Universidad Nacional Autónoma de México

Instituto de Fisiología Celular

Programa de Maestría y Doctorado
en Ciencias Biomédicas

Determinación de secuencias peptídicas antibacterianas

Tesis

que para obtener el grado de Doctor en Ciencias

presenta:

Carlos Polanco González

México, D.F. Octubre de 2009

Abstract

Antibacterial peptides represent a natural response to control infections in a wide variety of living forms and may have the potential to be used in the treatment of some human diseases. Our objective is to develop computer-based tools that may aid in the discovery of this class of peptides. Recognizing the astronomical number of possible peptide sequences, we developed computer programs to be executed in cluster of computers or in FPGA cards. One of our programs developed for high-performance computing named APAP II, classifies short Selective Antibacterial Peptides (SAPs) by their physical-chemical properties: Mean net charge, Mean hydrophobicity, Isoelectric point and Average helical hydrophobic moment. SAPs have no toxicity against human cells at least at the concentration these peptides are toxic against bacteria. Using this tool, one peptide out of 10^{11} peptides with 8 amino acids in length, was identified that presented mild antibacterial activity and no human toxicity.

Additionally, we implemented a Hidden Markov model (HMMs) based on 30 known SAPs. Using this method, a cluster of peptides with similar ranges of values characteristic of SAPs was detected from diverse peptide databases. These included 9 known SAPs, 6 synthetic antibacterial peptides formed by `Cecropin A` and `Magainin 2`, 19 peptides from the `Cecropin A` family, 4 peptides from the `Brevinin` family, 3 peptides from the `Cathelin` family and 2 peptides from the `Moricin` family. While all these peptides are known antibacterial peptides, most of these peptides were not included in our training set, supporting the usefulness of our implementation to identify potential SAPs.

Thus, innovative technology was developed to aid in the de novo identification of peptides based on high-performance computing systems and a machine learning method.

Prefacio

La toxicidad de los péptidos antibacterianos hallados en la naturaleza hacia los diversos patógenos, ha sido motivo de múltiples investigaciones en las últimas décadas, particularmente en la generación de nuevos fármacos.

Como proyecto doctoral se emprendió la tarea de responder, para un subgrupo de estos péptidos, tres preguntas:

> ¿La naturaleza ha considerado péptidos antibacterianos catiónicos y anfipáticos en el intervalo de 8 a 10 aminoácidos de longitud (la longitud promedio de los hallados en la naturaleza es de 23 aminoácidos), y que además se distingan por ser selectivos hacia membranas bacterianas?
>
> Por selectividad se entenderá que los péptidos presenten alta toxicidad hacia bacterias Gram-positivas y Gram-negativas y muy poca toxicidad extracelularmente, hacia células de eucariotes.
>
> ¿Sigue la naturaleza algún patrón observable en la estructura primaria de estos péptidos, de manera que sea posible distinguirlos dentro del espacio completo de todos los péptidos construibles?
>
> ¿Es viable su localización sin que ello involucre una costosa inversión en recursos computacionales y humanos?

El presente reporte describe:

> Un método matemático–computacional que aproxima la localización de estos péptidos en las longitudes referidas, evaluando uno a uno todos los éllos. Parte de la metodología inplicó predecir los péptidos antibacterianos selectivos hallados en la naturaleza referidos en bases de datos.
>
> Un método estocástico-computacional que aproxima estos péptidos buscándolos en el conjunto de péptidos antibacterianos hallados en la naturaleza.
>
> Un método computacional, que intentó predecir las regiones de actividad antibacteriana contenidas en los péptidos `magainina 2` y `cecropina A`.

Los resultados aquí expuestos muestran la importante participación de los métodos matemáticos y computacionales en el campo de la generación de nuevos fármacos, debido a que reducen la verificación experimental a sólo aquellos que se suponen los más probables.

Los métodos aquí expuestos permiten así efectuar exploraciones totales de espacios peptídicos para las longitudes referidas, a un bajo costo y aceptable efectividad.

Dedicatoria

Dedico esta tesis a mi madre que siempre nos espera con una sonrisa y cuya fuerza me inspira, a mi padre y a mis hermanos Miguel y José Manuel con quienes he aprendido lo que significa el trabajo constante y a Diana mi pareja de vida y apoyo incondicional.

Agradecimientos

A mi tutor Gabriel del Rio Guerra por su generosidad e inteligencia al permitirme participar en este proyecto doctoral el cual abrió camino al campo de la Bioinformática microcomputacional en México.

A mi comité tutorial, Lourival Domingos Possani Postay del Instituto de Biotecnología de la UNAM y Raúl Mancilla Jímenez del Instituto de Investigaciones Biomédicas de la UNAM por sus consejos y paciencia.

A Miguel Arias-Estrada del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) del CONACyT que creyó en este proyecto y facilitó todo tipo de recursos para lograr conocer y aplicar los dispositivos FPGA.

A José Lino Samaniego Mendoza del Departamento de Matemáticas de la Facultad de Ciencias de la UNAM, que cuidó cada aspecto matemático de este proyecto.

A mi jurado de Tesis, Ernesto Pérez Rueda, Gabriel del Río Guerra, Ignacio Méndez Ramírez, Lorenzo Segovia Forcella y Marco José Valenzuela por la revisión de este reporte.

A la Dirección General de Cómputo Académico (DGSCA) de la UNAM por facilitarme el uso de la supercomputadora Kam Balam para los procesos definitivos de esta investigación.

Al Instituto de Biotecnología de la UNAM (IBT) por permitirme el uso de su cluster lo cual contribuyó al desarrollo de las pruebas computacionales.

Al proyecto "Macroproyecto Tecnologías para la Universidad de la Información y la Computación", dirigido por el Dr. Humberto Carrillo del Departamento de Matemáticas de la Facultad de Ciencias de la UNAM, por su colaboración para la adquisición de los dispositivos electrónicos FPGAs y servidores que fueron utilizados en este proyecto doctoral.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) que económicamente me apoyó durante todo el proyecto de investigación.

Índice general

Índice de cuadros

Índice de figuras

Glosario

Notación	Descripción	
(a, b)	pareja ordenada, $(a, b) = \{a, \{a, b\}\} \in \mathcal{A} \times \mathcal{A}$	14
$[a, b]$	intervalo cerrado, $[a, b] = \{x \ : \ a \leq x \leq b \, a, b \in \mathcal{A}\}$	7
$\|\mathcal{V}\|$	norma de un vector $\mathcal{V}$	16
$\approx$	$a \approx b$, a es muy cercano a b	24
$\mathcal{V}_n^r$	variaciones con repetición de n elementos tomados de r en r. $\mathcal{V}_n^r = n^r$	2
σ	desviación estandar	24
$\sim$	$f \sim N(\mu, \sigma)$, f se distribuye como $N(\mu, \sigma)$	24
μ	esperanza matemática	24
e	$e \approx 2.7182$, número irracional	24
$F(x)$	función de distribución normal	24
$N(0, 1)$	distribución normal estandar	24
$f(x)$	función de probabilidad	24
$\forall$	$\forall \ a \in A$, todo elemento $a \in A$ cumple una condición	16
$\int_I f(t) \, dt$	integral de f en t sobre el intervalo $\mathcal{I} \subset \mathbf{R}$	24
χ^2_{gl}	distribución Ji-cuadrada con gl grados de libertad	25
$\mathcal{M}_{m \times n}(\mathcal{F})$	matriz de orden $m \times n$, sobre el campo $\mathbf{R}$	15
$N(\mu, \sigma)$	distribución Normal	24
$\mathbf{N}$	$\mathbf{N} = \{1, 2, 3, \ldots, n\}$, conjunto de números naturales	14
$\otimes$	operador binario, véase en el Índice modelos ocultos de Markov	17

Notación	Descripción	
$\in$	$a \in \mathcal{A}$, a pertenece al conjunto $\mathcal{A}$	7
$\mathcal{A} \times \mathcal{B}$	producto cartesiano, $\mathcal{A} \times \mathcal{B} = \{(x, y) : x \in \mathcal{A},\ y \in \mathcal{B}\}$	14
$\phi(z)$	función de distribución normal estandar	24
π	$\pi \approx 3.1416$, número irracional	24
$\mathbf{R}$	conjunto de números reales	14
$\mathcal{R}$	relación, subconjunto del producto cartesiano $\mathcal{R} \subset A \times B$	15
$\mathcal{S}_n$	sucesión, función $\mathcal{S}(n) : \mathbf{N} \to \mathcal{A}$, $\mathcal{A}$ conjunto arbitrario	2
$\subset$	$\mathcal{A} \subset \mathcal{B}$, $\mathcal{A} \subset \mathcal{B} \Longleftrightarrow \forall a \in \mathcal{A} \Rightarrow a \in \mathcal{B}$	14
$\sum_{i=1}^{n} S_n$	suma de los primeros n términos de la sucesión S_n	25
$\mathcal{V}_s$	vector de s términos sobre el campo $\mathbf{R}$	15
σ^2	varianza, cuadrado de la desviación estandar σ	25

Capítulo 1

Introducción

Los péptidos antimicrobianos constituyen la primera línea de defensa de los seres vivos ante diversos patógenos.

El espectro de acción de los péptidos es amplio [107] (hongos, bacterias, etcétera), al igual que su tamaño [21, 36, 49] (desde 6 aminoácidos (aa) hasta más de 100aa), siendo su longitud promedio 23aa (dato calculado a partir de los datos reportados en [2]).

Un ejemplo de péptido de amplio espectro es `Gambicina`: [Control de la malaria por medio de una metodología que permite un efectivo bloqueo del mosquito Anófeles MKQVCIILAVLLCTAAVADAMVFAYAPTCARCKSIGARYCGYGYLNRKGVSCDGQTTI NSCEDCKRKFGRCSDGFITECFL Anopheles arabiensis (Mosquito de la malaria originario de África del Sur) ACA05579.1—167861760 [80]], cuya longitud es de 85aa y se le ha reportado toxicidad tanto en hongos como en bacterias.

Si bien los péptidos antimicrobianos sólo se habian detectado en pocos organismos, incluyendo procariotes [79], hoy prácticamente se han localizado experimentalmente en todos los organismos eucariotes [30, 107] y sus aplicaciones son diversas, desde su uso como anticancerígenos [30, 34], hasta en el control de la obesidad [55].

Pensando en los péptidos antimicrobianos como una alternativa en la generación de nuevos fármacos [21, 24, 29, 30, 88]. Algunos esfuerzos se han orientado hacia su búsqueda en la naturaleza, de ello dan cuenta las bases de datos públicas [9, 15], donde pueden observarse centenas de estos péptidos.

Otras vertientes parten de los péptidos expuestos en las bases de datos referidas: se selecciona a alguno de ellos y se evalua su toxicidad repetidamente al sustituir o retirar algunos aminoácidos de su estructura primaria. Estos nuevos péptidos reciben el nombre de péptidos híbridos.

Algunos ejemplos son: `cecropina A(1--8)-magainina 2(1-12)` [Precursor de las magaininas que inhibe el crecimiento de numerosas especies de bacterias y hongos e induce lisis osmótica en protozoarios. Las magaininas son agentes de promueven la destrucción de la membrana. KWKLFKKIGIGKFLHSAKKF P11006.1—MAGA_XENLA [91]], construido por medio de la unión de los primeros 8 residuos de `cecropina`

A [Las cecropinas tienen una actividad antibacteriana y de lisis hacia varias bacterias Gram-positivas y Gram-negativas KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK Hyalophora cecropia P01507—CECA_HYACE [85]], y de los primeros 12 de magainina 2 [Las magaininas inhiben el crecimiento de numerosas especies de bacterias y hongos e inducen la lisis osmótica en los protozoarios. Las magaininas son agentes que promueven la destrucción de la membrana. GIGKFLHSAKKFGKAFVGEIMNS P11006.1—MAGAXENLA [38]], y (TPk: VRRFkWWWkFLRR) [108], que resulta de la mutación de los aminoácidos 5 y 9 del péptido catiónico triptripticin TP [Precursor de Prophenin-2 (PF-2) (PR-2) (C12) (Similar a Prophenin-1). Sus scrofa (cerdo) VRRFPWWWPFLRR P51525.1—PF12_ PIG [108]], por lisina (K).

También la búsqueda de péptidos antimicrobianos se ha orientado a partir de patrones fisicoquímicos, basados en la medición de estos parámetros para todos los péptidos contenidos en las bases de datos. Sin embargo, la producción de péptidos por esta vía no es práctica debido al costo de verificación experimental involucrado en la evaluación de todos los candidatos.

Baste suponer que si quisieramos ensayar todas las posibles mutaciones de cada aminoácido de un péptido antimicrobiano de longitud 45 aminoácidos el total de secuencias (o sucesiones) posibles sería las variaciones de $\mathcal{V}_{20}^{45}$ (equivalente a 20^{45}, véase la definición de $\mathcal{V}_n^r$ en Glosario).

Si sintetizar cada péptido emplea 1 USD entonces evaluar todos los péptidos involucrados en este supuesto requeriría 20^{45} USD.

No sólo el costo de síntesis bastaría en el análisis, también el tiempo que se emplearía para ensayarlos es una limitante, debido a que si 0.08s es el tiempo requerido para analizar un péptido entonces 20^{45}s es el tiempo necesario para analizar a todos ellos.

De lo anterior se concluye que el tiempo y el costo empleados para el ejemplo expuesto induce cifras no prácticas bajo métodos convencionales, y hace necesario considerar el uso de métodos computacionales que permita identificar a los más tóxicos entre todos los péptidos posibles.

Este trabajo comprende como objetivo el empleo de patrones fisicoquímicos, sobre la estructura primaria del péptido, incorporando herramientas computacionales y matemáticas, para la detección de péptidos con acción preferente a membranas bacterianas, catiónicos y anfipáticos, denominados péptidos antibacterianos selectivos, y con una longitud no mayor a diez aminoácidos.

Los métodos computacionales utilizados para la detección de estos péptidos hace uso de procesos probabilísticos como lo son los modelos ocultos de Markov (MOM) [11, 23, 34, 59, 62, 87]; modelos estadísticos [50] y de cómputo intensivo [75].

La metodología usada para responder el objetivo de este trabajo consiste en evaluar a través de métodos computacionales las variaciones de péptidos en longitud 8aa, 9aa y 10aa y detectar aquellos que se identifican con ciertos parámetros fisicoquímicos (Carga neta promedio, Hidrofobicidad promedio, Punto Isoelectrico, Momento

hidrofóbico y AGADIR), y parcialmente su selección se basa en la investigación de Del Rio G. [30] concerniente a la detección de péptidos proapoptóticos (péptidos que favorecen la apoptosis o muerte celular programada) los cuales presentan actividad antibacteriana selectiva.

Estos péptidos proapoptóticos se buscan computacionalmente entre los péptidos de longitud 8aa a 10aa, seleccionando aquellos que cumplen las características que se describen en los siguientes tres pasos:

Paso 1

- Ser catiónicos, es decir con presencia mayoritaria de aminoácidos básicos, lo cual contracta con el caracter aniónico de las membranas bacterianas.

Paso 2

- Adoptar una configuración anfipática, es decir, los aminoácidos de los péptidos se agrupan en regiones hidrofílicas e hidrofóbicas, independientemente de la estructura terciaria que adopten [31].

Paso 3

- Tener una actividad tóxica selectiva contra bacterias Gram-positivas y Gram-negativas y ser muy poco tóxicos hacia células de eucariotes.

A los péptidos antibacterianos con las características anteriores se les conoce como péptidos antibacterianos selectivos (PAS).

Cabe aclarar que el modo de acción de los PAS es aún desconocida; se sabe que afectan la membrana bacteriana [60, 101] causando la muerte de ésta pero el mecanísmo mediante el cual llevan a cabo la muerte de la célula no esta claro.

Los esfuerzos para caracterizar a los PAS en este trabajo se basaron en el procedimiento reportado por Del Rio G. [30] el cúal hace uso del índice terapéutico [30].

El índice terapéutico de un péptido, se define como la relación entre la concentración inhibitoria que se observa en células de mamíferos y la concentración inhibitoria que se encuentra en células bacterianas. Entre más alto sea el valor, más específico es el péptido en su acción contra membranas de organismos procariontes [30].

De acuerdo a ese procedimiento si el índice terapéutico de un PAS es mayor a 75 denominaremos a los PAS como péptidos antibacterianos selectivos anfipáticos catiónicos (PASAC).

El programa APAP fue diseñado suponiendo que las propiedades de los candidatos PASAC resultan equivalentes a la verificación de las siguientes propiedades fisicoquímicas:

- Tendencia a no formar una estructura $\alpha-$helicoidal en solución acuosa (AGADIR [57], por sus siglas en inglés).
- Punto isoeléctrico alto (PI).
- Momento hidrofóbico (MH) [32, 33].

Es importante señalar que las propiedades fisicoquímicas referidas en el trabajo de Del Rio G. [30], hacen notar que son indicativas del patrón de los péptidos antimicrobianos los cuales presentan: una baja tendencia a formar una estructura $\alpha-$helicoidal en solución acuosa, un alto Punto isoeléctrico y un alto Momento hidrofóbico en presencia de membranas bacterianas negativamente cargadas.

Particularmente, Del Rio G. [30] refiere que el Momento hidrofóbico es característico de los péptidos antibacterianos con baja toxicidad hacia celulas de mamíferos, y concluye que estas propiedades fisicoquímicas son características de los péptidos antimicrobianos que muestran una selectividad hacia membranas bacterianas.

Un primer reto del proyecto fue mejorar el tiempo de procesamiento del programa APAP ya que consume en el análisis de un péptido de longitud 9aa 0.08 segundos, ejecutándolo en una computadora personal, ello implica que la evaluación de 20^9 péptidos requiere $1,298$ años.

A partir de ello, evaluar el espectro total de PASAC y si éstos conforman cúmulos (agrupamientos de péptidos con características similares), en el espacio donde residen todos los péptidos, de manera que sea posible caracterizarlos.

La presente tesis esta organizada de manera que muestre los diferentes resultados obtenidos por medio de métodos matemáticos–computacionales, pretendiendo coadyuvar a la exploración efectiva de espacios peptídicos a un bajo costo y de aceptable efectividad.

Capítulo 2

Materiales y métodos

Las divisiones principales que integran esta Sección son: plataformas computacionales donde se da cuenta de las arquitecturas de cómputo usadas en el proyecto; programas computacionales, donde se describe el orden en que se plantearon todos los programas y archivos de péptidos y proteínas involucrados en las pruebas y, aplicaciones donde se describe cada plan de pruebas computacionales, aproximaciones matemáticas y/o ensayos biológicos efectuados.

2.1. Plataformas computacionales.

2.1.1. PC-Intel-686

Computadora Personal compuesta por un Procesador Intel Pentium 4 (2.80 GHz); RAM instalado: 512 MB DDR SDRAM; Disco duro: 80 GB estandar [4].

Las siguientes tres arquitecturas computacionales corresponden a las denominadas clusters y consisten de múltiples computadoras interconectadas que actúan como una sola computadora.

2.1.2. Cluster-20-Intel-686

Cluster compuesto por 20 procesadores Pentium IV 2.4Ghz. Un procesador por nodo. Memoria RAM: 512Mb por nodo. Disco duro: 80Gb Serial ATA. Switch: HP Procurve 10/100Mbps [7].

2.1.3. Cluster-14-Xeon

Cluster compuesto por 14 procesadores Xeon AMD Athlon(tm), 64×2 Procesador Dual Core 4200+ [5].

2.1.4. Cluster-20-AMD

Partición de 20 procesadores del sistema HP 4000. Procesadores: 1,368 (cores AMD Opteron de 2.6 GHz). Capacidad de procesamiento 7.113 Teraflops, memoria RAM: 3,000 Gbytes y un sistema de almacenamiento masivo de 160 Terabytes [8].

Esta siguiente plataforma provee el entorno para microprogramación.

2.1.5. FPGA-Xilinx

- Procesador–A–FPGA [69]. Modelo RC1000-PP V4.0 XCV2000E [1].
- Procesador–B–FPGA [69]. Modelo ADM-XRC 2VP7/2VP20 66MHz [1].
- Procesador–C–FPGA [69]. Modelo ADC-XRC-4FX Supports 3.3V PC o PCI-X a 64 bits.31.25MHz y 625MHz [1].

Las dos primeras plataformas conectadas a la empresa Xilinx [19] a través de INAOE [6], bajo responsabilidad de Arias-Estrada M. y la última Del Rio G.

Cabe señalar que un FPGA es una plataforma, que permite simular un cluster en su interior, y cuyo desempeño esta en relación directa al número de veces que un programa pueda ser copiado dentro de él.

2.2. Programas computacionales.

2.2.1. Rutinas computacionales del programa APAP

APAP es un conjunto de tres programas computacionales cuya selección corresponde a Del Rio G. [30], los cuales se ejecutaron en la plataforma PC-Intel-686, bajo el sistema operativo Linux [13]. Cada uno de esos programas representa una de las siguientes propiedades fisicoquímicas (para un mayor entendimiento de la selección de estas propiedades fisicoquímicas consúltese la Sección [1]), cuyos intervalos de aceptación son resultado de los ensayos experimentales de Del Rio G. [30]:

- Punto isoeléctrico alto PI. El punto isoeléctrico es el pH al que un aminoácido alcanza carga neta cero. El intervalo considerado para esta propiedad fue de 10.8 a 11.8. Este programa se encuentra codificado en lenguaje C [54].
- Momento hidrofóbico MH [32, 33]. El momento hidrofóbico es una medida de la tendencia del péptido a formar una $\alpha-$hélice. Estas $\alpha-$hélices son de tipo

anfipático lo cual lleva a que se distribuyan separadamente los residuos de aminoácidos con propiedades hidrofóbicas e hidrofílicas.

Esta distribución de los residuos de aminoácidos a lo largo de la $\alpha-$hélice está directamente relacionado con valores del momento hidrofóbico. El intervalo considerado para esta propiedad fue de 0.4 a 0.6. Este programa se codificó en lenguaje Fortran-77 [76].

