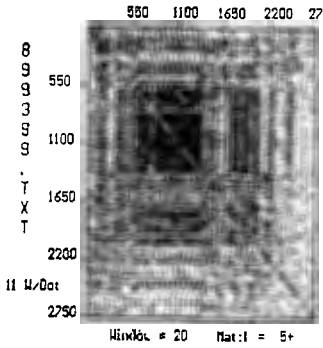
>SW:CHIT_YEAST SW:CHIT_YEAST P29029 saccharomyces cerevisiae (baker's yeast). endochitinase precursor (ec 3.2.1.14). 10/96; gi|596043 (U17243) Endochitinase 2 [Saccharomyces cerevisiae]



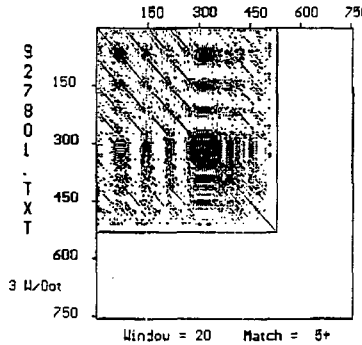
>GP:D1020704 gi|2077900|gnl|PID|d1020704 (AS003521) flocculin [Saccharomyces cerevisiae]



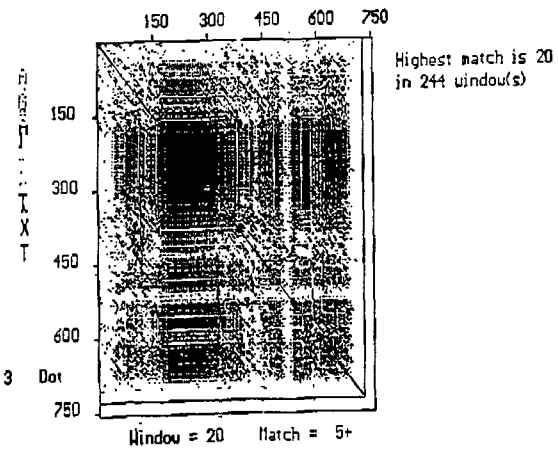
>SW:DR48_YEAST SW:DR48_YEAST P18899 saccharomyces cerevisiae (baker's yeast). dr48 stress protein (dna damage-responsive protein 48) (ddrp 48)



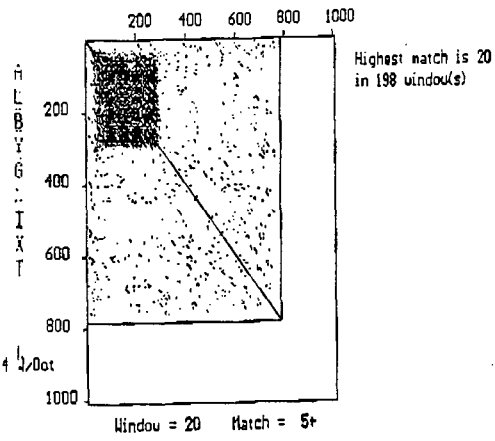
>GP-899399 gi|899399 (Z50016) Mumip [Saccharomyces cerevisiae]



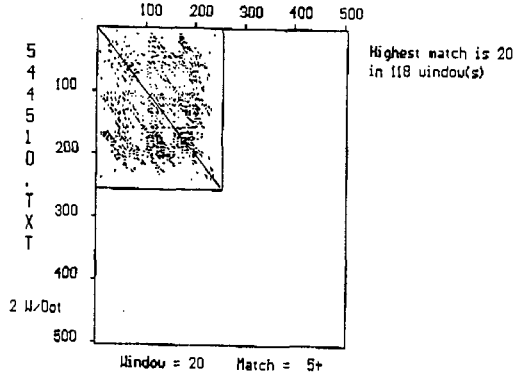
>GP-927801 gi|927801 (U33057) Ydr534cp; CAI: 0.26 [Saccharomyces cerevisiae]