- AGADIR [57]. Tendencia a no formar una estructura $\alpha-$helicoidal en solución acuosa o contenido de $\alpha-$hélices interacciones de corto rango. El intervalo considerado para esta propiedad fue de 0.0 a 10.0. Este programa se encuentra codificado en lenguaje C.

2.2.2. Rutinas computacionales del programa APAP–I.

El programa APAP–I es un programa elaborado en el lenguaje de programación Fortran-77, y se diseñó para ejecutarse en la plataforma PC-Intel-686 en el sistema operativo Linux. Contiene como subprogramas las siguientes cuatro propiedades fisicoquímicas:

- Punto isoeléctrico alto PI. Este programa es equivalente al programa correspondiente descrito para el programa APAP (véase la Sección [2.2.1]).
- Momento hidrofóbico MH [32, 33]. Este programa es equivalente al programa correspondiente descrito para el programa APAP (véase la Sección [2.2.1]).
- Hidrofobicidad promedio H [103]. Es el promedio de las hidrofobicidades de los aminoácidos normalizados a 1 sobre todos los aminoácidos del péptido. (Algoritmo proporcionado por el área técnica de ExPASy [9], debido a inconsistencia en cálculos, entre el descrito por Uversky V.N. [103] y la empresa ExPASy [9]). El intervalo considerado para esta propiedad fue de 0.35 a 0.55.
- Carga neta promedio C [103]. Ésta se encuentra determinada por la Ecuación [2.1]. (Algoritmo proporcionado por Uversky V.N.).

$$C(R,K,D,E) = \frac{1}{n}(R_i + K_i - D_i - E_i), i \in [1,n],\ n = \textit{ longitud del péptido.} \tag{2.1}$$

y cuyas variables R_i, K_i, D_i y E_i representan el número de veces que ocurren los aminoácidos: arginina (R), lisina (K), ácido aspártico (D) y ácido glutámico (E), aceptando aquellos péptidos cuya $C(R,K,D,E)$ de la Ecuación [2.1] sean mayores o iguales a $C_0(H)$ de la Ecuación [2.2], donde H es la Hidrofobicidad promedio.

$$C_0(H) = 45.896H^4 - 47.528H^3 + 13.324H^2 + 2.302H - 1.291 \qquad (2.2)$$

2.2.3. Rutinas computacionales del programa APAP–II.

El programa APAP–II esta constituido de los mismos cuatro subprogramas descritos para el programa APAP–I y escrito en el lenguaje de programación Fortran-77 y se modificó para que pudiera hacer uso de los clusters: Cluster-20-Intel-686, Cluster-14-Xeon y Cluster-20-AMD y ejecutarse en el sistema operativo Linux.

Para lograr ello se adicionaron las rutinas de paso de mensajes entre programas, correspondientes al software denominado (MPICH [75] Message-Passing Interface Standard Chamaleon, por sus siglas en inglés). El programa bajo MPICH se ejecutó en el cluster Cluster-20-AMD, mientras que a través de scripts de Linux se simuló el proceso paralelo referido en Cluster-20-Intel-686 y Cluster-14-Xeon.

De manera indistinta se simuló el proceso paralelo (MPICH y scripts de Linux), como se verá en la sección correspondiente a la exposición de resultados de esta sección, debido a que no influye en en estudio de rendimiento porque no se contempla como uso definitivo la plataforma Cluster-20-AMD.

Cabe hacer notar que existen otros programas anteriores a MPICH como (PVM [65] Parallel Virtual Machine, por sus siglas en inglés) que persiguen la misma finalidad.

Los programas que se diseñan para un cluster se conocen como programas paralelos o de cómputo intensivo.

2.2.4. Rutinas computacionales del programa APAP–III.

El programa APAP–III es un conjunto de comandos del sistema operativo Linux junto con el programa APAP y APAP–I (o APAP–II dependiendo de si se ejecuta en la plataforma PC-Intel-686 o en los clusters: Cluster-20-Intel-686, Cluster-14-Xeon y Cluster-20-AMD), ello dependerá del número de secuencias a ser analizadas.

APAP–III se diseñó para la detección de regiones antibacterianas probándose en:

- Los péptidos antibacterianos naturales no hemolíticos `magainina 2` y `cecropina A`.
- Un péptido antibacteriano hemolítico denominado `melitina` [Precursor de melitina. Apis cerana (Abeja india) Apis cerana (Abeja india)Main Toxina proveniente del veneno de abeja con una alta actividad hemolítica. Se integra a la membrana de la célula causando múltiples efectos, probablemente, como resultado de su interacción con fosfolípidos cargados negativamente .Inhibe el transporte de Na(+)-K(+)-AT-pase y the H(+)-K(+)-AT-pase. Incrementa la permeabilidad de la membranas

celulares a los iones, particularmente Na+ e indirectamente de Ca(2+), debido al intercambio Na(+)-Ca(2+) (por similitud). GIGAVLKVLTTGLPALISWIKRKRQ Q8LW54.1—MEL_APICE [58]].

- Un péptido de acción no específica llamado `gambicina`.

2.2.5. Rutinas computacionales del programa APAP–C.

Programa equivalente al programa APAP–I el cual fue codificado en el lenguaje de bajo nivel denominado Handel-C [69] en una interfaz Xilinx [19] para el sistema operativo Windows [14].

Handel-C fue seleccionado entre otros lenguajes de bajo nivel como: Verilog [48] y VHDL [77] por su semejanza al lenguaje de alto nivel denominado C y su control de memoria para codificación.

Todos estos lenguajes de bajo nivel referidos permiten programar un FPGA y ser ejecutados en una plataforma FPGA-Xilinx.

La generación del programa APAP–C comprende los siguientes dos pasos:

- Su codificación y pruebas en el simulador de Xilinx en Procesador–A–FPGA y Procesador–B–FPGA (véase la Sección [2.1.5].
- Una vez que el primer paso se resuelve se modifica el programa para sintetizarlo y probarlo físicamente en el Procesador–B–FPGA y Procesador–C–FPGA.

Es importante subrayar que un código codificado y probado en un simulador de FPGA no ofrece garantia alguna de arrojar los mismos resultados cuando éste mismo es sintetizado para correrse físicamente en el FPGA, usualmente es necesario cambiar rutinas de flujo de información o formatos de variables.

2.3. Aplicaciones

2.3.1. Variaciones en una secuencia peptídica.

Las variaciones de una secuencia peptídica son el conjunto generado por todas las secuencias que induce $\mathcal{V}_n^r$ (véase el Glosario), que son resultado de sustituir en cada una de las posiciones señaladas con la letra X en un péptido de longitud finita, cualquiera de los aminoácidos posibles.

Un péptido se expresa como una sucesión finita [ACDGIRSAYST] construida a partir de 20 aminoácidos diferentes. La lista total de aminoácidos es: A,C,D,E,F,G,H,I,K,L, M,N,P,Q,R,S,T,V,W,Y.

En este trabajo se utilizará una expresión abreviada para representar un conjunto finito de secuencias de determinada longitud (expresada ésta en x número de aminoácidos). A través de la expresión exponencial 20^n, donde n expresa el número de posiciones (al leer el péptido de derecha a izquierda), que serán sustituidas por cualquiera de los veinte posibles aminoácidos.

De manera que los 20^n péptidos de longitud x equivale a 20^n secuencias de longitud x cuyas últimas n posiciones variarían entre los veinte aminoácidos mencionados. En todos los casos la secuencia inicial tomará como aminoácidos iniciales el primero que aparece en la lista de aminoácidos aquí indicada, variando los mismos de derecha a izquierda.

Como ejemplo de ello, si se cita 20^4 péptidos de longitud 8aa, equivale a todos los péptidos posibles de longitud 8aa entre [AAAAAAAA] y [AAAAYYYY]. Esto es: [AAAAAAAA], [AAAAAAAC], [AAAAAAAD],. . ., [AAAAYYYY]. Note que aquí se resumen 20^4 secuencias peptídicas de longitud 8aa, seleccionadas bajo el criterio aquí descrito.

2.3.2. Análisis del programa APAP.

El programa APAP esta constituido por tres programas, y cada uno de ellos evalua una propiedad fisicoquímica del péptido y verifica que coincida con el intervalo de cada una de las propiedades fisicoquímicas (véase Sección [2.2.1]).

El primer reporte de desempeño del programa APAP se atribuye a Del Rio G. [30] quien previo a este proyecto mejora el rendimiento del programa en un 600 % reprogramando en el lenguaje VHDL, a través de arquitecturas computacionales específicas al algoritmo, donde se pueden implantar aceleración con base en la optimización de bajo nivel en los elementos de cómputo, y el paralelismo que se pueda definir en la arquitectura, sobre los programas PI y MH en un FPGA.

Las arquitecturas computacionales son un conjunto de elementos que hacen eficiente el balance memoria-velocidad, maximizando el desempeño de la mayoría de los programas. Su uso se remonta al lenguaje de bajo nivel denominado Assembler [3].

La verificación del programa APAP requirió del siguiente conjunto

- 20^5 péptidos de longitud 8aa y 9aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

Estas secuencias no se pueden considerar como una muestra representativa y sólo tuvo la finalidad de verificar el tiempo de procesamiento del programa APAP por cada péptido. La plataforma usada fue PC-Intel-686.

Cabe señalar que dentro de este invertalo se observó en procesos de prueba que existe un porcentaje significativo de candidatos PASAC, por lo que el tiempo de

procesamiento promedio por secuencia peptídica se puede considerar relevante para efecto de estimación de tiempo de procesamiento.

2.3.3. Significancia de AGADIR en péptidos cortos.

El programa APAP contiene un programa que verifica la propiedad AGADIR. Debido al considerable tiempo de procesamiento que emplea éste en la evaluación de cada péptido se verificó su efectividad discriminativa.

Esto equivale a verificar si la propiedad AGADIR esta en intervalo de aceptación para sólo una fracción de los péptidos cortos. Cabe hacer notar que en péptidos largos (mayores a 13aa), se reporta a AGADIR como un discriminante efectivo de PASAC [30].

AGADIR se verificó a partir de los siguientes dos conjuntos:

- Todos los péptidos construibles en longitud 9aa a partir de cada uno de los 442 genomas secuenciados y depositados en la base de datos NCBI [15].
- 442 genomas en su longitud original extraidos de la base de datos NCBI [15].
- 20^5 secuencias contenidas en los péptidos construibles en longitud 8aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

La plataforma usada fue PC-Intel-686.

2.3.4. AGADIR versus péptidos nativamente no estructurados.

La distinción entre las proteínas nativamente no estructuradas y las proteínas estructuradas lo establece Uversky V.N. [104, 103] a través de la Ecuación [2.3].

Es importante mencionar que las proteinas no estructuradas se agrupan en un espacio bien definido, dentro de todas las proteínas conocidas, bajo condiciones bien definidas de carga neta promedio-hidrofobicidad promedio [104], a través de una función lineal. La importancia de este hecho radica en la posibilidad de poder distinguir a las proteínas no estructuradas de las que si lo son, a partir de su estructura primaria.

$$C_1(H) = 2.785H - 1.150 \tag{2.3}$$

El procedimiento para calificar a una proteína como nativamente no estructurada consiste en determinar la carga neta C de acuerdo a la Ecuación [2.1], posteriormente sustitutir la hidrofobicidad promedio H en la Ecuación [2.3] y si C resulta ser mayor

o igual a $C_1(H)$ entonces esa proteína es no estructurada, de otra manera se dirá que es estructurada.

A partir de este hecho, se extrajeron de la base de datos ExPASy [9] los siguientes conjuntos con el fin de verificar si la relación a nivel de proteínas se podría llevar a péptidos cortos y de esta manera, poder calificar un péptido como nativamente no estructurado (PNNE) o estructurado (PNE).

- 91 proteínas naturales nativamente no estructuradas reportadas por Uversky V.N. [104], y extraidas de la base de datos ExPASy [9].

- 9 de los 31 PAS que fueron reportados por Del Rio G. [30] , los cuales se caracterizan por haber sido predichos por APAP y tener un índice terapéutico mayor a 75.

- 56 péptidos antimicrobianos extraidos de la base de datos ExPASy [9], detectados con estructura tridimensional por el método de resonancia magnética nuclear (NRM) [18]. La acción de estos péptidos se reporta en la base de datos sobre bacterias Gram-positivas y Gram-negativas, hongos, células de mamíferos y virus.

- 28 péptidos antimicrobianos extraidos de la base de datos ExPASy [9], detectados con estructura tridimensional por el método de resonancia magnética nuclear (NRM) [18]. La acción de estos péptidos se reporta en la base de datos sobre bacterias Gram-positivas y Gram-negativas únicamente, por lo que se les considera de acción selectiva.

  Además se verificó la acción de éstos consultando la literatura especializada, con objeto de verificar si existe algún modo de acción reportado, pero no incluido en la base de datos [9].

- 3 péptidos antimicrobianos extraidos de la base de datos ExPASy [9], detectados con estructura tridimensional por el método de difracción de rayos x (Rx) [18]. La acción de estos péptidos se reporta en la base de datos sobre bacterias Gram-positivas y Gram-negativas, hongos, células de mamíferos y virus.

- 1 péptido antimicrobiano extraido de la base de datos ExPASy [9], detectado con estructura tridimensional por el método de difracción de rayos x (Rx) [18]. La acción de este péptido se reporta en la base de datos sobre bacterias Gram-positivas y Gram-negativas únicamente, por lo que se le considera de acción selectiva.

  Además se verificó la acción de éste consultando la literatura especializada, con objeto de verificar si existe algún modo de acción reportado, pero no incluido en la base de datos [9].

- 30 % de 20^9 posibles péptidos de longitud 9aa seleccionados de manera aleatoria usando como generador de aminoácidos un algoritmo propio ya reportado [81].
- 2 péptidos [Precursor del escorpión Pandinus imperator GWINEEKIQKKIDERMGN TVLGGMAKAIVHKMAKNEFQCMANMDMLGNCEKHCQTSGEKGYCHGTKCKC GTPLSY P56972.1—SCRP_PANIM [27]] y [HADRURIN Hadrurus aztecus GILD-TIKSIASKVWNSKTVQDLKRKGINWVANKLGVSPQAA Actividad antimicrobiana hacia S.Typhimurium, L. Pheumoniae, E. Cloacae, P. Aeruginosa, E. Coli y S. Marcescens. P82656.1—HADR_HADAZ [99]], reportados por Possani L.D.

La plataforma usada fue PC-Intel-686.

Posteriormente se tomaron 20^5 péptidos aleatorios dentro de los posibles construibles en longitud 9aa y se evaluaron con AGADIR y con la curva correspondiente a la Ecuación [2.2] (véase la Sección [2.2.2] y la Sección [2.3.1].), comparándose los rechazados de ambos conjuntos.

2.3.5. Comparativo de los programas APAP y APAP–I.

La equivalencia entre los programas APAP y APAP–I se desarrolló sobre los siguientes conjuntos de péptidos y proteínas:

- Todos los péptidos construibles en longitud 9aa a partir de cada uno de 442 genomas secuenciados y depositados en la base de datos NCBI [15].
- 442 genomas en su longitud original, extraidos de la base de datos NCBI [15].

Ambos programas fueron ejecutados en la plataforma PC-Intel-686, comparándose todos los péptidos que cada programa produjo.

2.3.6. Comparativo de los programas APAP–I y APAP–II.

La equivalencia entre ambos programas requirió de la preparación de los siguientes conjuntos:

- Todos los péptidos construibles en longitud 9aa a partir de cada uno de 442 genomas secuenciados y depositados en la base de datos NCBI [15].
- 442 genomas en su longitud original, extraidos de la base de datos NCBI [15].
- 20^8 péptidos construibles de longitud 8aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

- 20^9 péptidos construibles de longitud 9aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

Es importante notar que es posible llevar a cabo el cálculo completo de los péptidos de longitud 8aa y 9aa, en la plataforma PC-Intel-686 debido a que hay una diferencia importante en el tiempo de procesamiento entre calificar un péptido y grabarlo en un disco y sólamente contarlo sin grabarlo [10].

La comparación se efectuó a nivel de los totales que ambos programas produjeron, al contar tanto los candidatos como los no candidatos PASAC, comparándose el número de péptidos que cada uno de los programas (APAP–I y APAP–II produjo).

El programa APAP–I se ejecutó en la plataforma PC-Intel-686 y el programa APAP–II en los clusters: Cluster-20-Intel-686, Cluster-14-Xeon y Cluster-20-AMD.

2.3.7. Frecuencia de aminoácidos contiguos en péptidos cortos.

Una vez que se tuvieron todos los candidatos PASAC en longitud 9aa se buscó un patrón ordenándolos en $\mathbf{N}^2 \subset \mathbf{R}^2$

El procedimiento consiste en graficar $(x, y) \in \mathbf{N}^2$, asignando a la variable $x \in \mathbf{N}$ el aminoácido o combinación de aminoácidos, y a la variable y la frecuencia relativa correspondiente.

De esta forma se construyó la distribución de los 20 aminoácidos (A,C, ..., Y), las 20^2 parejas de aminoácidos (AA, AC, ..., YY) y así sucesivamente, hasta las 20^5 parejas de aminoácidos (AAAAA, AAAAC,..., YYYYY).

Note que el plano cartesiano no es $\mathbf{R}^2$ debido a que sólo se seleccionaron los números enteros positivos $\mathbf{N}^2 \subset \mathbf{R}^2$.

2.3.8. Momento hidrofóbico versus Hidrofobicidad promedio.

La equivalencia entre ambas propiedades fisicoquímicas requirió de la preparación de los siguientes conjuntos de péptidos y proteínas:

- Todos los péptidos construibles en longitud 9aa a partir de cada uno de los 442 genomas secuenciados y depositados en la base de datos NCBI [15].
- 442 genomas en su longitud original, extraidos de la base de datos NCBI [15].
- 20^7 péptidos aleatorios en longitud 9aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

El programa APAP–I se ejecutó en la plataforma PC-Intel-686.

La comparación se efectuó contando y comparando cada secuencia aceptada o rechazada tanto para Momento hidrofóbico como para Hidrofobicidad promedio.

2.3.9. Carga neta promedio versus Punto isoeléctrico.

La equivalencia entre ambas propiedades fisicoquímicas requirió de la preparación de los siguientes conjuntos de péptidos:

- 20^7 péptidos aleatorios en longitud 9aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

El programa APAP–I se ejecutó en la plataforma PC-Intel-686.

La comparación se efectuó contando y comparando cada secuencia aceptada o rechazada tanto para Carga neta promedio como para Punto isoeléctrico.

2.3.10. Péptidos cortos y modelos observables de Markov.

Desde el inicio de este proyecto se trabajó en el diseño de una relación analítica que pudiera predecir un candidato PASAC sin necesidad de analizar todos los existentes.

La relación analítica propuesta adoptó una estructura matricial recursiva que corresponde al algoritmo conocido como modelos observables de Markov (véase la Ecuación [2.4]).

Es importante tener presente que los modelos markovianos o de Markov comprenden al menos tres estados: observables, semiocultos y ocultos. La diferencia estriba principalmente en el número de matrices estocásticas a ser empleadas. El autor consideró oportuno evaluar el modelo más simple (que no sencillo), para efecto de constatar si el primero de estos modelos podría predecir eficientemente el perfil buscado.

$$\mathcal{R}(\mathbf{X_n}) = \mathbf{A^k X_{n-1}} \tag{2.4}$$

Donde $\mathcal{R}(\mathbf{X})$ es un vector coordenado de orden n y $\mathbf{A}$ es una matriz cuadrada de orden $n \times n$, de tipo aleatorio. Con $k \in \mathbf{N},\ k > 0$. Véase la Ecuación [2.5] con objeto de una interpretación desglosada de la Ecuación [2.4].

$$\mathcal{R}(\mathbf{X_n}) = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,20} \\ a_{2,1} & a_{2,2} & \dots & a_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ a_{20,1} & a_{20,2} & \dots & a_{20,20} \end{pmatrix}^{\mathbf{k}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \end{pmatrix}_{\mathbf{n-1}} \tag{2.5}$$

Definiciones:

- $\mathcal{R}(\mathbf{X_n})$ es el vector coordenado de estado final o de resultados cuyos elementos $x_i \in \mathbf{R},\ i \in \mathbf{N}$, son las probabilidades finales de cada uno de los 20 aminoácidos.

- $\mathcal{R}(\mathbf{X_{n-1}})$ es el vector coordenado de estado inmediato anterior cuyos elementos $x_i \in \mathbf{R},\ i \in \mathbf{N}$, son las probabilidades de cada uno de los 20 aminoácidos.

- El vector coordenado $\mathbf{X_0}$ inicial es equiprobable. Esto quiere decir que el vector $\mathbf{X_n}$ tendrá cada elemento una probabilidad de 1/20.

- $\mathbf{A_{n\times n}}$ es una matriz que se distingue por ser aleatoria. Ello significa que sus elementos $a_{i,j} \in \mathbf{R}^2,\ i,j \in \mathbf{N}$; se interpretan como "la probabilidad de que un elemento del estado-j pase al estado-i". La característica de aleatoriedad de esta matriz debe cumplir: $a_{i,j} \in [0,1], \forall a_{i,j} \in A$ y $\sum_{i=1}^{n} a_{i,j} = 1$ para cada columna j.

  La matriz cuadrada $\mathbf{A}$ de orden 20 contiene la frecuencia relativa de las 20^2 variaciones posibles de los 20 aminoácidos tomados de dos en dos de todos los candidatos PASAC de longitud 9aa generados por el programa APAP–I.

  Una vez que se cuenta con la matriz $\mathbf{A}$ se multiplica k veces consigo misma, donde $k \in \mathbf{N}$ hasta encontrar una tolerancia $\beta > 0, \beta \in \mathbf{R}$, tal que la distancia $\|a_{i,j}^{n} - a_{i,j}^{n-1}\| < \beta$, $\forall a_{i,j} \in \mathbf{A}$ y ello lleva a la matriz $\mathbf{A^k}$, referida en la Ecuación [2.4].