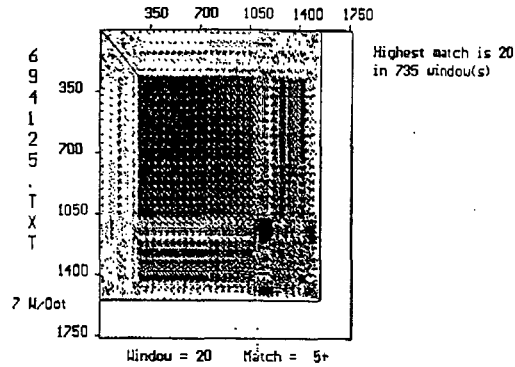
>SW-AGA1_YEAST SW:AGA1_YEAST P32323 saccharomyces cerevisiae (baker's yeast). a-agglutinin attachment subunit precursor. 10/96;



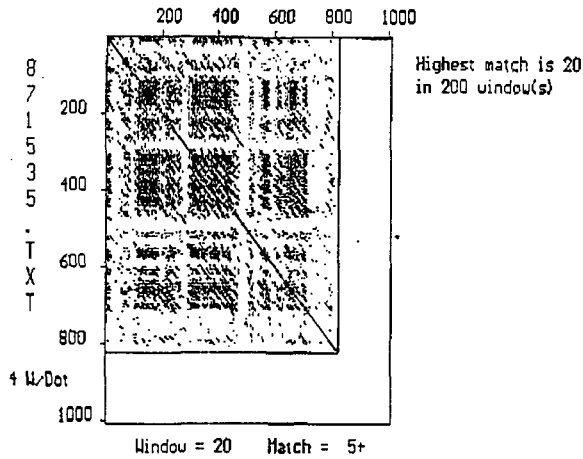
>PIR-ALBYG PIR:ALBYG glucan 1,4-alpha-glucosidase (EC 3.2.1.3) precursor



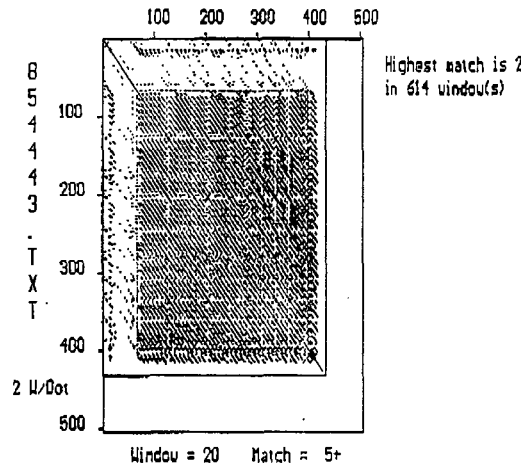
>GP-544510 gi|544510 (U14913) Yir194cp [Saccharomyces cerevisiae]



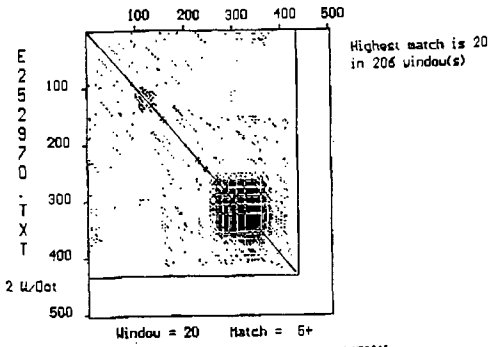
>GP-694125 gi|694125 (L28920) 1890bp repeat elements added, composed of 14 repeats of 135bp repeat element from Watari, J. et al., Yeast 10, 211-225 (1994) Molecular cloning and analysis of the yeast



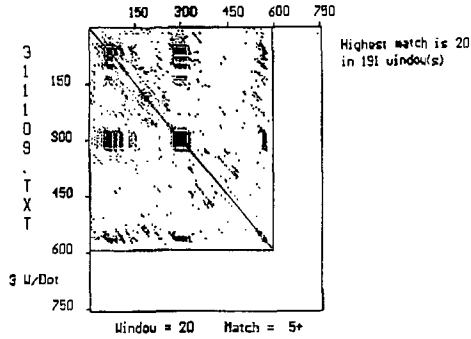
>GP-871535 gi|871535 (X87806) verprolin [Saccharomyces cerevisiae]