2.3.11. Péptidos cortos y modelos ocultos de Markov.

El algoritmo modelos ocultos de Markov (MOM) [11, 23, 87], se usa en forma intensiva en Bioinformática particularmente en la caracterización de familias de proteínas e identificación de genes.

MOM es un método analítico aleatorio que forma parte del área del conocimiento Sistemas Dinamicos Discretos [45] cuyo dominio de acción es el campo $\mathbf{N}$ lo cual lo distingue de los Sistemas Dinámicos Continuos los cuales actuan sobre el campo $\mathbf{R}$.

Su objetivo es describir la dinámica de fenómenos que no actuan en un espacio de tiempo continuo, sino como un conjunto de fotografías que permiten predecir la tendencia sin observar la transformación precisa del objeto en estudio, y se caracteriza porque no requiere del pasado total del objeto a evaluarse, sino su historia inmediata anterior maximizando las cualidades del perfil buscado.

El perfil que MOM buscará se explica en forma matricial, siendo uno de sus componentes dos matrices aleatorias $\mathbf{A}$ y $\mathbf{B}$ y un vector coordenado $\mathbf{X}$.

Una descripción analítica de MOM adopta la Ecuación [2.6].

$$\mathcal{R}(\mathbf{X_n}) = \mathbf{A} \otimes \mathbf{B}\mathbf{X_{n-1}} \tag{2.6}$$

Véase la Ecuación [2.7] para una interpretación desglosada de la Ecuación [2.6].

$$\mathcal{R}(\mathbf{X_n}) = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,20} \\ a_{2,1} & a_{2,2} & \dots & a_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ a_{20,1} & a_{20,2} & \dots & a_{20,20} \end{pmatrix} \otimes \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{20} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \end{pmatrix}_{\mathbf{n-1}} \tag{2.7}$$

Definiciones:

- $\mathcal{R}(\mathbf{X_n})$ es el vector coordenado de estado final o de resultados cuyos elementos $x_i \in \mathbf{R},\ i \in \mathbf{N}$, son las probabilidades finales de cada uno de los 20 aminoácidos.

- $\mathcal{R}(\mathbf{X_{n-1}})$ es el vector coordenado de estado inmediato anterior cuyos elementos $x_i \in \mathbf{R}, i \in \mathbf{N}$, son las probabilidades de cada uno de los 20 aminoácidos.

- El vector coordenado $\mathbf{X_0}$ inicial es equiprobable. Esto quiere decir que el vector $\mathbf{X_n}$ tendrá cada elemento una probabilidad de 1/20.

- $\mathbf{A_{n\times n}}$ es una matriz que se distingue por ser aleatoria. Ello significa que sus elementos $a_{i,j} \in \mathbf{R}^2,\ i,j \in \mathbf{N}$; se interpretan como "la probabilidad de que un elemento del estado-j pase al estado-i". La característica de aleatoriedad de esta matriz debe cumplir: $a_{i,j} \in [0,1], \forall a_{i,j} \in A$ y $\sum_{i=1}^{n} a_{i,j} = 1$ para cada columna j.

  La matriz cuadrada $\mathbf{A}$ de orden 20 contiene la frecuencia relativa de las 20^2 variaciones posibles de los 20 aminoácidos tomados de dos en dos de todos los candidatos PASAC de longitud 8aa generados por el programa APAP–II.

- $\mathbf{B}$ es una matriz rectangular de orden 20×2 que aporta un segundo perfil lo cual minimiza los posibles candidatos que son parte de la solución. Para este caso se optó por asignar la frecuencia relativa de aparición de los aminoácidos presentes en los péptidos reportados por Del Rio G. [30] en su primera columna y su complemento, considerando como el espacio probabilístico el cerrado $[0,1]$, en su segunda columna.

- $\otimes$ Es un operador que adiciona el elemento $b_i \in B_n$ al elemento $a_{i,j} \in \mathbf{A}$ pero guardando registro sólo de aquellos elementos $a_{i,j}$ que recorren un camino óptimo (véase preferéntemente [23] para una explicación detallada de este operador).

La matriz $\mathbf{A}$ se construyó a partir de los siguientes conjuntos:

- Los 20^9 candidatos PASAC que el programa APAP–II produjo.
- Todos los péptidos antibacterianos depositados en la base de datos BBCM [2] en su longitud original.

La matriz **B** se construyó a partir de los siguientes conjuntos:

- Los candidatos PASAC con índice terapéutico mayor a 75 reportados por Del Rio G. [30].

En opinión del autor, a continuación se listan unas desventajas de MOM las cuales son representativas de un Sistema Dinámico, al cual corresponde un MOM.

Desventajas del modelo oculto de Markov:

- Si el perfil suministrado es escaso, sus calificaciones serán imprecisas. Esto significa que un número pequeño de elementos para entrenar el método MOM o datos que no reflejen el perfil buscado, conduciran a resultados erroneos, no habiendo manera de que MOM lo advierta.
- MOM asume que la caracterización es independiente de los hechos remotos y ello no se puede asegurar. Esta advertencia es muy importante debido a que el uso de MOM es la base de muchos programas en Bioinformática [93, 102, 106].
- De manera particular, no se puede precisar el valor óptimo. MOM arroja un conjunto de valores los cuales pertenecen al intervalo cerrado [0, 1], queda en manos del usuario ordenarlos y aceptar como valor máximo el equivalente a la probabilidad máxima 1, y como valor mínimo al equivalente a la probabilidad mínima 0.

2.3.12. Péptidos antibacterianos y Modelos ocultos de Markov

Para una consulta detallada de esta metodología consúltese el Apéndice F. El algoritmo modelos ocultos de Markov (MOM) [11, 23, 34, 87] descrito en la Sección [2.3.11] se usó también en esta sección para predecir los PASAC ya conocidos entre los péptidos antibacterianos registrados en la base de datos [105] a septiembre, 2007.

Para ello se requirió de los siguientes conjuntos:

> **Conjunto $\mathcal{A}$**: 59 péptidos antibacterianos e híbridos extraidos del (**conjunto $\mathcal{C}$**), los cuales actuan exclusivamente contra bacterias, hongos, virus y células cancerígenas de mamíferos, cuya estructura tridimensional es conocida, reportada en la base de datos APD [105] a septiembre, 2007.

Cuadro 2.1

Elementos del vector $\mathsf{X_0}$. Distribución de frecuencia absoluta de los aminoácidos presentes en los péptidos antibacterianos, cuya acción patógena es exclusivamente contra bacterias, hongos, virus y células cancerígenas de mamíferos (**conjunto** $\mathcal{A}$) del vector $\mathsf{X_0}$. Las letras en la tabla se refieren a los 20 aminoácidos (una letra por aminoácido), y los números representan la correspondiente frecuencia del aminoácido del conjunto.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
103	132	23	32	61	182	39	129	146	101	9	52	67	49	135	87	53	85	31	57

Conjunto $\mathcal{B}$: 28 péptidos antibacterianos e híbridos extraidos del (**conjunto** $\mathcal{C}$), los cuales actuan exclusivamente contra bacterias y cuya estructura tridimensional es conocida, reportada en la base de datos APD [105] a septiembre, 2007.

Conjunto $\mathcal{C}$: 500 péptidos antibacterianos que no presentan acción específica contra algún patógeno y cuya estructura tridimensional es conocida, reportada en la base de datos APD [105] a septiembre, 2007.

Conjunto $\mathcal{D}$: 3 péptidos antibacterianos e híbridos extraidos del (**conjunto** $\mathcal{C}$): `gambicina`, `melitina` y `temporin H (XXA,frog)` [92], reportados en la base de datos APD [105] a septiembre, 2007.

Conjunto $\mathcal{E}$: 392,836 proteinas naturales reportados en la base de datos UNIPROT [17] a agosto, 2008.

Implementación

El vector coordenado $\mathsf{X_0}$ inicial fue construido a partir del promedio $\frac{1}{n}\sum_{i=1}^{n}\mathsf{x}_{0i}$, donde n es la longitud del péptido a ser examinado y x_{0i} es la distribución de frecuencia relativa de los aminoácidos del mismo péptido. Para ello se usó el (**conjunto** $\mathcal{A}$) (véase el Cuadro [2.1]).

La matriz A representa la distribución de frecuencia relativa de los 400 pares de posibles aminoácidos formados por contiguidad. Estos pares fueron tomados en dos direcciones dentro de la matriz: $(a_{i,j}, a_{i+1,j})$ o $(a_{i-1,j}, a_{i,j})$, para una j específica. Esta matriz fue construida a partir del (**conjunto** $\mathcal{C}$) (véase el Cuadro [2.2]).

La matriz B muestra la probabilidad condicionada del péptido evaluado como resultado de dos condiciones:

- Que el péptido evaluado por los programas APAP y APAP-I es aceptado como PASAC inspecccionando únicamente los primeros 8 aminoácidos del péptido.
- Que el $\mathbf{\acute{I}ndice_A}$ (Ecuación [2.8]) es mayor o igual 0.08.

Cuadro 2.2

Elementos de la matriz A. Distribución de frecuencia absoluta de los péptidos antibacterianos contiguos que no tienen una acción específica en contra de bacterias (**conjunto** $\mathcal{C}$); donde el método usado para detectar su estructura tridimensional no se restringue a espectroscopía NMR, rayos X o dicróismo circular. Cada letra es equivalente a un aminoácido, de tal manera que la ocurrencia de la pareja de aminoácidos (Ac_i, Ac_j) es construida a partir del aminoácido del renglón (i) y el aminoácido de la columna (j).

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	165	38	11	29	28	134	17	85	205	116	6	40	15	24	34	76	50	64	9	2
C	50	87	15	7	34	32	24	30	85	43	4	17	41	7	142	41	30	43	4	54
D	24	17	8	3	15	10	2	32	25	27	6	2	4	7	4	16	20	28	9	9
E	14	8	5	8	6	20	15	9	48	24	4	7	4	8	44	19	8	13	3	3
F	27	69	17	4	23	48	14	25	62	109	1	14	33	18	49	34	9	26	3	10
G	103	51	19	39	61	107	25	137	164	185	15	39	74	55	102	62	64	89	33	53
H	14	18	5	11	15	18	19	17	15	29	6	3	10	4	25	19	14	57	0	4
I	95	40	13	25	43	143	31	53	108	69	10	31	53	21	59	68	28	45	6	19
L	105	55	34	25	60	155	24	55	143	129	9	23	108	26	65	80	22	51	28	10
M	15	4	6	5	2	11	0	6	12	18	0	8	1	5	12	6	2	7	1	2
N	30	17	6	8	27	37	10	14	29	36	11	9	23	4	36	12	23	36	3	7
P	43	16	8	7	45	47	16	79	45	36	6	21	55	17	69	30	18	64	13	17
Q	44	23	3	4	12	47	12	27	24	7	5	12	21	20	16	8	16	16	4	2
R	40	62	32	17	45	94	23	60	73	69	7	48	97	30	118	35	20	54	20	21
S	63	67	13	16	20	78	21	43	74	52	14	15	12	17	33	31	28	52	10	15
T	51	79	8	5	17	34	5	38	29	47	8	4	14	15	44	11	13	38	4	13
V	107	59	14	13	36	133	13	49	63	103	2	22	47	11	46	47	32	60	8	12
W	15	8	5	8	5	16	3	9	31	22	3	14	9	9	5	5	2	4	4	0
Y	10	51	3	2	6	30	1	13	22	20	1	11	9	5	39	14	19	11	0	8

El $\mathbf{\acute{I}ndice_A}$ (Ecuación [2.8]) esta formado por la distribución de frecuencia relativa de los aminoácidos A_i del péptido ensayado; ello se deriva de la distribución de la frecuencia absoluta de los péptidos antibacterianos del **conjunto** $\mathcal{B}$ (véase el Cuadro [2.3]).

Cuadro 2.3

Elementos del vector $\mathsf{Index_A}$. Frecuencia absoluta de los péptidos antibacterianos que actuan exclusivamente contra bacterias (**conjunto** $\mathcal{B}$) en el vector $\acute{\mathsf{I}}\mathsf{ndice_A}$. Las letras representan los 20 aminoácidos (una letra por aminoácido), y los números representan la frecuencia relativa correspondiente a cada aminoácido.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
49	68	10	16	35	106	22	71	112	52	2	30	33	19	60	42	24	45	17	27

$$\mathbf{\acute{I}ndice_A} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{A}_i, i \in [1, n] \tag{2.8}$$

La matriz B muestra la probabilidad condicionada de $\mathbf{P}(o_i|h_{i|_{\acute{\mathsf{I}}\mathsf{ndice_A}}})$ a ser candidato PASAC si $(o_i = verdadero)$ o sea $\mathbf{P}(o_i = verdadero|h_{i|_{\acute{\mathsf{I}}\mathsf{ndice_A}}}) = 0.95$, y su complemento $(o_i = falso)$ $\mathbf{P}(o_i = falso|h_{i|_{\acute{\mathsf{I}}\mathsf{ndice_A}}}) = 0.05$.

Donde o_i es la condición observable y h_i es la condición oculta, de manera que se preserve la Ecuación [2.9]

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \mathbf{P}(h_i|o_i) = \frac{\mathbf{P}(o_i|h_i)\mathbf{P}(h_i)}{\mathbf{P}(o_i)} \tag{2.9}$$

2.3.13. Distribución de péptidos en longitud 8aa, 9aa y 10aa.

Para determinar la frecuencia relativa de todos los péptidos en longitud 8aa, 9aa y 10aa por cada una de las propiedades fisicoquímicas de APAP–II, se optó por dividir cada intervalo de cada propiedad fisicoquímica en 9 intervalos iguales, este particionamiento sólo tuvo como objetivo detectar la existencia de cúmulos en los rangos de cada propiedad fisicoquímica y se consideró que no fuera necesario reducir aún mas el mísmo. Ello permitió observar la existencia de cúmulos agrupados por intervalos y en consecuencia abrir la posibilidad de reducir el intervalo correspondiente.

Bajo este procedimiento se analizaron todos los péptidos en longitud 8aa, 9aa y 10aa, formando dos grupos: aquellos que fueron candidatos PASAC y aquellos que no lo fueron.

Es importante notar que es posible llevar a cabo el cálculo completo particularmente, de los 20^8 péptidos en longitud 8aa en la plataforma PC-Intel-686 debido a que hay una diferencia importante, en el tiempo de procesamiento, entre calificar un péptido y grabarlo en un disco y sólamente contarlo sin grabarlo.

Para calcular la distribución de los 20^9 péptidos en longitud 9aa y los 20^{10} péptidos en longitud 10aa se usó los clusters: Cluster-20-Intel-686, Cluster-14-Xeon y Cluster-20-AMD, asignando $20^8/20$ péptidos a cada procesador de la plataforma Cluster-20-Intel-686, $20^9/20$ péptidos a cada procesador de la plataforma Cluster-14-Xeon y $20^{10}/20$ péptidos a cada procesador de la plataforma Cluster-20-AMD.

2.3.14. Distribución de péptidos antibacterianos en longitud 8aa.

Para determinar la frecuencia relativa de todos los candidatos PASAC en longitud 8aa se requirió del siguiente conjunto:

- Todos los péptidos antibacterianos depositados en la base de datos BBCM [2] en longitud 8aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

Distribuyendo cada candidato por propiedad fisicoquímica a través del programa APAP–I, y dividiendo cada una de éstas en 9 intervalos iguales (véase la Sección [2.3.13]).

Se separaron aquellos que fueron candidatos a PASAC de aquellos que no lo fueron.

Este proceso de desarrolló utilizándo la plataforma PC-Intel-686.

2.3.15. Control negativo usando péptidos antibacterianos naturales.

La prueba negativa sobre el ensayo expuesto en la Sección [2.3.13] requirió de los conjuntos $\mathcal{C}$, $\mathcal{D}$ **y** $\mathcal{E}$.

El **conjunto** $\mathcal{C}$ se analizó a través del MOM y se observó el número de péptidos que formaron un cúmulo, conteniendo a aquellos que previamente se habian identificado como PASAC **conjunto** $\mathcal{B}$.

El **conjunto** $\mathcal{D}$ se observó si era o no aceptado formando cúmulo con los candidatos PASAC.

El **conjunto** $\mathcal{E}$ se usó para construir nuevamente las matrices **A** y **B** y luego se evaluó con MOM el **conjunto** $\mathcal{C}$.

2.3.16. Detección de candidatos PASAC en longitud 8aa.

Para minimizar los candidatos PASAC en longitud 8aa dentro de los 20^8 posibles, se generaron los siguientes conjuntos:

- Todos los candidatos PASAC en longitud 8aa que el programa APAP–I detectó.
- Todos los péptidos antibacterianos depositados en la base de datos BBCM [2] en longitud 8aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].
- Todos los candidatos PASAC en longitud 8aa usados para ensayar el programa APAP y reportados por Del Rio G. [30].

Posteriormente, se designó como candidatos PASAC en esa longitud al conjunto intercepción de los tres conjuntos.

Los clusters usados fueron: Cluster-20-Intel-686, Cluster-14-Xeon y Cluster-20-AMD.

2.3.17. Candidatos PASAC en `magainina 2` y `cecropina A`.

Para predecir candidatos PASAC con longitud 8aa contenidos en `magainina 2` y `cecropina A`, se leyeron las secuencias del amino-terminal al carboxilo-terminal seleccionando las subsecuencias de longitud 8aa encontradas, moviendo un aminoácido a la derecha cada vez, hasta el final de la misma.

Finalmente, este grupo de subsecuencias o péptidos fue evaluado por medio del programa APAP–III.

La plataforma utilizada fue PC-Intel-686.

2.3.18. Identificación de regiones antibacterianas en péptidos cortos.

Los conjuntos mínimos donde todos los candidatos PASAC de longitud 8aa predichos en `magainina 2` y `cecropina A` (véase Sección [2.3.17]), fueron definidos como las regiones antibacterianas del péptido antibacteriano natural.

Dado un péptido de longitud n. Al mínimo conjunto que contiene a todas las subsucesiones de longitud inferior a n que hayan sido calificadas como candidato PASAC, tanto por APAP como por APAP–I, se le denomina región antibacteriana (véase los Apéndices [B, C, D y E])

Para dicho proceso se utilizó la plataforma PC-Intel-686 y el programa APAP–III.

2.3.19. Control negativo usando no candidatos PASAC cortos.

La prueba negativa requirió de los siguientes conjuntos:

- 20^8 péptidos de longitud 8aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].
- Un no candidato PASAC en longitud 8aa, seleccionado entre los que no coincidieron como candidato PASAC y que fueron reportados por Del Rio G. [30] (véase el Apéndice [A] y la Sección [2.3.18] para la definición de región antibacteriana).
- El péptido `gambicina` (véase el Apéndice [E]).
- El péptido hemolítico `melitina` (véase el Apéndice [D]).
- El péptido hemolítico `melitina` subdividido en subpéptidos de longitud 8aa hasta su longitud total, (véase el Apéndice [D] y la Sección [2.3.1]). La subdivisión tiene como objetivo verificar si subconjuntos del mismo péptido guardan alguna correlación con las propiedades fisicoquímicas del péptido completo.

Se interceptaron (esto es, se formó un grupos con aquellos elementos que estuvieron presentes en ambos conjuntos), los primeros dos conjuntos y se seleccionó aquel no candidato PASAC que mostró menor similitud entre ellos.

La plataforma utilizada fue PC-Intel-686.

2.3.20. Ensayo experimental de candidatos PASAC en longitud 8aa.

Dos pruebas experimentales se desarrollaron con la finalidad de verificar la efectividad del pétidos `CE1`:

- Se determinó la concentración mínima inhibitoria de (`CE1: KWKLFKKI`) en el intervalo de concentración de 200 μg/ml a 400 μg/ml.
- Se comparó, usando como sustancia control H_2O, el efecto de cuatro péptidos distintos: Los péptidos (`PAP1: KKLKLLLKLL`) y (`SAP3:KLKLKKKLK`), (`MA3: GKFLHSAK`) y (`CE1: KWKLFKKI`) péptidos diseñados con el programa APAP–III.

2.3.21. Estadística no paramétrica Ji-cuadrada en péptidos cortos.

La prueba estadística no paramétrica Ji-cuadrada [56] se usó para verificar si cada propiedad fisicoquímica de los candidatos PASAC en longitud 8aa se dispersan de acuerdo a la distribución Normal.

Una variable aleatoria continua X se dice que sigue una distribución Normal $X \sim N(\mu, \sigma)$ si su función de probabilidad $f(x)$, corresponde a la Ecuación [2.10], donde $\mu, \sigma \in \mathbf{R}, \sigma > 0$.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbf{R} \tag{2.10}$$

Los valores μ y σ se llaman parámetros de la distribución de la función $f(x)$, la cual representaremos por $N(\mu, \sigma)$, siendo μ y σ la esperanza matemática y la desviación estandar, respectivamente de la variable aleatoria X.

Si integramos la Ecuación [2.10], obtendremos la función de distribución Normal $F(x)$ Ecuación [2.11].

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{2.11}$$

Particularmente, la gráfica de una distribución normal estandar $N(0, 1)$ se obtiene a partir de la gráfica de $f(x)$ correspondiente a $N(\mu, \sigma)$, efectuando sobre esta última, una traslación de ejes situando el origen en el punto $(\mu, 0)$, y tomando como unidad en el eje-x el parámetro σ (es decir, haciendo $\sigma = 1$).

Aplicando la transformación expresada en la Ecuación [2.12] en la Ecuación [2.11] resultará la función de distribución normal estandar $\phi(z)$ Ecuación [2.13].

$$z = \frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.12}$$

El parámetro σ^2 representa la varianza.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}u^2}\, du \tag{2.13}$$

La independencia y aleatoriedad de la población de candidatos PASAC radica en lo expresado en el Teorema Central del Límite (véase el Teorema [2.3.1]).