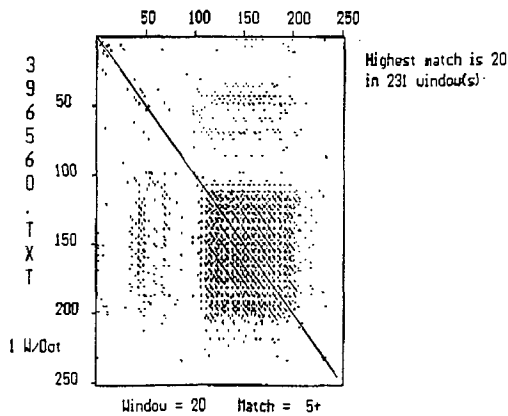
GP-854443 gi|854443 (Z49808) Ydr48p [Saccharomyces cerevisiae]; gi|6850151|bbs|155542 (S73336) FS2=flocculent specific protein



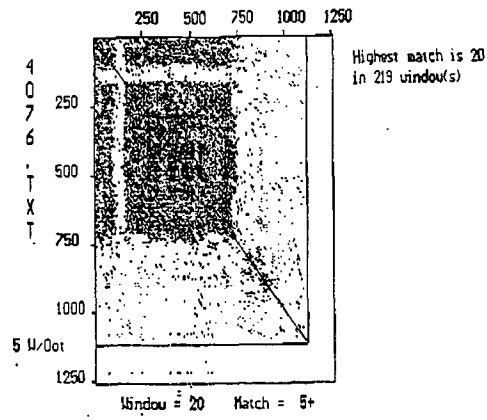
>GP-E252970 gi|14309831|pir||DID|a252970 (E74053) ORF YDL005c [Saccharomyces cerevisiae]; gi|683692 (L48432) unknown [Saccharomyces



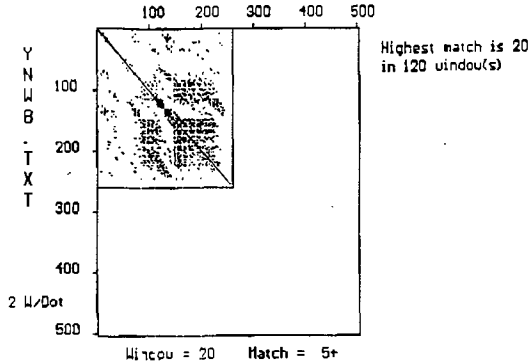
>GP-311109 gi|311109 (L16900) intrastrand crosslink recognition protein [Saccharomyces cerevisiae]



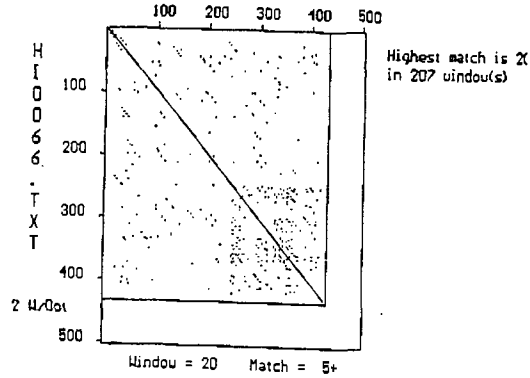
>GP-396560 gi|396560 (X74428) serine-rich protein [Saccharomyces cerevisiae]



>GP-4076 gi|4076 (S15036) nuclear pore complex protein NUP116 [Saccharomyces cerevisiae]

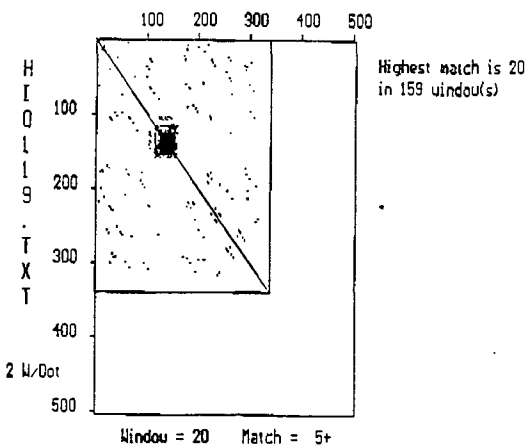


>SW:YNW8_YEAST SW:YNW8_YEAST P53862 saccharomyces cerevisiae (baker's yeast). hypothetical 28.6 kd protein in urn2-ssu72 intergenic region

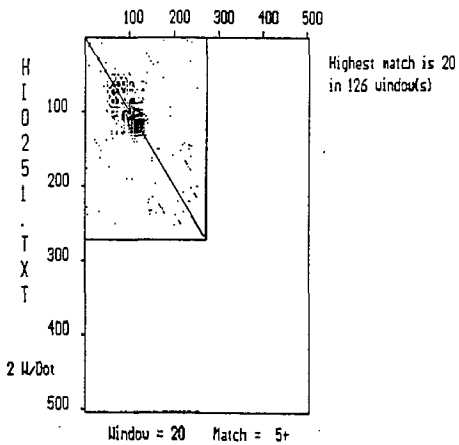


AMIB OM HI0066.

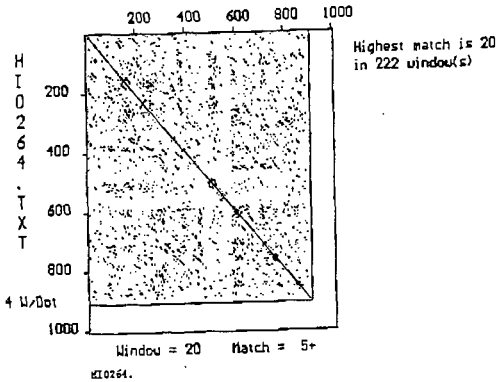
Haemophilus influenzae



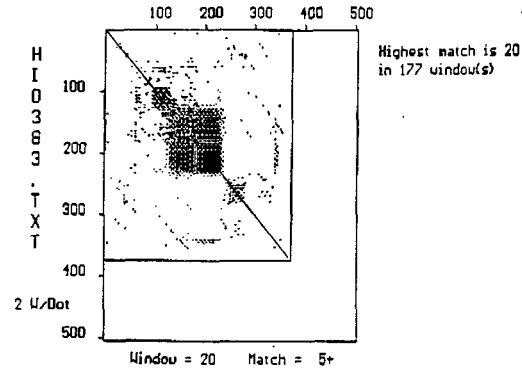
·ET0119,



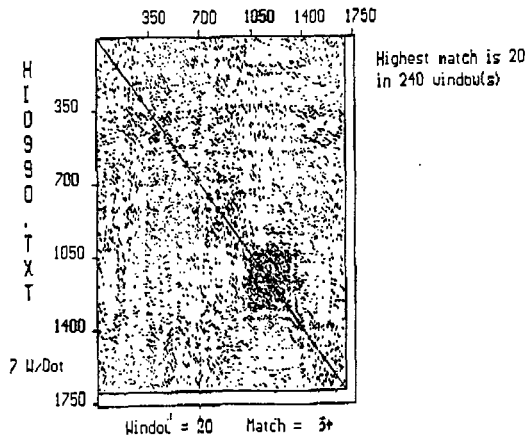
TOMS OR HI0251.



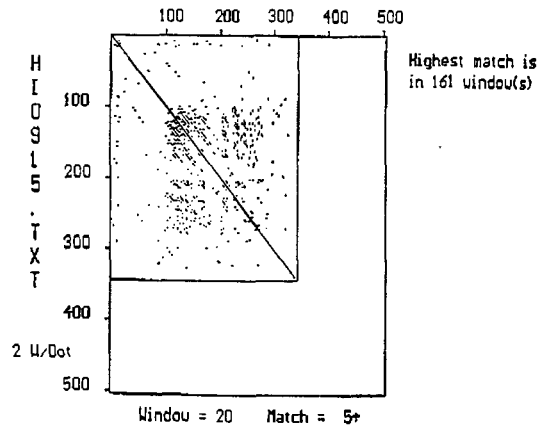
HI0264.