Teorema 2.3.1 (Teorema Central del Límite) "Sea $X_1, X_2, \cdots, X_n$ una muestra aleatoria de tamano n variables independientes e identicamente distribuidas tomadas de una población infinita, con media μ y varianza σ^2, entonces la distribución límite de la Ecuación [2.11] es la distribución Normal Estandar $N(0,1)$ cuando $n \to \infty$, (independiente de la distribución de $X_1, X_2, \cdots, X_n$)".

De acuerdo a este teorema si la población de candidatos PASAC no se observa una distribución Normal, entonces su población no corresponde a una variable independiente y aleatoria y se concluiría que presenta un patrón de comportamiento sesgado.

Verificar que la distribución de frecuencias relativas por propiedad fisicoquímica de los PASAC se distribuye de acuerdo a la función $F(x)$ equivale a verificar la convergencia a cero de la Ecuación [2.14], donde las x_j son las frecuencias relativas observadas sobre la variable aleatoria X, y np_j son las frecuencias relativas esperadas para una muestra de tamaño n.

$$\sum_{j=1}^{n} \frac{(x_j - np_j)^2}{np_j} \tag{2.14}$$

Los conjuntos requeridos para la verificación estadística fueron las frecuencias absolutas y relativas por propiedad fisicoquímica de los candidatos PASAC en longitud 8aa.

La metodología seguida para efectuar la prueba no paramétrica Ji-cuadrada se basó en [94] y el cálculo final se resolvió en [84].

La plataforma utilizada para determinar las distribución de péptidos de longitud 8aa fue PC-Intel-686.

2.3.22. Análisis del programa APAP–C.

El programa APAP–C esta constituido por cuatro subprogramas, y cada uno de ellos evalua una propiedad fisicoquímica del péptido y verifica que coincida con el intervalo de cada una de las propiedades fisicoquímicas codificadas para el programa APAP–I y escrito en el lenguaje Handel-C (véase Sección [2.2.5]).

Su verificación requirió del siguiente conjunto de péptidos.

> 20^5 péptidos de longitud 9aa, generados bajo el mismo criterio que el expresado en la Sección [2.3.1].

Los procesadores FPGA-Xilinx usados fueron Procesador–A–FPGA y Procesador–B–FPGA. Al momento de escribir esta tesis no se habia confirmado la síntesis del programa APAP–C en Procesador–C–FPGA.

2.3.23. Estadística no paramétrica Wilcoxon, Mann y Whitney en péptidos antibacterianos.

La prueba estadística no paramétrica Wilcoxon, Mann y Whitney se usó para verificar si cada propiedad fisicoquímica de los candidatos PASAC hallados entre los péptidos antibacterianos reportados en la base de datos APD [105] a septiembre, 2007, se dispersan de acuerdo al conjunto que los contiene.

La prueba consiste en hallar los valores críticos c_1 y c_2 y suponer que ambas distribuciones de datos se dispersan de manera semejante, entonces la variabe aleatoria W sobre las poblaciones descritas será aproximadamente Normal con media y varianza (Ecuaciones [2.15] y [2.16])

$$\mu_W = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{2.15}$$

$$\sigma_W^2 = \frac{n_1 n_2(n_1 + n_2 + 1)}{12} \tag{2.16}$$

donde c_1 y c_2 se obtienen sustituyendo μ_W y σ_W en las Ecuaciones [2.17] y [2.18]

$$P(W \leq c_1) = \Phi\left(\frac{c_1 - \mu_W}{\sigma_W}\right) = 1.25\,\% \tag{2.17}$$

$$P(W \geq c_2) = 1 - \Phi\left(\frac{c_2 - \mu_W}{\sigma_W}\right) = 1.25\,\% \tag{2.18}$$

Capítulo 3

Resultados

En esta sección se exponen cronológicamente los resultados computacionales, matemáticos y biológicos de los diferentes métodos expuestos en el Capítulo 2.

Debido a que el presente reporte es resultado de un proyecto multidisciplinario se han clasificado los resultados de la manera siguiente:

- Resultado computacional. Se reporta cronológicamente el rendimiento de la familia de los programas involucrados con respecto a las diferentes plataformas computacionales.

- Resultado matemático. Se reporta exclusivamente el beneficio del aporte matemático enfocado en la localización de los péptidos o regiones antibacterianas ya explicadas, sin pretender entrar a demostración alguna de la herramienta matemática usada.

- Resultado biológico. Describe los péptidos y las regiones antibacterianas halladas en relación a los aminoácidos que los constituyen y sus actividades ensayadas en presencia de bacterias y células humanas.

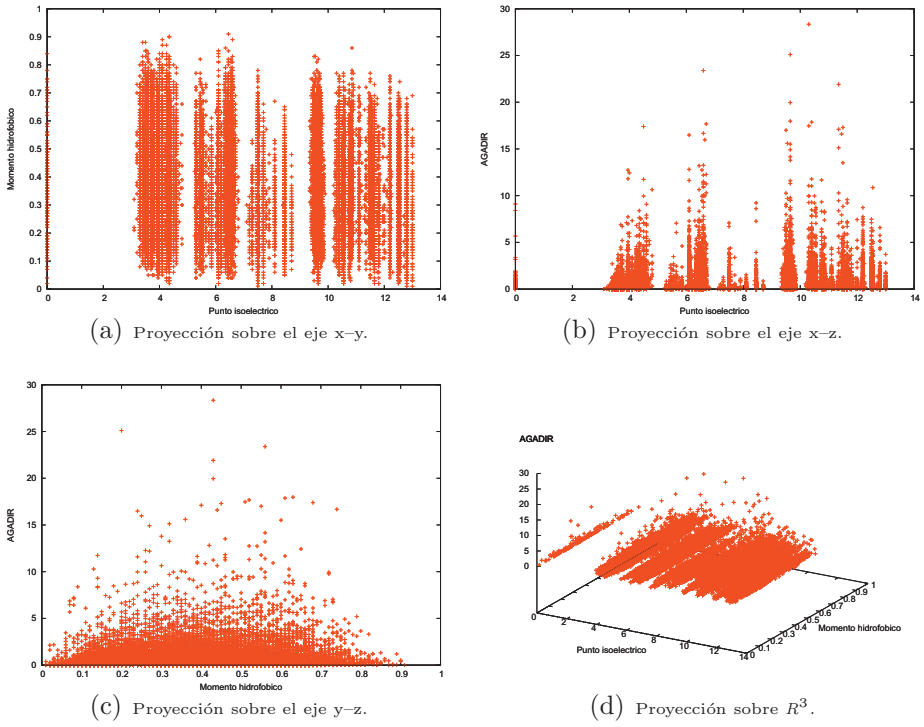
3.1. Cronología.

3.1.1. Relación entre AGADIR, Punto isoeléctrico y Momento hidrofóbico.

Todos los proteomas ensayados, forman cúmulos de candidatos PASAC que al ser evaluados con el programa APAP llevan a la suposición de que este agrupamiento no es una característica aislada (véase la Sección [2.2.1]) y en consecuencia las propiedades fisicoquímicas permiten distinguir PASAC.

Un ejemplo geométrico de la correspondencia bidimensional y tridimensional de las propiedades fisicoquímicas expresadas por el programa APAP, se puede observar

para el proteoma de `Aeropyrum Pernix NC-000854` (véase la Figura [3.1]).



(a) Proyección sobre el eje x–y. (b) Proyección sobre el eje x–z.

(c) Proyección sobre el eje y–z. (d) Proyección sobre R^3.

Figura 3.1

Cúmulos de candidatos PASAC en longitud 9aa, detectados por el programa APAP sobre el proteoma `Aeropyrum Pernix NC-000854`. Proyección sobre el eje x–y (Figura [3.1(a)]), proyección sobre el eje x–z (Figura [3.1(b)]), proyección sobre el eje y–z (Figura [3.1(c)]) y proyección sobre $\mathbf{R}^3$ (Figura [3.1(d)]).

3.1.2. ■ Significancia de AGADIR en péptidos cortos.

El parámetro fisocoquímico AGADIR no es un discriminante confiable ya que al evaluarse sobre todos los péptidos de longitud 9aa referidos en la Sección (2.3.3), sólo el 0.001 % es rechazado.

Ello significa que mayoritariamente la propiedad AGADIR acepta en su intervalo a prácticamente todos los péptidos en esa longitud.

Es importante indicar que si bien AGADIR no es un discriminante efectivo para péptidos cortos, en el caso de los péptidos antibacterianos extraidos de la base de datos ExPASy [9] si lo és [30].

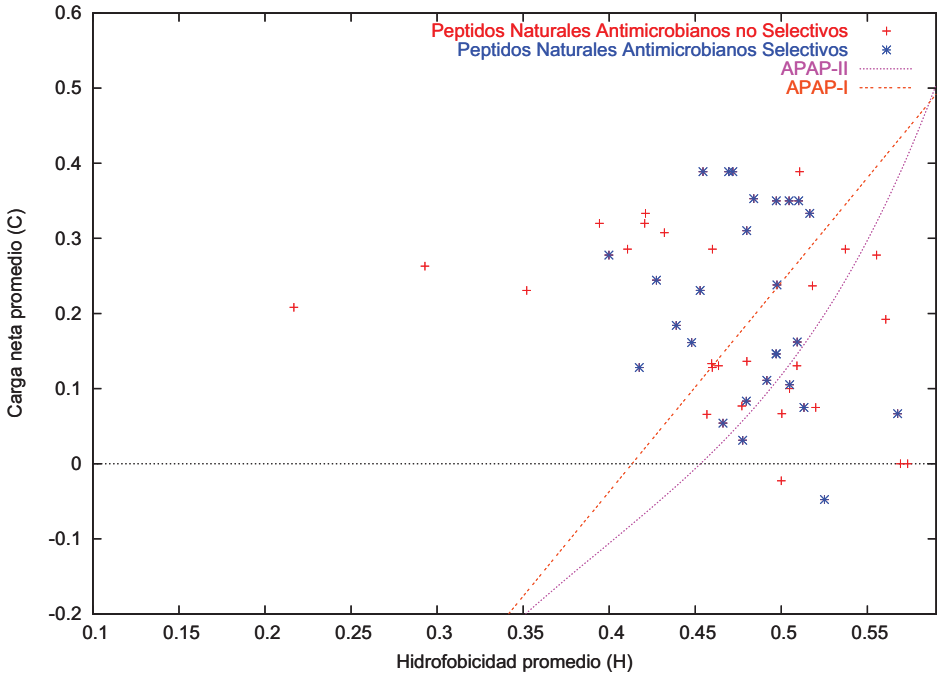


Figura 3.2

Relación entre los péptidos naturales antibacterianos con/sin actividad selectiva y nativamente estructurados, con respecto al programa APAP–I y la ecuación lineal Uversky V.N.

3.1.3. ■ Significancia de la no estructurabilidad en los candidatos PASAC.

Se determinó (véase la Sección [2.3.4]) que el porcentaje de similitud entre las propiedades AGADIR y la no estructurabilidad de un péptido era de 25 % de acuerdo a la Ecuación [2.3].

Como consecuencia de ello se diseñó analíticamente una curva de grado cuatro (véase la Ecuación [2.2]) por el método de mínimos cuadrados, que detectara hasta el 70 % de los candidatos PASAC que cumplieran con ser PNNE (véase la Figura [3.3] y la Sección [2.3.4]).

3.1.4. ■ Discriminación entre AGADIR y la estructurabilidad en péptidos cortos.

Los péptidos de longitud 9aa (véase la Sección [2.3.3]) rechazados tanto por AGADIR como por la curva correspondiente a la Ecuación [2.2] coinciden en 93 %.

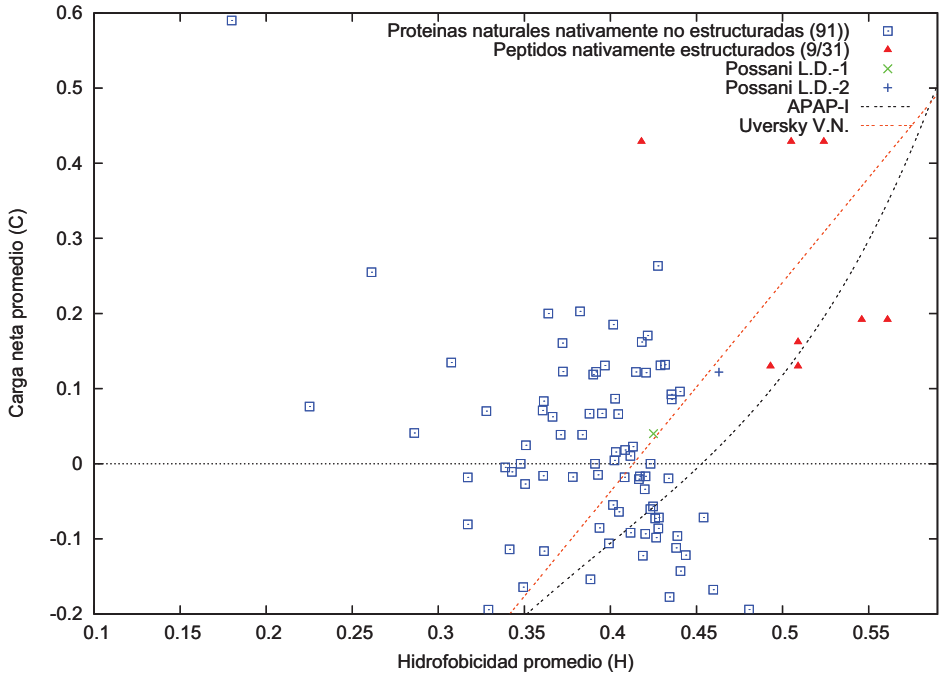


Figura 3.3
Relación entre péptidos nativamente no estructurados PNNE y péptidos nativamente estructurados PNE.

3.1.5. ■ Redundancia entre Momento hidrofóbico e Hidrofobicidad promedio en péptidos cortos.

Las pruebas correspondientes (véase la Sección [2.3.8]), mostraron una disparidad de 45 % entre ambas propiedades fisicoquímicas para los péptidos ensayados en longitud 9aa a partir de los 442 genomas secuenciados, por lo que se concluye que no son equivalentes ambas propiedades fisicoquímicas.

Tomando los péptidos aleatorios se encontró la región de equivalencia $[0.1, 0.6] \times [0.4, 0.7]$ (véase la Figura [3.4]), equivalente al 70 % de la región de correlación.

3.1.6. ■ Redundancia Carga neta promedio y Punto isoeléctrico en péptidos cortos.

Las pruebas correspondientes (véase la Sección [2.3.9]), mostraron una disparidad de 90 % entre ambas propiedades fisicoquímicas para los péptidos ensayados en longitud 9aa, por lo que se concluye que no son equivalentes ambas propiedades fisicoquímicas (véase la Figura [3.5]).

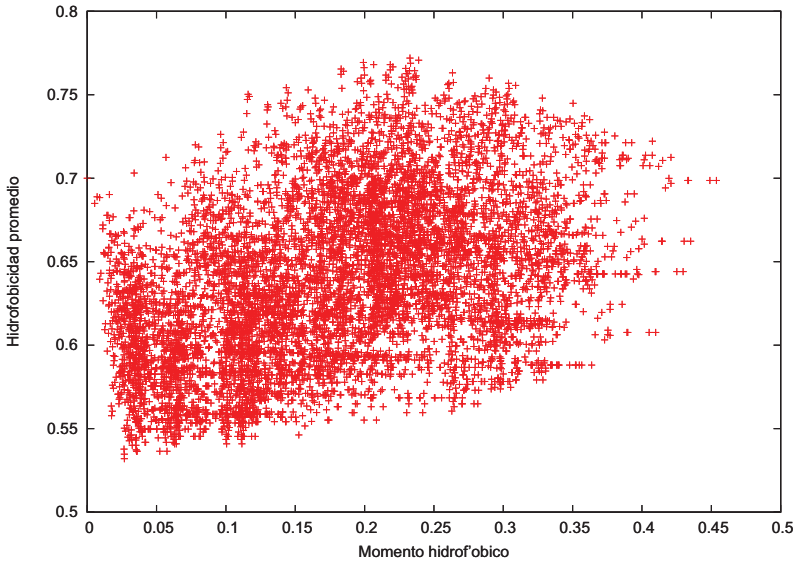


Figura 3.4

Relación entre las propiedades fisicoquímicas: Momento hidrofóbico e Hidrofobicidad promedio en péptidos de longitud 9aa.

3.1.7. ■ Contigüidad de los aminoácidos en los candidatos PASAC cortos.

Los candidatos PASAC en longitud 9aa distribuidos de acuerdo a su frecuencia indujeron la serie descrita en la Figura [3.6], que no mostró una evidente correlación geométrica entre los aminoácidos del péptido y su contigüidad (véase la Sección [2.3.7]). Como consecuencia de ello se descartó la frecuencia de contigüidad de los aminoácidos en los candidatos PASAC cortos.

3.1.8. ■ Candidatos PASAC en longitud 8aa, 9aa y 10aa versus péptidos antibacterianos.

La distribución de frecuencias relativas (véase la Sección [2.3.13]) para candidatos PASAC en longitud 8aa, 9aa y 10aa (véase el Cuadro [3.3]) muestran regiones coincidentes entre si, asi como con la distribución de péptidos antibacterianos (véase el Cuadro [3.4]). Lo anterior muestra que la subdivisión de cada intervalo asignado a cada propiedad fisicoquímica en nueve intervalos iguales, permitiría mejorar la detección de candidatos PASAC al reducirse los intervalos de aceptación.

El Cuadro [3.6] muestra un resumen gráfico por propiedad fisicoquímica de las frecuencias relativas en las longitudes indicadas, tanto para los candidatos APAP como para los no candidatos APAP.

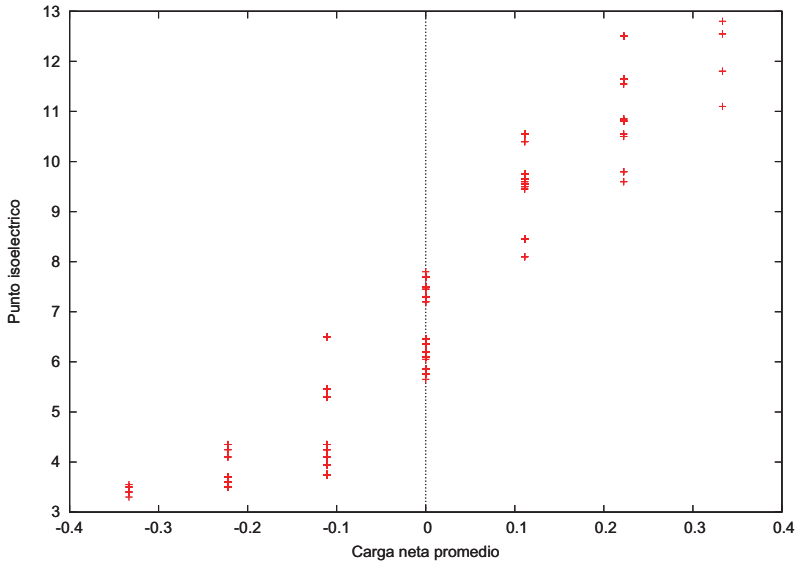


Figura 3.5

Relación entre las propiedades fisicoquímicas: Carga neta promedio y Punto isoeléctrico en péptidos de longitud 9aa.

Una de las consecuencias de este resultado es la reducción de los intervalos de aceptación para las propiedades fisicoquímicas H y MH a quedar: H de 0.35 a 0.46 y MH de 0.4 a 0.5 y la consecuente aceleración de los programas APAP–I y APAP–II, lo que conduciría a poder explorar expectros de péptidos de mayor longitud que 10aa.

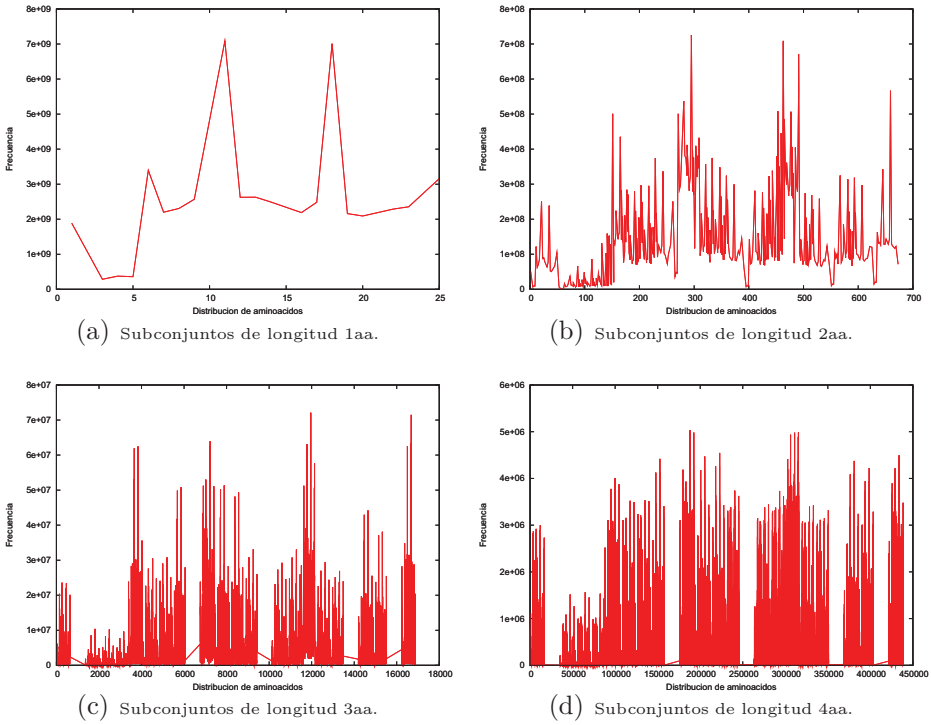
La observación de las distribuciones de los 20^8 candidatos PASAC y los 501 péptidos antibacterianos no se distribuyen en los mismos intervalos (véanse los Cuadros [3.3 y 3.4] y la Sección [2.3.14]), a excepción de la propiedad carga neta promedio C para los candidatos PASAC (+). En cuanto a los no candidatos PASAC (-) no se observa ninguna coincidencia.