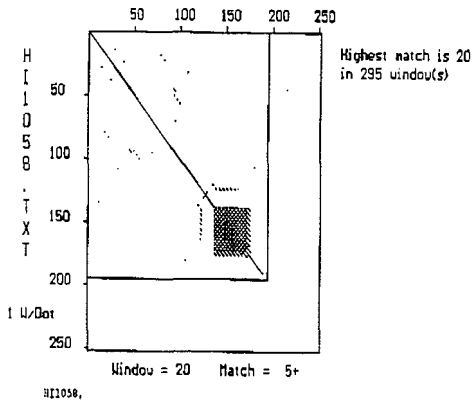
TOLA OR HI0383.



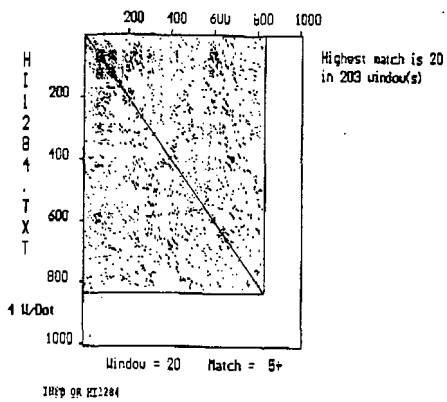
IGA OR IGAL OR HI0990.



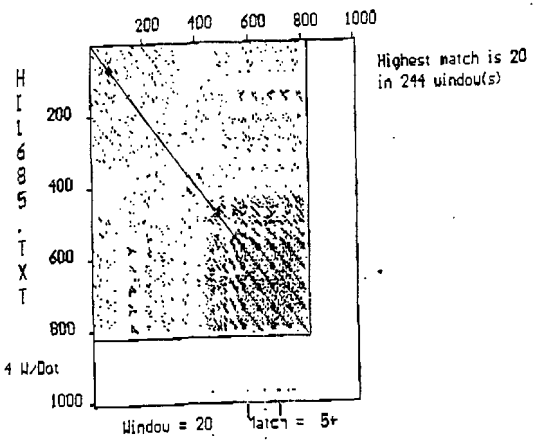
HI0915.



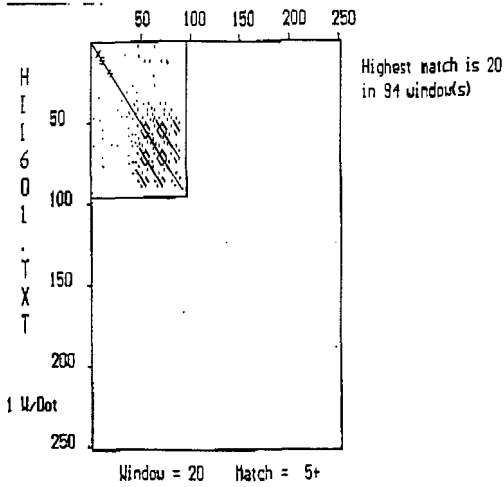
HI1058,



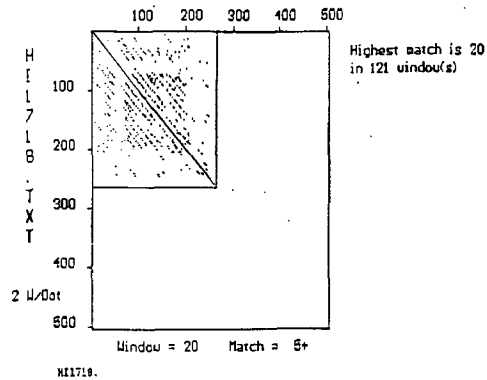
[illegible] OR HI1284

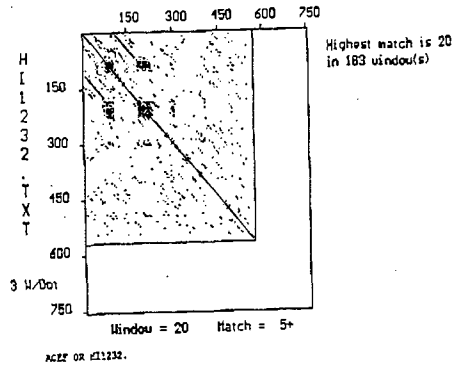


>HI1685 outer membrane integrity protein (tolA) (Escherichia coli)