Lo anterior permite concluir que la distribución estadística de los candidatos PASAC no corresponde a la de los péptidos antibacterianos.

3.1.9. ■ Rendimiento computacional.

El Cuadro [3.1] muestra el mejoramiento del tiempo de procesamiento del programa APAP con respecto al tiempo de procesamiento que consumen las versiones de los programas APAP–I y APAP–II, después de analizar cada péptido de longitud 9aa.

Note que el programa APAP–II en la plataforma Cluster-20-AMD produjo una mejora de 278,985 veces en su tiempo de procesamiento, con respecto al programa referido, mientras que APAP y APAP–I difieren en 4,582. La mejora de APAP con respecto a APAP–II se explica primordialmente por el hardware usado y las técnicas



(a) Subconjuntos de longitud 1aa. (b) Subconjuntos de longitud 2aa.

(c) Subconjuntos de longitud 3aa. (d) Subconjuntos de longitud 4aa.

Figura 3.6

Serie de frecuencias de aminoácidos contiguos de candidatos PASAC de longitud 9aa, tomados en longitud 1aa (Figura [3.6(a)]), 2aa (Figura [3.6(b)]), 3aa (Figura [3.6(c)]) y 4aa (Figura [3.6(d)]).

de programación paralela.

Las cifras reportadas en el Cuadro [3.1] fueron determinadas a través de scripts de Linux, salvo el rendimiento de la plataforma Cluster-20-AMD la cual se resolvió usando la interfaz MPICH (Véase la Sección [2.2.3]), debido a que esa plataforma sólo acepta esa interfaz y no scripts de Linux.

Cuadro 3.1

Tiempo de procesamiento de los programas computacionales con respecto al programa APAP.

Programa	Plataformas computacionales	Tiempo (s)	Velocidad (veces)
APAP	PC-Intel-686	0.080000000000	1
APAP–I	PC-Intel-686	0.000017457346	4,582
APAP–II	Cluster-20-Intel-686	0.000000725000	110,344
	Cluster-14-Xeon	0.000000543750	147,126
	Cluster-20-AMD	0.000000286753	278,985

Programa: Programa computacional usado. Plataforma computacional: Para una descripción detallada de las plataformas computacionales véase la Sección [2.1]. Tiempo: Tiempo de procesamiento en segundos que requieren las diferentes versiones del programa APAP para evaluar cada péptido de longitud 9aa. Velocidad: Tasa de aprovechamiento de las diferentes plataformas computacionales en número de veces con respecto al programa APAP.

3.1.10. ■ Evaluación de candidatos PASAC en longitud 8aa seleccionados.

Cinco candidatos PASAC fueron hallados (véase el Cuadro [3.2]) en `magainina 2` y `cecropina A`, cuando éstos fueron distribuidos sobre los 9 subintervalos (véase las Secciones [2.3.13, 2.3.17, 2.3.18 y 2.3.19]), en que se dividió cada propiedad fisicoquímica (véase el Cuadro [3.5]), se observó que los candidatos PASAC (`MA3: GKFLHSAK`) y (`CE1: KWKLFKKI`) representaron la mejor correlación con respecto a los péptidos antibacterianos naturales y los 20^8 posibles péptidos en longitud 8 aminoácidos (véase los Cuadros [3.3 y 3.4]), mientras que (`MA4: KFLHSAKK`), (`MA5: FLHSAKKF`) y (`MA7: HSAKKFGK`) mostraron una menor coincidencia con respecto a los dos conjuntos referidos.

Aqui el término coincidencia debe de entenderse en el sentido de la representatividad de un péptido en todos los subintervalos de mayor cúmulo, en que se dividió cada propiedad fisicoquímica.

Se ensayaron los péptidos aquí referidos, encontrándose que `CE1` es antibacteriano mientras que el resto no lo és. `CE1` se probó satisfactoriamente con un cultivo de células epiteliales de prepucio de humano como control de que el péptido no es tóxico a células de humano.

La toxicidad hacia células de humano siguió el siguiente procedimiento: Los fibroblastos del prepucio humano se hicieron crecer en una solución con alta concentración de glucosa (Invitrogen, Carlsbad, CA) adicionada con 10 % de suero bovino (Sigma, St. Louis, MO) y penicilina/strectomicina 100 U/ml (Invitrogen, Carlsbad, CA).

El cultivo fue incubado a una temperatura de 37C, en un ambiente a 95 % aire y 5 % dioxido de carbono con 95 % de humedad. Después de 24 horas, 15 microlitros de cada péptido en solución fue adicionado y 72 horas después se estimó la viabilidad de éstos usando LIVE/DEAD cytotoxicity kit (Molecular Probes L-3224) de acuerdo a las instrucciones del proveedor (4 .M EthD-1 y 2 .M Calcein AM fueron usadas).

Para este ensayo, las soluciones fueron preparadas con las siguientes concentraciones: `PH(Cecropin)1` = 4 mg/ml, `PH(Magainin)1` = 4 mg/ml, `PH(CeMa)1` = 4 mg/ml, `alpha-Pheromone-PH(Cecropin)1` = 4.8 mg/ml, `alpha-Pheromone-PH(CeMa)1` = 4.8 mg/ml, `CE1` = 4 mg/ml, `MA3` = 4 mg/ml, `MA4` = 12 mg/ml, `MA5` = 12 mg/ml, `MA7` = 20 mg/ml, `PH(SCAP*)6` = 12 mg/ml.

Estas concentraciones fueron seleccionadas al menos tres veces antes de detectar el índice terapéutico de cada péptido. Todas estas concentraciones fueron determinadas como el peso seco de cada péptido.

Cuadro 3.2

Candidatos a péptidos antibacterianos selectivos anfipáticos catiónicos (PASAC), encontrados en magainina 2 y cecropina A.

Tipo	Péptido	MH	PI	Eq.(1)	Eq.(2)	A
Magainina 2						
MA3	GKFLHSAK	0.42	10.8	25.0	11.0	7.0
MA4	KFLHSAKK	0.45	11.10	37.0	1.0	15.0
MA5	FLHSAKKF	0.56	10.80	25.0	4.0	19.0
MA7	HSAKKFGK	0.40	11.10	37.0	20.0	54.0
Cecropina A						
CE1	KWKLFKKI	0.49	11.30	50.0	3.0	1.66

Tipo: Nombre que identifica al péptido evaluado. Péptido: Las secuencias péptidicas están representadas por letras. Los péptidos referidos fueron aceptados tanto por el programa APAP como por el programa APAP–I. APAP–I: acepta el péptido si las propiedades fisicoquímicas siguientes se encuentran en los intervalos indicandos. MH: momento hidrofóbico [0.4,0.6], PI: punto isoeléctrico [10.8,11.8] y Eq. (1) $\geq$ Eq. (2), (véanse las Ecuaciones [2.1] y [2.2]). APAP: acepta el péptido si las propiedades fisicoquímicas siguientes se encuentran en los intervalos indicados MH [0.4,0.6], PI [10.8,11.8] y A (AGADIR) [0,10].

3.1.11. ■ Pruebas experimentales de candidatos PASAC en longitud 8aa.

De los cinco candidatos PASAC seleccionados (véase el Cuadro [3.2]) se ensayaron CE1 y MA3.

Las pruebas experimentales mostraron que (CE1: KWKLFKKI) tiene una Concentración Mínima Inhibitoria (CMI)de 350μg/ml (véase la Figura [3.7]).

(MA3: GKFLHSAK) no resultó tóxico a una concentración mínima inhibitoria de 400μg/ml (véase la Figura [3.8]), y (CE1: KWKLFKKI) resultó ser el mejor péptido en longitud 8aa con respecto a (MA3: GKFLHSAK), SAP3 y PAP1 (véase la Sección [2.3.20]).

3.1.12. ■ Regiones antibacterianas en magainina 2 y cecropina A.

La región antibacteriana mínima detectada por APAP–III en magainina 2 fue magainina 2(3-14) y en cecropina A cecropina A(1-8) (véanse los Apéndices B y C). La región cecropina A(1-8) se verificó experimentalmente con actividad antibacteriana. La experimentación de la región magainina 2(3-14) fue negativa (véase la Sección [3.1.10] acerca de la prueba experimental).

Cuadro 3.3

Distribución porcentual de 20^8, 20^9 y 20^{10} candidatos a péptidos antibacterianos selectivos anfipáticos catiónicos (PASAC) en longitud 8aa, 9aa y 10aa respectivamente. Reportando separadamente candidatos PASAC y no candidatos PASAC, en 9 intervalos (véase la Sección [2.3.13]).

Tipo	1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9	9/9
Longitud 8aa									
(+)									
(+)									
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.1	0.9
H	8.3	11.3	17.4	17.3	17.5	16.2	10.2	1.6	0.0
PI	0.0	51.4	0.1	6.6	0.3	5.5	32.5	3.2	0.0
MH	0.0	0.0	90.8	9.0	0.1	0.0	0.0	0.0	0.0
(–)									
C	8.7	0.0	23.1	0.0	34.0	0.0	23.0	10.7	0.3
H	7.9	9.7	11.2	11.9	12.0	11.4	10.1	8.5	6.8
PI	9.5	31.3	0.1	5.4	0.6	7.2	38.3	6.5	0.8
MH	69.9	23.9	5.3	0.7	0.0	0.0	0.0	0.0	0.0
Longitud 9aa									
(+)									
C	0.0	0.0	0.0	0.0	0.0	0.0	72.0	25.0	3.0
H	12.0	15.0	17.0	17.0	16.0	13.0	9.0	1.0	0.0
PI	0.0	39.0	0.0	8.0	1.0	9.0	37.0	7.0	0.0
MH	0.0	0.0	26.0	28.0	22.0	16.0	8.0	0.0	0.0
(–)									
C	0.0	10.0	23.0	0.0	32.0	23.0	9.0	2.0	0.0
H	9.0	11.0	13.0	13.0	14.0	13.0	11.0	9.0	7.0
PI	14.0	28.0	0.0	6.0	1.0	9.0	34.0	7.0	1.0
MH	21.0	19.0	16.0	13.0	10.0	8.0	6.0	4.0	2.0
Longitud 10aa									
(+)									
C	0.0	0.0	0.0	0.0	0.0	0.0	66.0	29.0	5.0
H	11.0	15.0	17.0	18.0	16.0	14.0	9.0	1.0	0.0
PI	0.0	38.0	0.0	10.0	2.0	10.0	33.0	7.0	0.0
MH	0.0	0.0	27.0	29.0	22.0	15.0	8.0	0.0	0.0
(–)									
C	3.0	10.0	0.0	22.0	30.0	22.0	9.0	3.0	1.0
H	8.0	11.0	13.0	14.0	14.0	13.0	11.0	9.0	7.0
PI	16.0	27.0	0.0	7.0	1.0	10.0	30.0	7.0	2.0
MH	22.0	20.0	16.0	13.0	10.0	7.0	5.0	3.0	2.0

(+): candidatos PASAC. (–): No candidatos PASAC. C: carga neta promedio. H: hidrofobicidad promedio. PI: punto isoeléctrico. MH: momento hidrofóbico. Los intervalos de igual longitud son resultado de dividir por nueve cada intervalo de las propiedades fisicoquímicas (véase la Sección [2.3.13]) del programa APAP–II. Los candidatos PASAC de valor máximo están sombreados en color ■ y los no candidatos PASAC de valor máximo sombreados en color ■.

3.1.13. ■ Control negativo en péptidos cortos.

Como control negativo se seleccionó un péptido de longitud 8aa, `AVVGQATQ` (véase el Apéndice [A]), el cual es representantivo del 75 % de la tendencia de los que no cumplen con ser candidatos PASAC (véase el Cuadro [3.7]).

Note que el péptido `AVVGQATQ` esta inserto en `cecropina A`, lo cual muestra que las regiones antibacterianas y no antibacterianas conviven en el péptido, de acuerdo al método expuesto. Sin embargo, este resultado experimentalmente no fue verificado.

Cuadro 3.4

Distribución porcentual de 501 péptidos antibacterianos naturales en longitud 8aa en 9 intervalos (véase la Sección [2.3.13]), reportando separadamente candidatos PASAC y no candidatos PASAC.

Tipo	1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9	9/9
(+)									
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	92.1	7.8
H	2.5	7.9	11.7	15.9	15.5	21.0	17.2	7.9	0.0
PI	0.0	42.9	0.1	27.2	7.4	3.5	14.3	4.1	0.0
MH	0.0	0.0	16.7	23.0	27.2	22.2	16.6	0.0	0.0
(–)									
C	0.8	0.0	5.4	0.0	24.4	0.0	37.0	32.1	1.7
H	5.8	8.2	10.7	12.4	13.9	15.5	15.9	17.3	15.7
PI	9.6	52.8	0.0	13.4	2.7	3.4	12.9	4.8	3.3
MH	15.4	14.5	13.2	13.8	13.0	11.1	10.4	8.1	5.3

(+): candidatos PASAC. (–): No candidatos PASAC. C: carga neta promedio. H: hidrofobicidad promedio. PI: punto isoeléctrico. MH: momento hidrofóbico. Los intervalos de igual longitud son resultado de dividir por nueve cada intervalo de las propiedades fisicoquímicas, (véase la Sección [2.3.13]) del programa APAP–II (véase la Sección [2.3.14]) Los candidatos PASAC de valor máximo están sombreados en color y los no candidatos PASAC de valor máximo sombreados en color .

Cuadro 3.5

Distribución de cinco candidatos a péptidos antibacterianos selectivos anfipáticos catiónicos (PASAC) en longitud 8aa, encontrados en `magainina 2` (MA) y `cecropina A` (CE) en 9 intervalos (véase la Sección [2.3.13]).

Tipo	2/9	3/9	4/9	5/9	8/9	9/9
C	0	0	0	0	0	MA3--7,CE1
H	MA7	0	0	MA4,5	MA3,CE1	0
PI	MA3,5	MA7	MA4,CE1	0	0	0
MH	0	MA3,7	MA4,CE1	0	MA5	0

C: carga neta promedio. H: hidrofobicidad promedio. IP: punto isoeléctrico. HM: momento hidrofóbico. Los intervalos de igual longitud son resultado de dividir por nueve cada intervalo de las propiedades fisicoquímicas (véase la Sección [2.3.13]), del programa APAP–II. `MA3`, `MA4`, `MA5`, `MA7` y `CE1`. Los candidatos PASAC que mejor interceptan los Cuadros [3.3] y [3.4] están sombreados en color

3.1.14. Control negativo en `gambicina`

El resultado con `gambicina` (véase el Apéndice [E]) fue negativo, esto es, APAP–II no lo seleccionó como candidato PASAC (véase la Sección [2.3.16]). Lo anterior concuerda con lo esperado ya que el péptido `gambicina` no es selectivo, por lo que no es candidato PASAC (véase la Sección [1]).

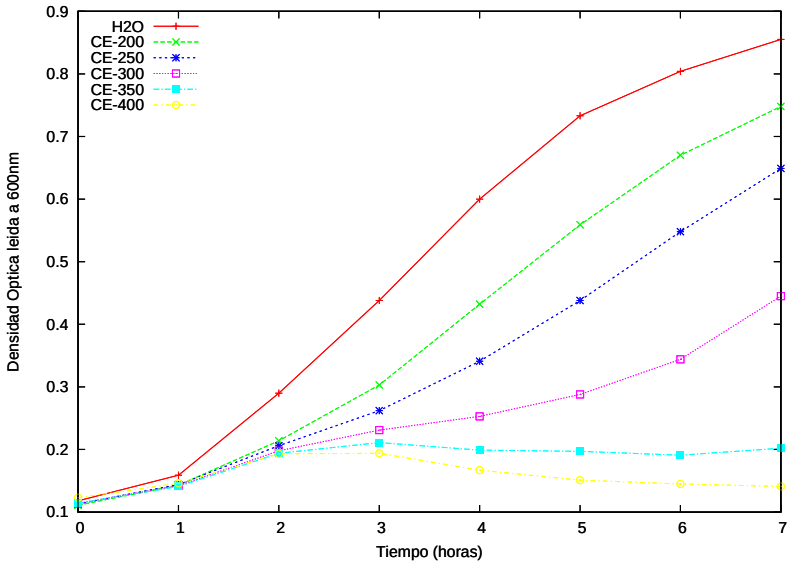


Figura 3.7

H_2O: sustancia control. `CE1-200: CE1` a una CMI de 200μg/ml. `CE1-250: CE1` a una CMI de 250μg/ml. `CE1-300: CE1` a una CMI de 300μg/ml. `CE1-350: CE1` a una CMI de 350μg/ml. `CE1-400: CE1` a una CMI de 400μg/ml. Concentración: μg/ml. Tiempo: horas.

Cuadro 3.6

Tendencia de la distribución porcentual de los candidatos y no candidatos PASAC en el intervalo de 8aa a 10aa de longitud. Por propiedad fisicoquímica e intervalos (véase la Sección [2.3.13]).

Tipo	1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9	9/9
Candidatos PASAC									
C									
H									
PI									
MH									
No candidatos PASAC									
C									
H									
PI									
MH									

C: carga neta promedio. H: hidrofobicidad promedio. PI: punto isoeléctrico. MH: momento hidrofóbico. Los intervalos de igual longitud son resultado de dividir por nueve cada intervalo de las propiedades fisicoquímicas (véase la Sección [2.3.13]), del programa APAP–II. Candidatos PASAC: sombreados en color . No candidatos PASAC: sombreados en color .

3.1.15. Regiones antibacterianas en `melitina`

No se detectó región antibacteriana alguna en `melitina` (véase el Apéndice [D]), esto es, APAP–II no seleccionó ninguna región como candidato PASAC (véase la

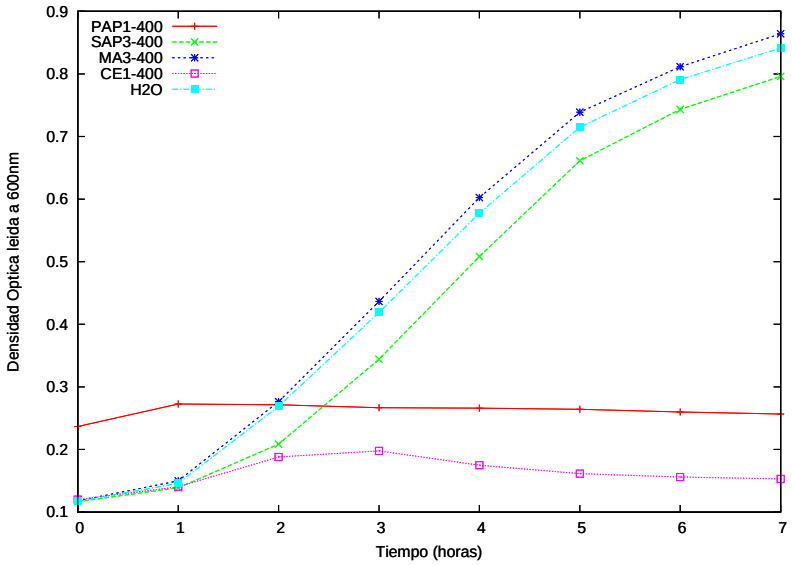


Figura 3.8

H_2O: sustancia control. `CE1-400`: (`CE1: KWKLFKKI`) péptido determinado por el programa APAP–III a una CMI de 400μg/ml. `MA3-400`: (`MA3: GKFLHSAK`) péptido determinado por el programa APAP–III a una CMI de 400μg/ml. SAP3-400: péptido experimental del grupo de investigación Del Rio G. PAP1-400: péptido experimental del grupo de investigación Del Rio G. Concentración: μg/ml. Tiempo: horas.

Cuadro 3.7

Distribución porcentual de no candidatos a péptidos antibacterianos selectivos anfipáticos catiónicos (PASAC) en longitud 8 aminoácidos, a partir de `magainina 2` y `cecropina A`.

Tipo	1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9	9/9
(–)									
C	0.0	0.0	4.0	0.0	39.0	0.0	31.0	24.0	0.0
H	0.0	3.0	18.0	18.0	6.0	12.0	3.0	28.0	9.0
IP	33.0	16.0	0.0	50.0	0.0	0.0	0.0	0.0	0.0
HM	11.0	14.0	5.0	2.0	0.0	20.0	17.0	14.0	11.0

(–): no candidatos PASAC. C: carga neta promedio. H: hidrofobicidad promedio. PI: punto isoeléctrico. MH: momento hidrofóbico. Los intervalos de igual longitud son resultado de dividir por nueve cada intervalo de las propiedades fisicoquímicas(véase la Sección [2.3.13]), del programa APAP–II. Los no candidatos PASAC de valor máximo están sombreados en color .

Sección [2.3.16]). Lo anterior concuerda con lo esperado ya que el péptido `melitina` es hemolítico, por lo que no es candidato PASAC (véase la Sección [1]).

3.1.16. Candidatos PASAC cortos hallados por modelos observables de Markov.

La relación referida (véase la Sección [2.3.10] y el Cuadro [3.9]) no aportó ningún conjunto de candidatos PASAC significativo, debido a que calificó como candidatos a todos los péptidos con una diferencia apenas observable en la posición decimal veinte.

Se entenderá por ello que este modelo observable de Markov no creó una matriz bien definida, de manera que pudiera identificarse candidatos PASAC, ya que asignó valores probabilísticos iguales hasta las primeras veinte cifras decimales.

3.1.17. Candidatos PASAC cortos hallados por modelos ocultos de Markov.

Este método (véase la Sección [3.9] y el Cuadro [3.9]), aportó la detección del candidato PASAC (`CE1: KWKLFKKI`) dentro de los péptidos antibacterianos de longitud 9aa usados para la prueba y no identificó, entre los primeros 31 mejores candidatos PASAC a ningún PASAC (véase el Cuadro [3.8]).

3.1.18. Candidatos PASAC hallados entre péptidos antibacterianos naturales.

Se detectó un cluster de 57 péptidos antibacterianos (véase el Cuadro [3.11]), que actuan exclusivamente contra bacterias, hongos, virus y células cancerígenas de mamíferos y cuya estructura tridimensional es conocida. Indicados en la base de datos APD [105] (**conjunto** $\mathcal{C}$).

Se hallaron 9 PASAC (9,14,15,16,19, 32, 57, 458 y 459), 6 péptidos antibacterianos híbridos formados por `Cecropina A` y `Magainina 2` (3,9, 14,15 490 y 493), 19 péptidos pertenecientes a la familia `Cecropina A` (9,14,15,16 58, 61, 119, 172, 173, 174, 175, 259, 424, 425, 434, 435, 454, 490 y 493), 4 péptidos de la familia `Brevinina` (125, 127, 380 y 386), 3 péptidos de la familia `Catelina` (19, 176, y 459) y 2 péptidos de la familia `Moricina` (32 y 52).

El péptido 32 (posición 20 en el Cuadro [3.11]), no fue aceptado por los programas APAP y APAP-I pero si lo fue por MOM (véase el Apéndice [F]).

Los resultados anteriores permiten establecer que asignar la verificación de las propiedades fisicoquímicas (véase la Sección [2.2.2]), como probabilidad condicional en MOM resultó ser favorable para la detección de candidatos PASAC entre los péptidos antibacterianos.

Cuadro 3.8
Evaluación de MOM sobre candidatos PASAC en longitud 9aa hallados en la naturaleza.

Candidato MOM	Evaluación MOM
WKKKIAKIG	0.000000000000000391249525781302912
WKKIAAKIG	0.000000000000000385884967519710394
WKKIAKKIG	0.000000000000000383063899077588438
FLKKKIAAK	0.000000000000000371580367977863585
FLKKIAKKK	0.000000000000000363089439844539676
WKKKIAKKI	0.000000000000000360478728537742206
FLKKIAAKK	0.000000000000000359789190566239598
WKKKIGKIG	0.000000000000000355631267798748321
WKKKIKKIG	0.000000000000000355383635789972957
WKKIKKKIG	0.000000000000000354366088850178671
HVAKKIAAK	0.000000000000000350438958516449861
FLKKKIAKK	0.000000000000000346376154193748530
WKKIGKKIG	0.000000000000000345328257870828869
FLKKKIAAA	0.000000000000000336698814331712929
FLKKAKKIG	0.000000000000000333099079741747989
FLKAAKKIG	0.000000000000000330288113083173888
FLKKIGKKK	0.000000000000000330034796996606739
FLKKIKKKK	0.000000000000000329804988239205212
FLAKKIAAK	0.000000000000000329613174980219772
WKKKIGKKI	0.000000000000000327661757514862201
WKKKIKKKI	0.000000000000000327433601144601300
FLKKIAKIG	0.000000000000000327055674704106827
HVAKKIAKK	0.000000000000000326668761837874497
FLGLKKKIG	0.000000000000000325098280794748037
HVAAKIAKK	0.000000000000000323912059541861432
FLKKKIGKK	0.000000000000000322268243609051490
FLAKIAKKK	0.000000000000000322081232978589461
FLKIAKKIG	0.000000000000000320213096022092726
FLKKKIKKK	0.000000000000000319848066906225396
HVAKKIAAA	0.000000000000000317542023197410242
FLKKKAKKI	0.000000000000000312562289692513385

Candidato MOM: Candidato PASAC en longitud 9aa. Evaluación MOM: Valor asignado por modelos ocultos de Markov (MOM), el cual pertenece al intervalo cerrado [0,1]. Los péptidos que están sombreados en color fueron aceptados tanto por el programa APAP–I como por el programa APAP.

3.1.19. Pruebas no paramétricas Ji-cuadrada sobre péptidos cortos.

La prueba Ji-cuadrada (véase la Sección [2.3.21]), muestra valores muy lejanos al cero por ello se observa que la muestra de los candidatos PASAC en longitud 8aa no se asemeja a una distribución Normal (véase el Cuadro [3.10]).

De lo anterior se concluye que los candidatos PASAC no pueden inferirse a partir de ninguna muestra normal, representativa de la población total de péptidos en longitud 8aa. Quedando pendiente la caracterización de la distribución estadística de un PASAC.

Cuadro 3.9

Evaluación de MOM sobre los péptidos antibacterianos naturales en longitud 9aa.

Candidato MOM	Evaluación MOM
GRFKRFRKK	0.000000000000000043327113470206797
GLRKRLRKF	0.000000000000000018805192352439615
VGRFRRLRK	0.000000000000000011769683548138165
KIKWFKTMK	0.000000000000000005518613980181834
RGFRKHFNK	0.000000000000000003773976503217658
KLKLFKKIG	0.000000000000000003672289234316443
GWLKKLGKR	0.000000000000000003589871721907159
GWLRKIGKK	0.000000000000000003110033958853967
KNLRRITRK	0.000000000000000002935003833444232
RGLRRLGRK	0.000000000000000002714626650866182
KWKLFKKIP	0.000000000000000002679588819723156
AKIPIKAIK	0.000000000000000002654570345188928
KWKLFKKIS	0.000000000000000002630963647488743
KNLRRIIRK	0.000000000000000002618128697804314
KWKLFKKIG	0.000000000000000002605305625365854
GWLKKIGKK	0.000000000000000002596449240617689
GGLKKLGKK	0.000000000000000002585553969264380
GLLKRIKTL	0.000000000000000002244817819206022
KAKLFKKIG	0.000000000000000002237495301900073
AKRHHGYKR	0.000000000000000002212901132033290
GIFSKLGRK	0.000000000000000002208783786188830
RRIRPRPPR	0.000000000000000002054339825107147
KWKLFKKIL	0.000000000000000001925271512425512
GKPRPYSPR	0.000000000000000001565175819331401
PKRKSATKG	0.000000000000000001383620262230333
FKLGSFLKK	0.000000000000000001357956439152334
GGLRSLGRK	0.000000000000000001282004370960111
KRLFKKLLF	0.000000000000000001273444518525477
RFRPPIRRP	0.000000000000000001111264878623090
ALWKTMLKK	0.000000000000000001070570110947266
KIGAKIKIG	0.000000000000000001035879309238745

Candidato MOM: Péptido antibacteriano en longitud 9aa. Evaluación MOM: Valor asignado por modelos ocultos de Markov (MOM), el cual pertenece al intervalo cerrado [0,1]. Los péptidos que están sombreados en color fueron aceptados tanto por el programa APAP–I como por el programa APAP.

Cuadro 3.10

Estadística Ji-cuadrada de candidatos PASAC de longitud 8aa por propiedad fisicoquímica.

$\chi^2_{PI_{(8)}}$	$\chi^2_{MH_{(8)}}$	$\chi^2_{C_{(8)}}$	$\chi^2_{H_{(8)}}$
1,811,753	1,204,740	31,037,978,516	11,422,014

Valores de $\chi^2_{\mathsf{pf}_{(\mathsf{gl})}}$: Ji-cuadrada correspondiente a la propiedad fisicoquímica pf para gl grados de libertad, .C: carga neta promedio. H: hidrofobicidad promedio. PI: punto isoeléctrico. MH: momento hidrofóbico (véase la Sección [2.3.21]).

3.1.20. Pruebas no paramétricas Wilcoxon, Mann y Whitney sobre péptidos antibacterianos.

Con objeto de verificar la similitud estadística entre los **conjuntos** $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ **y** $\mathcal{E}$ (véase la Sección [2.3.12]), compararemos únicamente dos de los conjuntos más si-

milares (**conjunto $\mathcal{B}$ y $\mathcal{C}$**) con un p-value<0.025 bilateral (ello equivale a 0.0125 de tolerancia).

No se observó ninguna correlación Normal entre ambos conjuntos por lo que se concluye que no hay una relación estadística de este tipo entre los conjuntos ($\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ y $\mathcal{E}$) bajo una semejanza de 98.5 %.

NL	NP	F	APAP	APAP-I	Nombre de la secuencia peptídica	Refs.
1	454	C	•	•	Cecropin-B type 1 precursor (Cecropin-B1) [Contains: Cecropin-B (AalCecB); Cecropin-B amidated isoform].	[96]
2	417		•	•	Parabutoporin	[70]
3	16	S, C	•	•	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P3).	[73]
4	61	C	•	•	Cecropin B [Bombyx mori].	[97]
5	458	S	•	•	cathelin-like protein [Mus musculus].	[83]
6	172	C	•	•	Hyphancin-3D precursor (Hyphancin-IIID) (Cecropin-A).	
7	15	S, C	•	•	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P4).	[73]
8	58	C	•	•	Cecropin-B	[86]
9	174	C	•	•	Hyphancin-3F precursor (Hyphancin-IIIF) (Cecropin-A2).	
10	68		•	•	Defensin NP-3a [Oryctolagus cuniculus].	[61]
11	57	S	•	•	Cecropin-A precursor (Cecropin-C).	[40]
12	425	C	•	•	RecName: Full=Cecropin-A.	
13	175	C	•	•	Hyphancin-3G precursor (Hyphancin-IIIG) (Cecropin-A3).	
14	259	C			Cecropin-A1 precursor (Cecropin-A) (AalCecA).	[96]
15	356				Ranatuerin-2Lb.	[39]
16	173	C	•	•	Hyphancin-3E precursor (Hyphancin-IIIE) (Cecropin-A1).	
17	176	Ca	•	•	Cathelin-like protein [Mus musculus].	[83]
18	474				Sentrin/SUMO-specific protease [Plasmodium yoelii yoelii str. 17XNL].	[25]
19	67		•	•	Defensin NP-3a [Oryctolagus cuniculus].	[61]
20	32	S, M			Chain A, Solution Structure Of Antibacterial Peptide (Moricin).	[43]
21	52	M			Moricin [Bombyx mori].	[43]
22	9	S, C	•	•	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide.	[73]
23	74				GK14120 [Drosophila willistoni].	[109]
24	426	M			RecName: Full=Virescein.	
25	106		•	•	Xenopsin precursor protein [Xenopus laevis].	[71]
26	169		•	•	Antibacterial peptide PMAP-37 precursor (Myeloid antibacterial peptide 37).	[100] [100]
27	435	C			Cecropin 1 [Musca domestica].	[100]
28	424	C			Cecropin precursor.	[22]
29	75				Sarcotoxin-1B precursor (Sarcotoxin IB).	[52]
30	355		•	•	Hadrurin	[99]
31	434	C			Cecropin-1 precursor	[89]
32	493	C	•	•	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P2).	[73]
33	490	C	•	•	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P3).	[73]
34	14	S, C	•	•	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P2).	[73]
35	469		•	•	Ribosomal protein L1 [Helicobacter pylori G27].	
36	386	B			Brevinin-1SY.	[67]
37	102				Megakaryocyte stimulating factor [Trichomonas vaginalis G3].	[25]
38	127	B			RecName: Full=Brevinin-1.	[72]
39	405				Maximin-H14 antimicrobial peptide precursor [Bombina maxima].	
40	267				Neutrophil defensin 3 (HANP-3).	[64]
41	265				Neutrophil defensin 1 (HANP-1).	[64]
42	380	B			Brevinin 1Pb precursor [Rana pipiens].	[98]
43	119	C			Cecropin C CG1373-PA [Drosophila melanogaster].	[46]
44	56				Bombinin.	

(continuación)

Cuadro 3.11 (continuación)

NL	NP	F	APAP	APAP-I	Nombre de la secuencia peptídica	Refs.
45	263				Fabatin precursor [Vicia faba].	
46	262				Fabatin precursor [Vicia faba].	
47	125	B			Brevinin-1E.	[66]
48	346				Ponericin-W2.	[74]
49	345				Ponericin-W1.	[74]
50	132				Ceratotoxin A [Ceratitis capitata].	[89]
51	239				Gaegurin-6.	[78]
52	488				Nigrocin-2P precursor [Rana palustris].	
53	137				Ranalexin precursor.	[26]
54	392				Temporin-1Ca.	[42]
55	19	S, Ca			Cathelin-related peptide SC5 precursor 1 (Antibacterial peptide SMAP-29) (Myeloid antibacterial peptide MAP-29).	[63]
56	112				Defensin related cryptdin 4 [Mus musculus].	[95]
57	459	S, Ca			Cathelin-like protein [Mus musculus].	[83]

Cuadro 3.11

Cúmulo de péptidos antibacterianos naturales predichos por MOM y listados en orden descendente (**conjunto** $\mathcal{C}$). NL: Posición del péptido antibacteriano en la lista. NP: Número que corresponde a péptido antibacteriano de acuerdo a MOM. F: Familia. Si el PASAC es parte del (**conjunto** $\mathcal{B}$) [S]. Si pertenece a `Brevinina` (Brevinin) [B]. Si pertenece a `Catelina` (Cathelin) [Ca]. Si pertenece a `Cecropina` [C]. Si pertenece a `Moricina` (Moricin) [M]. APAP: Péptido aceptado por el programa APAP (Sección 2.2). APAP-I: Péptido aceptado por el programa APAP-I (Sección 2.2).

3.1.21. Prototipo del programa APAP–C.

Las rutinas correspondientes a las propiedades fisicoquímicas: punto isoeléctrico (PI), carga neta promedio (C) e hidrofobicidad promedio(H) se probaron separadamente en el Procesador–A–FPGA y en el Procesador–C–FPGA.

Posteriormente, se completaron todas las rutinas de manera que se creó el programa APAP–C, el cual es un prototipo del programa APAP–I.

Los 20^4 candidatos PASAC de longitud 9aa (véase la Sección [2.3.1]), que el programa APAP–C predijo se contrastaron con los que produjo el programa APAP–I. El conjunto resultante empató con los resultados del programa APAP–I y fue sintetizado en el Procesador–C–FPGA

El tiempo de ejecución promedio por secuencia peptídica que arrojó APAP–C en la plataforma Procesador–B–FPGA fue 0.000258s.

El desempeño de APAP–C es sensible a la plataforma utilizada; notará que APAP–C al ejecutarse en la plataforma Procesador–B–FPGA es mayor que el del programa APAP–I, sin embargo, en la plataforma Procesador–C–FPGA se estima sería inferior a 30 veces (véase el Cuadro [3.12]). Este último cálculo fue predicho con base en las estimaciones obtenidas debido a las réplicas y memoria de este modelo. Queda pendiente la implementación de APAP–C en la plataforma Procesador–C–FPGA.

El comparativo entre clusters (véase el Cuadro [3.13]) se lleva a cabo primeramente

Cuadro 3.12

Desempeño de APAP–C versus APAP–I con respecto a un procesador Intel y procesadores FPGA.

Programa	Plataformas computacionales	Tiempo (s)
APAP–I	PC-Intel-686	0.000017457346
APAP–C	Procesador–B–FPGA	0.000258000000
APAP–C	Procesador–C–FPGA	0.000006450000

Programa: Programa computacional usado. Plataforma computacional: Para una descripción detallada de las plataformas computacionales véase la Sección [2.1]. Tiempo: Tiempo de procesamiento en segundos que requieren las diferentes versiones del programa APAP–C para evaluar cada péptido de longitud 9aa en promedio.

comparando las plataformas, luego se midió el tiempo de procesamiento que cada una de éstas consumió en analizar cada secuencia peptídica de longitud 9aa. Note que sólo es superado el cluster de FPGAs por el cluster correspondiente a una partición de la supercomputadora Kam Balam.

Cuadro 3.13

Desempeño de clusters versus cluster-FPGA, comparándose plataformas FPGA, Intel, Xeon y AMD (supercomputadora).

Programa	Plataformas computacionales	Tiempo (s)
APAP–C	Cluster-Procesador–C–FPGA	0.000000645000
APAP–II	Cluster-20-Intel-686	0.000000725000
	Cluster-14-Xeon	0.000000543750
	Cluster-20-AMD	0.000000286753

Programa: Programa computacional usado. Plataforma computacional: Para una descripción detallada de las plataformas computacionales véase la Sección [2.1]. Tiempo: Tiempo de procesamiento en segundos que requieren las diferentes plataformas para evaluar cada péptido de longitud 9aa.

Capítulo 4

Discusión

Los péptidos antibacterianos selectivos anfipáticos y cationicos (PASAC) son una nueva clase de péptidos con potenciales aplicaciones en el tratamiento de la salud [29, 34, 37, 55]. Sin embargo, no más de mil péptidos antibacterianos son conocidos hasta ahora y estos podrían representar sólo una pequeña fracción del arsenal que los organismos contemplan en contra de las bacterias [88].

Es aceptado que el objetivo de los PASAC son las membranas bacterianas o algún objetivo intracelular[41, 44] pero en cualquiera de los casos se sabe que ocurre el reconocimiento membrana-péptido (véase la Sección [1]). La atracción de los PASAC por las membranas bacterianas podría usarse para diseño y/o clasificación de los distintos tipos de péptidos en familias de péptidos. Entre los mecanismos de acción de los PASAC en contra de las membranas bacterianas [24, 41, 47, 68] esta la afectación de la tensión superficial de la membrana exterior de la bacteria y el mecanismo de barril en el que aparentemente el péptido penetra la membrana.

Dado lo anterior, es aceptable suponer que el mecanismo de acción para reconocimiento de la membrana, tenga que ver con el efecto electrostático e hidrofóbico: péptidos cargados positivamente podrían ser atraídos por membranas negativamente cargadas y entonces la parte hidrofóbica del péptido podría ser atraída por los lípidos de la membrana.

Bajo esta hipótesis las propiedades fisicoquímicas podrían ser suficientes para determinar la actividad de un PASAC y diferentes enfoques se han reportado para hallar PASAC: modelos probabilísticos por medio de modelos ocultos de Markov [82], modelos estadísticos [50] y computacionales.

Esta tesis resume los esfuerzos desarrollados para hallar PASAC en el intervalo de 8aa a 10aa y la detección de algún patrón de manera que puedan ser identificados dentro de todos los posibles péptidos construibles (desde 20^8 hasta 20^{10}), a través de la detección de propiedades fisicoquímicas.

Como primer obstáculo el programa que da origen a este proyecto doctoral consume 0.08 segundos en evaluar cada péptido de longitud 9aa, ello significa que evaluar 20^9 péptidos de longitud 9aa con una computadora personal consumiría $1,298$ años.

Para ello se mejoró el programa computacional APAP [30] (véase la Sección [2.2.1]) el cual evalua tres propiedades fisicoquímicas: Punto isoeléctrico, Momento hidrofóbico y AGADIR, sustituyéndose esta última propiedad por la propiedad de que el péptido sea nativamente no estructurado (véase la Sección [2.3.4]). Cabe mencionar que originalmente Uversky V.N. [103, 104] resuelve ello a nivel de proteinas, separando las nativamente no estructuradas y las nativamente estructuradas a través de la Ecuación [2.3] (véase la Sección [2.3.4]), donde se relacionan dos propiedades fisicoquímicas: Carga neta promedio e Hidrofobicidad promedio.

Nuestra investigación llevó a una adecuación de la ecuación referida por una de grado cuarto (Ecuación [2.2]) (véase la Sección [2.2.1]), verificando que es viable, para péptidos cortos, calificar como nativamente no estructurados a los péptidos situados por arriba de la ecuación.

Cabe mencionar la relevancia de esta equivalencia en péptidos cortos, ya que independientemente a la reducción significativa en el tiempo de procesamiento del programa APAP, producto de la sustitución de AGADIR, se hace presente una equivalencia fisicoquímica que pretende caracterizar PASAC.

La sustitución de la propiedad AGADIR por la evaluación de la ecuación referida, además de las mejoras computacionales a la rutina Punto isoeléctrico, aceleró el programa APAP 4,572 veces (véase la Sección [3.1.9]), y el uso de cómputo intensivo a través del ensayo en tres diferentes clusters (véase la Sección [2.2.3]), permitió acelerar al programa APAP en 278,985 veces (véase el Cuadro [3.1] y la Sección [3.1]).

Como resultado de las mejoras anteriores el programa APAP modificado terminó constituyéndose por las siguientes propiedades fisicoquímicas (véase la Sección [2.2.2]):

- Punto isoeléctrico (PI). Rango: [10.8 a 11.8].
- Momento hidrofóbico (MH). Rango: [0.4 a 0.6].
- Hidrofobicidad promedio (H). Rango: [0.35 a 0.55].
- Carga Neta promedio (C). Esta se encuentra determinada por la Ecuación [2.1] la cual depende únicamente de la Hidrofobicidad promedio.

Las mejoras computacionales a las rutinas que integran APAP y la equivalencia biológica péptidos no estructurados–PASAC, además de la reprogramación para cómputo paralelo [75] de APAP, permitieron explorar en horas los espacios totales (desde 20^8 hasta 20^{10}) de péptidos construibles, tarea que no era posible con el programa APAP original, y lo cual constituye un hecho relevante ya que antes de esta optimización no se contaba con referencia alguna de una inspección exhaustiva de estos espacios para tal fin.

A partir de este nuevo programa APAP, se creó un programa escrito en el lenguaje Handel-C [48, 69, 77] de manera que pudiera ejecutarse en un FPGA. Esta tarea

corresponde al ámbito de la microprogramación y no se habia explorado la posibilidad de usarla para la predición de PASAC, aunque si se habia hecho uso de FPGAs para determinación de péptidos [20, 28]. Es importante aquí mencionar que el objetivo de usar FPGAs para reprogramar la nueva versión de APAP, proviene de la idea de proveer a futuro de un hardware-software independiente e instalable en un laboratorio el cual en horas resuelva la evaluación de diversos conjuntos de péptidos, sin necesidad de conectarse a cluster o supercomputadora alguna.