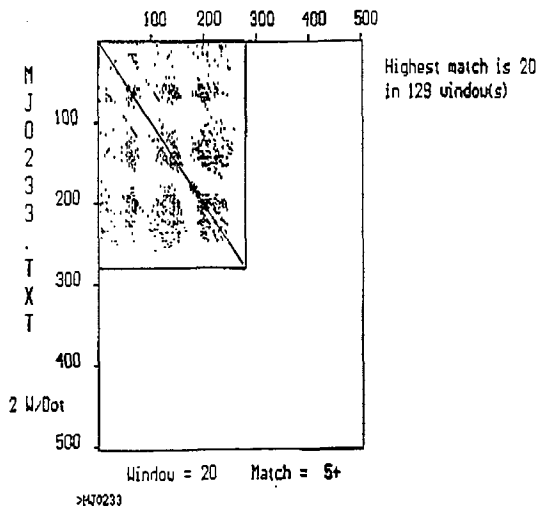


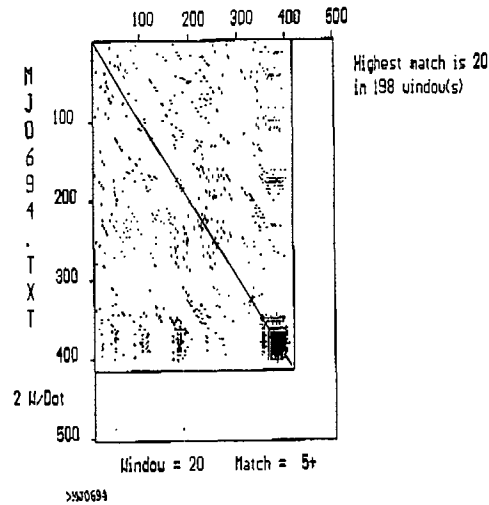
HI1601.

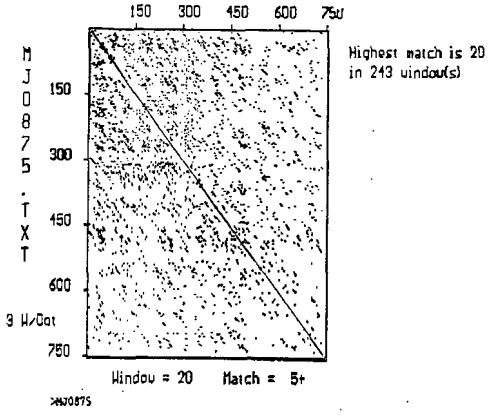


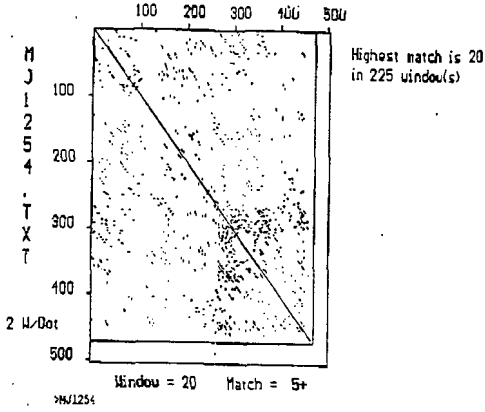


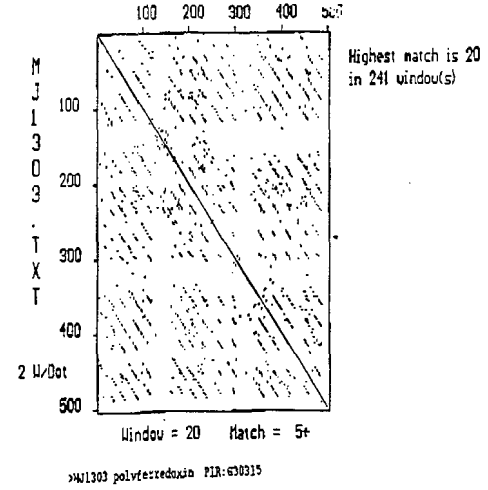
Methanococcus jannaschii

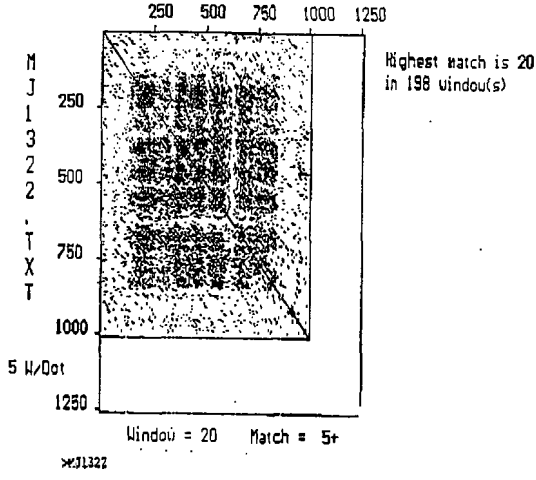


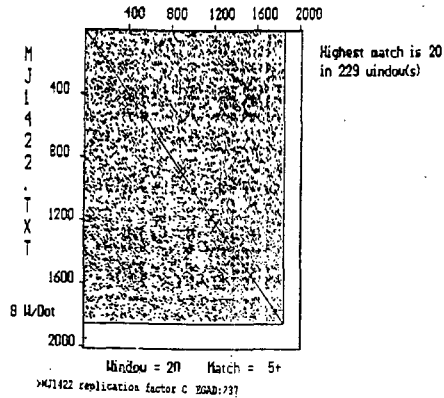


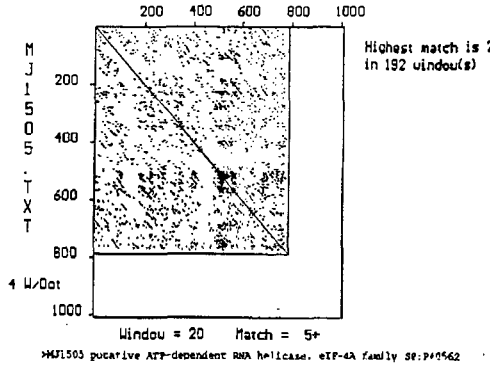

150 300 450 600 750
MJ0875.TXT
150
300
450
600
750
9 W/Dot
Highest match is 20
in 243 window(s)
Window = 20 Match = 5+

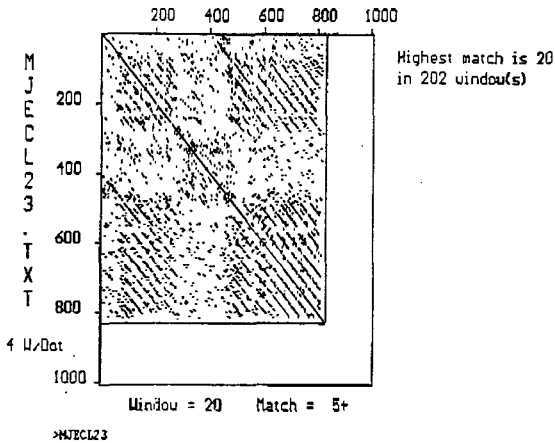

100 200 300 400 500
MJ1254.TXT
100
200
300
400
500
2 W/Dot
Highest match is 20
in 225 window(s)
Window = 20 Match = 5+

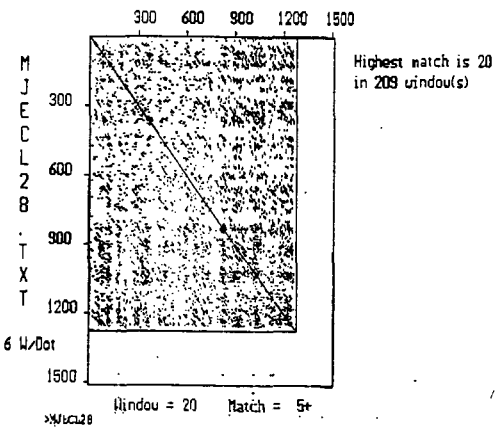

100 200 300 400
MJ1303.TXT
100
200
300
400
500
2 W/Dot
Highest match is 20
in 241 window(s)
Window = 20 Match = 5+