Ya reducido sustancialmente el tiempo de procesamiento del programa APAP se procedió a explorar los espacios peptídicos completos para 8aa,9aa y 10aa, encontrándose que sólo el 1.3 % de los péptidos que integran cada espacio son, de acuerdo al programa APAP, candidatos PASAC.

Dado que este porcentaje es una cifra impráctica para ser ensayada biológicamente se procedió tanto a reconocer con mayor cuidado los intervalos de las propiedades fisicoquímicas que integran a APAP, como a buscar en ciertos subconjuntos de péptidos a candidatos PASAC. La primera parte de estas acciones llevó a subdividir en igual longitud cada intervalo de las propiedades fisicoquímicas de APAP, y la otra medida llevó a buscar estos candidatos dentro de péptidos antibacterianos hallados en la naturaleza y conocidos por su propiedad antibacteriana: `magainina 2` y `cecropina A`.

La primera acción permitió reducir los intervalos de las propiedades fisicoquímicas Momento hidrofóbico y Punto isoeléctrico al observar la frecuencia de aparición por subintervalo en cada propiedad (véase la Sección [3.1.12]), además se determinó la no relación entre las parejas de propiedades fisicoquímicas: Momento hidrofóbico e Hidrofobicidad promedio (véase la Sección [3.1.5]) y entre Carga neta promedio y Punto isoeléctrico (véase la Sección [3.1.6]). Todo ello llevó no sólo a reconocer aun más la naturaleza de los PASAC, sino a acelerar el programa debido a que se redujeron los intervalos originales de las propiedades fisicoquímicas referidas.

La segunda acción llevó a la localización de cinco candidatos PASAC en longitud 8aa, cuya posterior verificación experimental, condujo a la identificación de uno de ellos (`CE1: KWKLFKKI`) como PASAC (véase las Secciones [3.1.10] y [3.1.11]), y aunque de baja toxicidad, pasó satisfactoriamente el ensayo de toxicidad hacia células de humano con un cultivo epitelial de células de bebé como control.

Lo anterior resulta relevante debido a que la determinación de este PASAC fué resultado de un análisis matemático–computacional sobre propiedades fisicoquímicas de la materia, lo cual es parte de los objetivos originales de este proyecto doctoral, además de haberse detectado un PASAC corto en una longitud (aa) que no se ajusta al promedio (23aa), correspondiente a los péptidos antibacterianos (véase la Sección [1].

El presente reporte da cuenta de dos pruebas estadísticas que acreditaron la no distribución Normal entre los PASAC y los péptidos antibacterianos, por lo que esta

aproximación desestimó alguna caracterización al respecto.

La primera prueba tuvo que ver con verificar los 20^8 péptidos de longitud 8aa construibles con respecto a todos los candidatos PASAC en esa longitud con la prueba no paramétrica Ji-cuadrada. Ello permitió observar que los candidatos PASAC no guardan una distribución Normal con respecto al espacio total de péptidos en esa longitud (véase la Sección [3.1.19]).

La segunda prueba se realizó con la prueba estadística no paramétrica Wilcoxon, Mann y Whitney (véase la Sección [3.1.20]), sobre los siguientes conjuntos de péptidos extraidos de la base de datos [105] a septiembre, 2007. La mísma corroboró la no distribución Normal entre ambos conjuntos:

> Los péptidos antibacterianos los cuales actuan exclusivamente contra bacterias y cuya estructura tridimensional es conocida.
>
> Los péptidos antibacterianos que no presentan acción específica contra algún patógeno y que no importó que método fue usado para detectar su estructura tridimensional.

Lo anterior permitió observar que las propiedades fisicoquímicas no guardan una distribución Normal, y en consecuencia no puede inferirse a los candidatos PASAC a partir de una muestra representativa de las variaciones de los posibles péptidos a determinada longitud. Se considera este resultado relevante para el entendimiento de los aminoácidos que emplea la naturaleza para la fabricación de PASAC.

Continuando la búsqueda de un patrón de PASAC se usó el programa APAP para examinar dos péptidos conocidos por su acción antibacteriana: `magainina 2` y `cecropina A`. Del amino-terminal al carboxilo-terminal se fueron extrayendo todos los péptidos hallados en longitud 8aa, luego en longitud 9aa y así sucesivamente hasta su longitud original.

Al ser todos estos péptidos evaluados por APAP, condujeron a la detección de dos regiones antibacterianas: `magainina 2(3-14)` y `cecropina A(1-8)`. Ésta última ya habia sido reportada como región antibacteriana [51, 73, 91] y la primera no mostró actividad antibacteriana. Nuévamente, es importante mencionar que estas regiones se obtuvieron por análisis matemático-computacional.

Pero las pruebas de detección y caracterización de PASAC no sólo ocuparon el aspecto matemático–computacional, casi desde el primer tercio del desarrollo de este proyecto doctoral se investigó el algoritmo modelos ocultos de Markov (MOM) [11, 23, 82, 87] (véase las Secciones [3.1.17 y 3.1.18]), el cual se usa en forma intensiva en Bioinformática [16], particularmente en la caracterización de familias de proteínas [62, 102] e identificación de genes.

Su objetivo es describir la dinámica de fenómenos que no actuan en un espacio de tiempo continuo (aunque existen esfuerzos en ese sentido [90]), sino como un conjun-

to de fotografías que permiten conjeturar la tendencia sin observar la transformación precisa del objeto en estudio, y se caracteriza porque no requiere del pasado total del objeto a evaluarse, sino su historia inmediata anterior, maximizando las cualidades del perfil buscado.

MOM se adecuó para su búsqueda sobre el espacio total de candidatos PASAC en longitud 8aa, hallando sólo uno (`CE1: KWKLFKKI`) (véase el Cuadro [3.9]). Posteriormente sobre los péptidos reportados por Del Rio G. [30] se detectaron candidatos PASAC contenidos principalmente en los péptidos antibacterianos `magainina 2` y `cecropina A`.

Dado estos resultados, se pensó en construir un MOM basado principalmente en propiedades fisicoquímicas [82], cuyo perfil de búsqueda fueran los péptidos antibacterianos localizados en la base de datos APD [105] a septiembre, 2007. Este ensayo produjo la detección de un cluster de 57 péptidos antibacterianos (véase el Cuadro [3.11]), que actuan exclusivamente contra bacterias, hongos, virus y células cancerígenas de mamíferos y cuya estructura tridimensional fué detectada por espectroscopía NRM y rayos X.

Entre éstos se hallaron 9 PASAC, 6 péptidos antibacterianos híbridos formados por `cecropina A` y `magainina 2`, 19 péptidos pertenecientes a la familia `Cecropina A`, 4 péptidos de la familia `Brevinina`, 3 péptidos de la familia `Catelina` y 2 péptidos de la familia `Moricina`.

Particularmente sólo la familia Catelina exhibe un número sobresaliente de cisteínas, aminoácido representativo en las secuencias predichas por MOM [53] y ninguna de esas familias muestra un número sobresaliente de triptofanos (es un aminoácido esencial ya que ayuda a regular los niveles adecuados de serotonina en el cerebro), aminoácido representativo en las secuencias predichas por QSAR (Quantitative structure-activity relationship, por sus siglas en ingles. Es un proceso mediante el cual una estructura química se identifica con un proceso biológico o químico) [35].

La relevancia de la detección de los péptidos referidos a través de MOM radica en la adecuación del algoritmo MOM para usar como probabilidad condicional la evaluación de las propiedades fisicoquímicas del péptido.

Finalmente, un trabajo que se caracterizó por un esfuerzo conjunto de diferentes colaboradores del equipo de investigación del Dr. Del Rio G. (véase la Sección [2.3.20]), llevó a observar que las propiedades fisicoquímicas son un elemento realmente discriminativo para distinguir PASAC, ya que se pudo observar que las propiedades fisicoquímicas: Punto Isoeléctrico y Momento hidrofóbico, en los intervalos referidos en este artículo, se preservan para PASAC de manera independiente a los aminoácidos que constituyen el péptido.

Ello fue resultado, particularmente del estudio de la preservación de la actividad antibacteriana de dos familias: `Cecropina` y `Magainina`.

Quedan sin embargo, muchas preguntas sin responderse. Una en particular que el

autor de esta tesis piensa que es viable responder en el corto plazo, es llevar el análisis de PASAC al nivel de la interacción entre aminoácidos, esto implica la recreación en el espacio tridimensional de cada PASAC, y a revisar particularmente la interacción actividad-antibacteriana-selectiva y distancia-de-interacción-entre-los-elementos-constitutivos-de-los-aminoácidos.

Para lograr la simulación automatizada de la interacción actividad-antibacteriana-selectiva y distancia-de-interacción-entre-los-elementos-constitutivos-de-los-aminoácidos, se cuenta con el hardware existente en diversas instituciones de investigación en México, de manera tal que se constituya un grid (el cual consiste de múltiples clusters interconectados que actúan como una única computadora) [12], de supercomputadoras que apoye el cómputo paralelo que es necesario para tal fin. Asimismo se cuenta con el conocimiento y experiencia adquirido en este proyecto en el área de microprogramación con FPGAs [6, 19] (véase la Sección [2.1.5]), con objeto de miniaturizar el sistema resultante de manera que ello derive en un sistema instalable, por su tamaño en diferentes laboratorios de investigación.

Capítulo 5

Conclusiones

En resumen, hemos presentado evidencias de que las propiedades fisicoquímicas de los PASAC pueden ser determinantes en la actividad de estos péptidos. Para alcanzar esta conclusión, hemos mostrado que es posible predecir PASAC basados únicamente en propiedades fisicoquímicas bajo dos vias de búsqueda:

- Una exhaustiva, a través de cómputo de alto rendimiento y microprogramación en FPGA.
- Una eminentemente matemática, basada en los modelos ocultos de Markov.

En ambos casos extrayendo los péptidos y proteinas de bases de datos públicas, reforzando así la conjetura de que el cómputo intensivo es una herramienta eficaz en la predicción de estructuras biológicas.

El aceleramiento logrado entre la versión original de APAP, hasta su versión mejorada APAP-II ($278,985$ veces), resultado de equivalentes bioquímicos y técnicas de programación paralela y secuencial, no cuenta con referente alguno en la literatura especializada de la cual, quien escribe este reporte, pueda dar cuenta.

Como último comentario quiero mostrar la relevancia del cómputo intensivo añadiendo que este proyecto que aquí se concluye, no hubiera sido posible sin el auxilio de este tipo de procesamiento masivo de información el cual permitió cuantificar no sólo que porcentaje de candidatos PASAC se hallaron bajo los criterios fisicoquímicos descritos, sino encontrar tambien cúmulos de éstos dentro de los intervalos de cada uno de estos criterios, y diseñar el algoritmo matemático computacional (referido como modelo oculto de Markov), que ocupó la primera publicación indexada resultado de esta investigación.

Bibliografía

[1] Alpha data. http://www.alpha-data.com/amd-xrc-4.html.

[2] Antimicrobial peptides laboratory. department of biochemistry, biophysics and macromolecular chemistry. university of trieste. copyright 2005 antimicrobial peptides laboratory. http://www.bbcm.units.it/~tossi/antimic.html.

[3] Assembler language. http://ulita.ms.mff.cuni.cz/pub/predn/swi\\119/AssemblerIBM_390.pdf.

[4] Dell dell dimension 4600. http://reviews.cnet.com/desktops/dell-dimension-4600-pentium/4505-3118_7-30529709.html?tag=contentBody;compare.

[5] Departamento de bioquímica. instituto de fisiología celular (ifc). 2008. unam. http://bionetics.ifc.unam.mx/ganglia/.

[6] Departamento de ciencias computacionales. instituto nacional de ástrofísica óptica y electrónica (inaoe). 2008. conacyt. http://www.inaoep.mx/.

[7] Departamento de cómputo. instituto de biotecnología (ibt). 2008. unam. http://cluster.ibt.unam.mx/ganglia/.

[8] Departamento de supercómputo. dirección general de cómputo académico. 2008. unam. http://www.super.unam.mx/index.php?op=e\-qhw.

[9] Expasy proteomics server (expasy). 2008. http://www.expasy.ch/sprot.

[10] Hard drive (hdd)-random access memory (ram). http://www.macfriends.com/what-is-the-difference-between-memory-ram-hard-drive-and-fl\ash-memory.aspx.

[11] Hidden markov models. a tutorial for the course computational intelligence. signal processing and speech communication laboratory. http://www.igi.tugraz.at/lehre/CI/tutorials/HMM/HMM.pdf.

[12] Intel software network. http://software.intel.com/en-us/articles/grid-supercomputer-demonstrates-intel-itaniumr-2-processor-prowess/.

[13] Linux: Gnu/linux operating system. 2008. free software foundation, inc. http://www.gnu.org/.

[14] Microsoft: Windows corporation. 2008. http://www.microsoft.com/windows//.

[15] National center for biotechnology information. u.s. national library of medicine. 8600 rockville pike, bethesda, md 20894. http://www.ncbi.nlm.nih.gov/.

[16] Ubc bioinformatics centre. http://bioinformatics.ubc.ca/resources/tools/hmmer.

[17] Uniprot swiss-prot. ftp://ftp.expasy.org/databases/swiss-prot/release_compressed/<>uniprot_sprot.fasta.gz.

[18] University of nebraska medical center. copyright 2008. http://www.unmc.edu/.

[19] Xilinx inc. copyright 2008. http://www.xilinx.com/.

[20] I A Bogdán, J Rivers, R J Beynon, and D Coca. High-performance hardware implementation of a parallel database search engine for real-time peptide mass fingerprinting. Bioinformatics., 24(13):1498–502, jul 2008.

[21] E Boix and M V Nogueés. Mammalian antimicrobial proteins and peptides: overview on the rnase a superfamily members involved in innate host defence. Mol. Biosyst., 3(5):317–335, may 2007.

[22] N Boulanger, R Brun, L Ehret-Sabatier, C Kunz, and Bulet. Immunopeptides in the defense reactions of glossina morsitans to bacterial and trypanosoma brucei brucei infections. Insect Biochem Mol Biol., 32(4):369–75, apr 2002.

[23] H Bourlard, Krstulović, and M Magimai-Doss. Introduction to Hidden Markov Models. Ecole Polytechnique Fédérale de Lausanne, Switzerlan, 2001.

[24] K A Brogden. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? Nat Rev Microbiol., 3(3):238–50, mar 2005.

[25] J M Carlton, S V Angiuoli, B B Suh, T W Kooij, M Pertea, J C Silva, M D Ermolaeva, J E Allen, J D Selengut, H L Koo, J D Peterson, M Pop, D S Kosack, M F Shumway, S L Bidwell, S J Shallom, S E van Aken, S B Riedmuller, T V Feldblyum, J K Cho, J Quackenbush, M Sedegah, A Shoaibi, L M Cummings, L Florens, J R Yates, J D Raine, R E Sinden, M A Harris, D A Cunningham, P R Preiser, L W Bergman, A B Vaidya, L H van Lin, C J Janse, A P Waters, H O Smith, O R White, S L Salzberg, J C Venter, C M Fraser, S L Hoffman, M J Gardner, and D J Carucci. Genome sequence and

comparative analysis of the model rodent malaria parasite plasmodium yoelii yoelii. Nature, 419(6906):512–9, oct 2002.

[26] D P Clark, S Durell, W L Maloy, and M Zasloff. Ranalexin. a novel antimicrobial peptide from bullfrog (rana catesbeiana) skin, structurally related to the bacterial antibiotic, polymyxin. J Biol Chem., 269(14):10849–55, apr 1994.

[27] R Conde, F Z Zamudio, M H Rodríguez, and L D Possani. Scorpine, an anti-malaria and anti-bacterial agent purified from scorpion venom. FEBS Lett., 471(2-3):165–168, apr 2000.

[28] Y S Dandass, S C Burgess, M Lawrence, and S M Bridges. Accelerating string set matching in fpga hardware for bioinformatics research. BMC Bioinformatics, 9:179, apr 2008.

[29] R M Dawson and C Q Liu. Properties and applications of antimicrobial peptides in biodefense against biological warfare threat agents. Crit Rev Microbiol., 34(2):89–107, 2008.

[30] G Del Rio, S Castro-Obregon, R Rao, M H Ellerby, and D E Bredesen. Apap, a sequence-pattern recognition approach identifies substance p as a potential apoptotic peptide. FEBS Lett., 494(3):213,219, apr 2001.

[31] B Deslouches, S M Phadke, V Lazarevic, M Cascio, K Islam, R C Montelaro, and T A Mietzner. De novo generation of cationic antimicrobial peptides: influence of length and tryptophan substitution on antimicrobial activity. Antimicrob Agents Chemother., 49(1):316–22, jan 2005.

[32] D Eisenberg, R M Weiss, and T C Terwilliger. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature, 23(299):371–4, sep 1982.

[33] D Eisenberg, R M Weiss, and T C Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. U.S.A., 81(1):140–144, jan 1984.

[34] H Mi Ellerby, W Arap, M L Ellerby, R Kain, R Andrusiak, G Del Rio, S Krajewski, R C Lombardo, R Rao, E Ruoslahti, E D Bredesen, and R Pasqualini. Anti-cancer activity of targeted pro-apoptotic peptides. Nature Medicine, 5(9):1032–8, sep 1999.

[35] C D Fjell, H Jenssen, Fries P, P Aich, P Griebel, K Hilpert, R E Hancock, and A Cherkasov. Identification of novel host defense peptides and the absence of alpha-defensins in the bovine genome. Proteins, 73(2):420–430, 2008.

[36] R M Gajardo and P G Sánchez. Péptidos antimicrobianos y su participación en la defensa contra infecciones bacterianas. Universidad de Concepción campus Chillán, 9(17), jan 2006.

[37] D M Gerlag, E Borges, P P Tak, H M Ellerby, D E Bredesen, R Pasqualini, E Ruoslahti, and G S Firestein. Suppression of murine collagen-induced arthritis by targeted apoptosis of synovial neovasculature. Arthritis Res, 3(6):357–61, sep 2001.

[38] J Gesell, M Zasloff, and S J Opella. Two-dimensional 1h nmr experiments show that the 23-residue magainin antibiotic peptide is an alpha-helix in dodecylphosphocholine micelles, sodium dodecylsulfate micelles, and trifluoroethanol/water solution. J Biomol NMR, 9(2):127–35, feb 1997.

[39] J Goraya, Y Wang, Z Li, M O'Flaherty, F C Knoop, J E Platz, J E, and J M Conlon. Peptides with antimicrobial activity from four different families isolated from the skins of the north american frogs rana luteiventris, rana berlandieri and rana pipiens. Eur J Biochem., 267(3):894–900, feb 2000.

[40] G H Gudmundsson, D A Lidholm, B Asling, R Gan, and H G Boman. The cecropin locus. cloning and expression of a gene cluster encoding three antibacterial peptides in hyalophora cecropia. J Biol Chem, 266(18):11510–11517, jun 1991.

[41] J D Hale and R E Hancock. Alternative mechanisms of action of cationic antimicrobial peptides on bacteria. Expert Rev Anti Infect Ther., 5(6):951–9, dec 2007.

[42] T Halverson, Y J Basir, F C Knoop, and J M Conlon. Purification and characterization of antimicrobial peptides from the skin of the north american green frog rana clamitans. Peptides., 21(4):469–76, apr 2000.

[43] H Hemmi, J Ishibashi, S Hara, and M Yamakawa. Solution structure of moricin, an antibacterial peptide, isolated from the silkworm bombyx mori. FEBS Lett., 518(1-3):33–8, may 2002.

[44] S T Henriques, M Melo, and M A Castanho. Cell-penetrating peptides and antimicrobial peptides: how different are they? Biochem J., 399(1):1–7, oct 2006.

[45] M W Hirsch and Smale S. Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, Inc., New York, USA, 1974.

[46] R A Hoskins, J W Carlson, C Kennedy, D Acevedo, M Evans-Holm, E Frise, K H Wan, S Park, M Mendez-Lago, F Rossi, A Villasante, P Dimitri, G H Karpen, and S E Celniker. Sequence finishing and mapping of drosophila melanogaster heterochromatin. Science, 316(5831):1625–8, jun 2007.

[47] H W Huang. Molecular mechanism of antimicrobial peptides: the origin of cooperativity. Biochim Biophys Acta., 1758(9):1292–302, sep 2006.

[48] C D Hyde. CSCI 320 Computer Architecture, Handbook on Verilog HDL. Computer Science Department, Bucknell University, Lewisburg, PA 17837, 1995.

[49] A Izadpanah. Antimicrobial peptides. Nature, 52(3 Pt 1):381–90; quiz 391–2, mar 2005.

[50] H Jenssen, C D Fjell, A Cherkasov, , and Hancock R E. Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward p. aeruginosa. Chem Biol Drug Des., 70(2):134–142, aug 2007.

[51] F Jin, X Xu, L Wang, W Zhanq, and D Gu. Expression of recombinant hybrid peptide cecropin(1–8)–magainin 2(1–12) in pichia pastoris: purification and characterization. Protein Expr Purif., 50(2):147–156, dic 2006.

[52] A Kanai and S Satori. Cloning of gene cluster for sarcotoxin i, antibacterial proteins of sarcophaga peregrina. FEBS Lett., 258(2):199–202, dec 1989.

[53] C Y Kao, Y Chen, Y H Zhao, and R Wu. Orfeome-based search of airway epithelial cell-specific novel human [beta]-defensin genes. Am J Respir Cell Mol Biol., 1:71–80, jul 2003.

[54] B W Kernighan and Ritchie D M. The C Programming Language. Prentice-Hall, Englewood Cliffs, New Jersey,USA, 1978.

[55] M G Kolonin, P K Saha, L Chan, R Pasqualini, and W Arap. Reversal of obesity by targeted ablation of adipose tissue. Nat Med., 10(6):625–32, 2004.

[56] E Kreyszig. Introductory Mathematical Statistics: Principles and Methods. John Wiley & Sons, Inc., New York, USA, 1970.