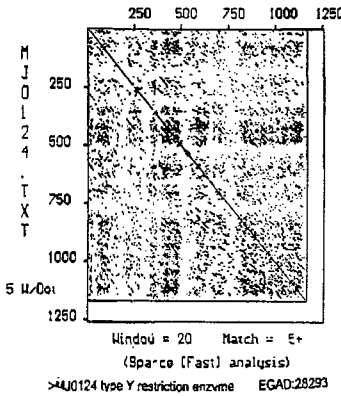

250 500 750 1000 1250
MJ1322.TXT
250
500
750
1000
1250
5 W/Dot
Highest match is 20
in 198 window(s)
Window = 20 Match = 5+

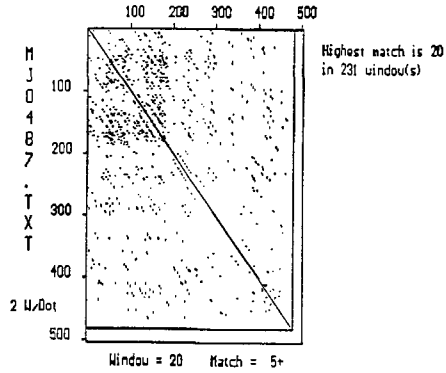

400 800 1200 1600 2000
MJ1422.TXT
400
800
1200
1600
2000
8 W/Dot
Highest match is 20
in 229 window(s)
Window = 20 Match = 5+
>MJ1422 replication factor C RGAD:73?


200 400 600 800 1000
MJ1505.TXT
200
400
600
800
1000
4 W/Dot
Highest match is 20
in 192 window(s)
Window = 20 Match = 5+
>MJ1505 putative ATP-dependent RNA helicase. eIF-4A family SP:P60562


200 400 600 800 1000
MJECL23.TXT
200
400
600
800
1000
4 W/Dot
Highest match is 20
in 202 window(s)
Window = 20 Match = 5+
>MJECL23


300 600 900 1200 1500
MJECL28.TXT
300
600
900
1200
1500
6 W/Dot
Highest match is 20
in 209 window(s)
Window = 20 Match = 5+
>MJECL28





8. GLOSARIO

1. Abreviaciones de los 20 aminoácidos.
   Alanina (**A**).
   Arginina (**R**).
   Asparagina (**N**).
   Ácido aspártico (**D**).
   Cisteína (**C**).
   Glutamina (**Q**).
   Ácido glutámico (**E**).
   Glicina (**G**).
   Histidina (**H**).
   Isoleucina (**I**).
   Leucina (**L**).
   Lisina (**K**).
   Metionina (**M**).
   Fenilalanina (**F**).
   Prolina (**P**).
   Serina (**S**).
   Treonina (**T**).
   Triptofano (**W**).
   Tirosina (**Y**).
   Valina (**V**).

2. Anonymous FTP: Permite tener acceso a información que se encuentra en internet para el dominio público basándose en una clave a la cual se le llama anonymous.

3. FTP: (File Transfer Protocol).Es un protocolo para la transferencia de archivos, que permite obtener o hacer copias de archivos hacia una computadora remota en internet.

4. Shell Unix: Es un programa en lenguaje Unix, este se encarga de leer los comandos que inserta el usuario y los interpreta como requerimientos para ejecutar ciertos programas que sean requeridos.

5. Tumover: Mecanismos que causan continuas fluctuaciones en el número de copias del DNA (variando en extensión desde dos a miles de nucleótidos), estos mecanismos pueden actuar simultáneamente en regiones del DNA.

6. Desigual crossing-over: Describe un evento de recombinación en el cual los dos sitios de recombinación, se sitúan en regiones no identificadas en las dos moléculas de DNA parental.

"La muerte es en la vida igual que el nacer;
como el andar está lo mismo en alzar el pie que en volverlo a la tierra".
Tagore.