[57] E Lacroix, A R Viguera, and L Serrano. Elucidating the folding problem of -helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of nmr parameters. Journal of Molecular Biology, 284(1):173–191, nov 1997.

[58] Y H Lam, V Guyen, E Fakaris, and F Separovic. Conformational studies of a melittin-inhibitor complex. J Protein Chem., 6(2):529–34, aug 2000.

[59] Y H Lam, V Guyen, E Fakaris, and F Separovic. Analysis of an optimal hidden markov model for secondary structure prediction. BMC Struct Biol., 6:25, dec 2006.

[60] A Lamaziére, F Burlina, C Wolf, G Chassaing, G Trugnan, and J Ayala-Sanmartini. Non-metabolic membrane tubulation and permeability induced by bioactive peptides. PLoS One., 2(2):e201, feb 2007.

[61] R Linzmeier, D Michaelson, L Liu, and T Ganz. The structure of neutrophil defensin genes. FEBS Lett., 321(2-3):267–73, 1993.

[62] M Madera. Profile comparer: a program for scoring and aligning profile hidden markov models. Bioinformatics., 24(22):2630–1, nov 2008.

[63] M M Mahoney, A Y Lee, D J Brezinski-Caliguri, and K M Huttner. Molecular analysis of the sheep cathelin family reveals a novel antimicrobial peptide. FEBS Lett., 377(3):519–22, dec 1995.

[64] P Mak, K Wójcik, I B Thogersen, and A Dubin. Isolation, antimicrobial activities, and primary structures of hamster neutrophil defensins. Infect Immun., 64(11):4444–9, nov 1996.

[65] M Maniezzo and A Sanna. A Deep Evaluation of Design Issues and Performances of PVM, MPICH and mpiGAMMA Libraries. Dipartimento di Informatica ed Automatica Politecnico di Torino, Italy, Torino, Italy, 2005.

[66] L Marenah, P R Flatt, D F Orr, C Shaw, and Y H Abdel-Wahab. Skin secretions of rana saharica frogs reveal antimicrobial peptides esculentins-1 and -1b and brevinins-1e and -2ec with novel insulin releasing activity. J Endocrinol., 188(1):1–9, jan 2006.

[67] B Matutte, K B Storey, F C Knoop, and Conlon J M. Induction of synthesis of an antimicrobial peptide in the skin of the freeze-tolerant frog, rana sylvatica, in response to environmental stimuli. FEBS Lett., 483(2-3):135–8, oct 2000.

[68] E Mátyus, C Kandt, and D P Tieleman. Computer simulation of antimicrobial peptides. Curr Med Chem., 14(26):2789–98, 2007.

[69] S Maya, R Reynoso, C Torres, and M Arias-Estrada. Compact spiking neural network implementation in fpga. lecture notes in computer science. In Heidelberg, editor, Field-Programmable Logic and Applications. The Roadmap to Reconfigurable Computing: 10th International Conference, FPL 2000, Villach, Austria, pages 27–30. Springer Berlin, 2000.

[70] L Moerman, S Bosteels, W Oppe, J Willems, E Clynen, L Schoofs, K Thevissen, J Tytgat, J Van Eldere, J Van Der Walt, and F Verdonck. Antibacterial and antifungal properties of alpha-helical, cationic peptides in the venom of scorpions from southern africa. Eur J Biochem., 269(19):4799–4810, oct 2002.

[71] K S Moore, C L Bevins, M M Brasseur, N Tomassini, K Turner, H Eck, and M Zasloff. Antimicrobial peptides in the stomach of xenopus laevis. J Biol Chem., 266(29):19851–7, oct 1991.

[72] N Morikawa, K Hagiwara, and T Akajima. Brevinin-1 and -2, unique antimicrobial peptides from the skin of the frog, rana brevipoda porsa. Biochem Biophys Res Commun., 189(1):184–90, nov 1992.

[73] D Oh, S Y Shin, J H Kang, K S Hahm, K L Kim, and Y Kim. Nmr structural characterization of cecropin a(1-8) - magainin 2(1-12) and cecropin a (1-8) - melittin (1-12) hybrid peptides. J. Pept. Res., 53(5):578–589, may 1999.

[74] J Orivel, V Redeker, J P Le-Caer, F Krier, A M Revol-Junelles, A Longeon, A Chaffotte, A Dejean, and J Rossier. Ponericins, new antibacterial and insecticidal peptides from the venom of the ant pachycondyla goeldii. J Biol Chem., 276(21):17823–9, may 2001.

[75] J L Ortega-Arjona and G Roberts. Architectural patterns for parallel programming. In Proceedings of the 3rd European Conference on Pattern Languages of Programming and Computing (EuroPLoP'98). Kloster Irsee, Germany, 1998.

[76] G C Page. Professional Programmer's Guide to Fortran 77. University of Leicester, United Kingdom, 2005.

[77] F C Pardo. VHDL, Lenguaje para descripción y modelado de circuitos. Departamento de Ingenieria Informática. Universidad de Valencia, Valencia, España, 1997.

[78] J M Park, J E Jung, and B J Lee. Antimicrobial peptides from the skin of a korean frog, rana rugosa. Biochem Biophys Res Commun., 205(1):948–54, nov 1995.

[79] A Parmakelis, M A Slotman, J C Marshall, P H Awono-Ambene, C Antonio-Nkondjio, F Simard, A Caccone, and J R Powell. Evidence for myxobacterial origin of eukaryotic defensins. Immunogenetics., 59(12):949–54, dec 2007.

[80] A Parmakelis, M A Slotman, J C Marshall, P H Awono-Ambene, C Antonio-Nkondjio, F Simard, A Caccone, and J R Powell. The molecular evolution of four anti-malarial immune genes in the anopheles gambiae species complex. BMC Evol. Biol., 8:79, 2008.

[81] C Polanco. Análisis comparativo de diversos generadores de números aleatorios. Departamento de Matemáticas, Facultad de Ciencias, UNAM, México, 2004.

[82] C Polanco and J L Samaniego. Detection of selective cationic amphipatic antibacterial peptides by hidden markov models. Acta Biochimica Polonica, 56(1):167–176, jan 2009.

[83] A E Popsueva, M V Zinovjeva, MV, J W Visser, J M Zijlmans, W E Fibbe, and A V Belyavsky. A novel murine cathelin-like protein expressed in bone marrow. FEBS Lett., 391(1-2):5–8, aug 1996.

[84] K J Preacher. Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [computer software]. `http://www.psych.ku.edu/preacher/chisq/chisq.htm`.

[85] Z Qu, H Steiner, A Engström, H Bennich, and H G Boman. Sequence and specificity of two antibacterial proteins involved in insect immunity. Nature, 292(5820):246–8, jul 1981.

[86] Z Qu, H Steiner, A Engström, H Bennich, and H G Boman. Insect immunity: isolation and structure of cecropins b and d from pupae of the chinese oak silk moth, antheraea pernyi. Eur J Biochem., 127(1):219–24, sep 1982.

[87] L R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of the IEEE. USA, 1989.

[88] K V Reddy, R D Yedery, and C Aranha. Antimicrobial peptides: premises and promises. Int J Antimicrob Agents., 25(5):448–9, may 2005.

[89] M Rosetto, A G Manetti, D Marchini, R Dallai, J L Telford, and C T Baldari. Sequences of two cdna clones from the medfly ceratitis capitata encoding antibacterial peptides of the cecropin family. Gene, 134(2):241–3, dec 1993.

[90] Y Shi, M Klustein, I Simon, T Mitchell, and Z Bar-Joseph. Continuous hidden process model for time series expression experiments. Bioinformatics., 23(13):459–67, jul 2007.

[91] S Y Shin, J H Kang, S Y Janq, Y Kim, K L Kim, and K S Kahm. Effects of the hinge region of cecropin a(1–8)–magainin 2(1–12), a synthetic antimicrobial peptide, on liposomes, bacterial and tumor cells. Biochin Biophys Acta, 1463:209–18, feb 2000.

[92] M Simmaco, G Mignogna, S Canofeni, R Miele, M L Mangoni, and D Barra. Temporins, antimicrobial peptides from the european red frog rana temporaria. Eur J Biochem., 242(3):788–92, dec 1996.

[93] J Söding. Protein homology detection by hmm-hmm comparison. Bioinformatics., 1-21(7):951–960, apr 2005.

[94] M R Spiegel. Theory and problems of Probability and Statistics. Schaum's outline series, New York, USA, 1975.

[95] R L Strausberg, E A Feingold, L H Grouse, J G Derge, F S Klausner, R D Collins, L Wagner, C M Shenmen, G D Schuler, S F Altschul, B Zeeberg, K H Buetow, C F Schaefer, N K Bhat, R F Hopkins, H Jordan, T Moore, S I Max, J Wang, F Hsieh, L Diatchenko, K Marusina, A A Farmer, G M Rubin, L Hong, M Stapleton, M B Soares, M F Bonaldo, T L Casavant, T E Scheetz, M J Brownstein, T B Usdin, S Toshiyuki, P Carninci, C Prange, S S Raha, N A Loquellano, G J Peters, R D Abramson, S J Mullahy, S A Bosak, P J McEwan, K J McKernan, P H Malek, J A Gunaratne, S Richards, K C Worley, S Hale, A M Garcia, L J Gay, S W Hulyk, D K Villalon, D M Muzny, E J Sodergren, X Lu, R A Gibbs, J Fahey, E Helton, M Ketteman, A Madan, S Rodrigues, A Sanchez, M Whiting, A Madan, A C Young, Y Shevchenko, G G Bouffard, R W Blakesley, J W Touchman, E D Green, M C Dickson, A C Rodriguez, J Grimwood, J Schmutz, R M Myers, Y S Butterfield, M I Krzywinski, U Skalska, D E Smailus, A Schnerch, J E Schein, S J Jones, and M A Marra. Generation and initial analysis of more than 15,000 full-length human and mouse cdna sequences. Proc Natl Acad Sci U S A., 99(26):16899–903, dec 2002.

[96] D Sun, Eccleston E D, and Fallon A M. Cloning and expression of three cecropin cdnas from a mosquito cell line. FEBS Lett., 454(1-2):147–51, jul 1999.

[97] K Taniai, K Kadono-Okuda, Y Kato, M Yamamoto, M Shimabukuro, S Chowdhury, J Xu, E Kotani, S Tomino, and M Yamakawa. Structure of two cecropin b-encoding genes and bacteria-inducible dna-binding proteins which bind to the 5'-upstream regulatory region in the silkworm, bombyx mori. Gene, 163(2):215–9, oct 1995.

[98] J A Tennessen and M S Blouin. Selection for antimicrobial peptide diversity in frogs leads to gene duplication and low allelic variation. J Mol Evol., 65(5):605–15, nov 2007.

[99] A Torres-Larios, G B Gurrola, F Z Zamudio, and L D Possani. Hadrurin, a new antimicrobial peptide from the venom of the scorpion hadrurus aztecus. Eur J Biochem., 267(16):5023–31, aug 2000.

[100] A Tossi, M Scocchi, M Zanetti, P Storici, and R Gennaro. Pmap-37, a novel antibacterial peptide from pig myeloid cells. cdna cloning, chemical synthesis and activity. Eur J Biochem., 228(3):941–6, mar 1995.

[101] S Troeira, N Melo, and A Castanho. Cell-penetrating peptides and antimicrobial peptides: how different are they? Biochem., 399:1–7, feb 2006.

[102] I Tsigelny, Y Sharikov, and L F Teni-Eyck. Hidden markov models-based system (hmmspectr) for detecting structural homologies on the basis of sequential information. Protein Eng, 15(5):347–52, may 2002.

[103] V Uversky. What does it mean to be natively unfolded? Eur. J. Biochem., 269(1):2–12, jan 2002.

[104] V Uversky, J R Gillespie, and A L Fink. Why are "natively unfolded"proteins unstructured under physiologic conditions? Proteins, 41(3):415–27, nov 2000.

[105] Z Wang and G Wang. Apd: the antimicrobial peptide database. Nucleic Acids Research, 32:590–592, sep 2004.

[106] K J Won, T Hamelryck, A Prügel-Bennett, and A Krogh. An evolutionary method for learning hmm structure: prediction of protein secondary structure. BMC Bioinformatics., 21(8):357, sep 2007.

[107] M Zasloff. Antimicrobial peptides of multicellular organisms. Nature, 415:389–395, jan 2002.

[108] W L Zhu, H Lan, Y Park, S T Yang, J I Kim, I S Park, H J You, J S Lee, Y S Park, Y Kim, K S Hahm, and S Y Shin. Effects of pro → peptoid residue substitution on cell selectivity and mechanism of antibacterial action of tritrpticin-amide antimicrobial peptide. Biochemistry., 45(43):13007–17, jan 2006.

[109] A V Zimin, D R Smith, G Sutton, and J A Yorke. Assembly reconciliation. Bioinformatics, 1(24):142–5, jan 2008.

Apéndice A

Candidato no PASAC para prueba negativa.

L	Tipo	Péptido	MH	PI	C	CU	AG
8		AVVGQATQ	0.36	6.05	0.00	0.26	0.04

L: longitud del péptido, Tipo: [*,/], Péptido: péptido evaluado, MH: momento hidrofóbico $[0.4, 0.6]$ (véase la Sección 2.2.1), PI: punto isoeléctrico $[10.8, 11.8]$ (véase la Sección 2.2.1), C: carga neta promedio $[-0.2, 0.5]$ (véase la Ecuación 2.1), CU: algoritmo modificado de Uversky $[-0.2, 0.5]$ (véase la Ecuación 2.2), AG: AGADIR $[0.0, 10.0]$ (véase la Sección 2.2.1). APAP debe cumplir: MH, PI y AG. Los aceptados estan marcados con [*]. APAP–I debe cumplir: MH, PI y C > CU. Los aceptados estan marcados con [/]. Los péptidos que están calificados con [*] y [/] y están presentes en todas las longitudes están sombreados en color ■. Los péptidos seleccionados para prueba negativa están sombreados en color ■. Los péptidos que están calificados con [*] y [/] y sólo están presentes en las primeras longitudes están sombreados en color ■.

Apéndice B

Candidatos PASAC hallados en `magainina 2`.

L	Tipo	Péptido	MH	PI	C	CU	AG
8		GIGKFLHS	0.60	9.65	0.12	0.25	0.14
		IGKFLHSA	0.56	9.65	0.12	0.29	0.29
	/*	GKFLHSAK	0.42	10.80	0.25	0.11	0.07
	/*	KFLHSAKK	0.44	11.10	0.37	-0.00	0.15
	/*	FLHSAKKF	0.56	10.80	0.25	-0.04	0.19
		LHSAKKFG	0.39	10.80	0.25	-0.12	0.36
	/*	HSAKKFGK	0.40	11.10	0.37	-0.20	0.54
		SAKKFGKA	0.34	11.10	0.37	-0.13	0.75
		AKKFGKAF	0.62	11.10	0.37	-0.05	0.29
		KKFGKAFV	0.63	11.10	0.37	0.04	0.04
		KFGKAFVG	0.65	10.80	0.25	0.28	0.11
		FGKAFVGE	0.57	6.45	0.00	0.53	0.32
		GKAFVGEI	0.55	6.45	0.00	0.70	0.05
		KAFVGEIM	0.66	6.45	0.00	0.80	0.06
		AFVGEIMN	0.73	3.95	-0.12	0.57	0.10
		FVGEIMNS	0.70	3.95	-0.12	0.37	0.20
9		GIGKFLHSA	0.58	9.65	0.11	0.30	0.18
	/*	IGKFLHSAK	0.49	10.80	0.22	0.21	0.42
	/*	GKFLHSAKK	0.43	11.10	0.33	0.00	0.16
	/*	KFLHSAKKF	0.50	11.10	0.33	-0.02	0.24
	/*	FLHSAKKFG	0.48	10.80	.22	-0.04	0.26
		LHSAKKFGK	0.39	11.10	0.33	-0.16	0.79
		HSAKKFGKA	0.37	11.10	0.33	-0.17	0.51
	/*	SAKKFGKAF	0.48	11.10	0.33	-0.06	0.73
		AKKFGKAFV	0.62	11.10	0.33	0.01	0.30
		KKFGKAFVG	0.64	11.10	0.33	0.10	0.07
		KFGKAFVGE	0.61	9.50	0.11	0.34	0.12
		FGKAFVGEI	0.56	6.45	0.00	0.63	0.29
		GKAFVGEIM	0.60	6.45	0.00	0.75	0.06
		KAFVGEIMN	0.69	6.45	0.00	0.57	0.11
		AFVGEIMNS	0.71	3.95	-0.11	0.41	0.20
10		GIGKFLHSAK	0.53	10.80	0.20	0.23	0.25
	/*	IGKFLHSAKK	0.47	11.10	0.30	0.08	0.75
	/*	GKFLHSAKKF	0.47	11.10	0.30	-0.01	0.25
	/*	KFLHSAKKFG	0.47	11.10	0.30	-0.03	0.29
	/*	FLHSAKKFGK	0.45	11.10	0.30	-0.09	0.62
		LHSAKKFGKA	0.38	11.10	0.30	-0.14	0.74
	/*	HSAKKFGKAF	0.45	11.10	0.30	-0.10	0.50
	/*	SAKKFGKAFV	0.53	11.10	0.30	-0.00	0.72
		AKKFGKAFVG	0.63	11.10	0.30	0.07	0.31
		KKFGKAFVGE	0.62	10.50	0.20	0.16	0.07
		KFGKAFVGEI	0.59	9.50	0.10	0.43	0.11
		FGKAFVGEIM	0.59	6.45	0.00	0.68	0.30
		GKAFVGEIMN	0.65	6.45	0.00	0.57	0.10

(continuación)

Cuadro B.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		KAFVGEIMNS	0.69	6.45	0.00	0.44	0.18
11	/*	GIGKFLHSAKK	0.50	11.10	0.27	0.11	0.46
	/*	IGKFLHSAKKF	0.50	11.10	0.27	0.05	1.03
	/*	GKFLHSAKKFG	0.45	11.10	0.27	-0.01	0.29
	/*	KFLHSAKKFGK	0.45	11.30	0.36	-0.07	0.61
	/*	FLHSAKKFGKA	0.42	11.10	0.27	-0.08	0.59
	/*	LHSAKKFGKAF	0.44	11.10	0.27	-0.09	0.74
	/*	HSAKKFGKAFV	0.50	11.10	0.27	-0.05	0.50
	/*	SAKKFGKAFVG	0.56	11.10	0.27	0.04	0.71
		AKKFGKAFVGE	0.62	10.50	0.18	0.12	0.30
		KKFGKAFVGEI	0.60	10.50	0.18	0.24	0.06
		KFGKAFVGEIM	0.61	9.50	0.09	0.49	0.11
		FGKAFVGEIMN	0.63	6.45	0.00	0.54	0.35
		GKAFVGEIMNS	0.66	6.45	0.00	0.45	0.16
12	/*	GIGKFLHSAKKF	0.52	11.10	0.25	0.08	0.64
	/*	IGKFLHSAKKFG	0.47	11.10	0.25	0.03	1.05
	/*	GKFLHSAKKFGK	0.44	11.30	0.33	-0.05	0.59
	/*	KFLHSAKKFGKA	0.43	11.30	0.33	-0.07	0.59
	/*	FLHSAKKFGKAF	0.46	11.10	0.25	-0.04	0.60
	/*	LHSAKKFGKAFV	0.48	11.10	0.25	-0.04	0.75
	/*	HSAKKFGKAFVG	0.53	11.10	0.25	-0.00	0.52
		SAKKFGKAFVGE	0.56	10.50	0.16	0.09	0.71
		AKKFGKAFVGEI	0.60	10.50	0.16	0.19	0.31
		KKFGKAFVGEIM	0.61	10.50	0.16	0.30	0.07
		KFGKAFVGEIMN	0.63	9.50	0.08	0.41	0.15
		FGKAFVGEIMNS	0.64	6.45	0.00	0.44	0.44
13	/*	GIGKFLHSAKKFG	0.50	11.10	0.23	0.06	0.67
	/*	IGKFLHSAKKFGK	0.46	11.30	0.30	-0.00	1.46
	/*	GKFLHSAKKFGKA	0.43	11.30	0.30	-0.05	0.57
	/*	KFLHSAKKFGKAF	0.46	11.30	0.30	-0.03	0.60
	/*	FLHSAKKFGKAFV	0.49	11.10	0.23	-0.00	0.61
	/*	LHSAKKFGKAFVG	0.50	11.10	0.23	-0.00	0.76
		HSAKKFGKAFVGE	0.53	10.50	0.15	0.03	0.51
		SAKKFGKAFVGEI	0.56	10.50	0.15	0.14	0.74
		AKKFGKAFVGEIM	0.61	10.50	0.15	0.24	0.30
		KKFGKAFVGEIMN	0.63	10.50	0.15	0.26	0.10
		KFGKAFVGEIMNS	0.64	9.50	0.07	0.35	0.20
14	/*	GIGKFLHSAKKFGK	0.48	11.30	0.28	0.01	1.00
	/*	IGKFLHSAKKFGKA	0.44	11.30	0.28	-0.01	1.41
	/*	GKFLHSAKKFGKAF	0.45	11.30	0.28	-0.02	0.58
	/*	KFLHSAKKFGKAFV	0.48	11.30	0.28	-0.00	0.61
	/*	FLHSAKKFGKAFVG	0.51	11.10	0.21	0.02	0.62
		LHSAKKFGKAFVGE	0.51	10.50	0.14	0.02	0.75
		HSAKKFGKAFVGEI	0.54	10.50	0.14	0.08	0.53
		SAKKFGKAFVGEIM	0.57	10.50	0.14	0.19	0.70
		AKKFGKAFVGEIMN	0.63	10.50	0.14	0.21	0.31
		KKFGKAFVGEIMNS	0.64	10.50	0.14	0.23	0.14
15	/*	GIGKFLHSAKKFGKA	0.46	11.30	0.26	0.00	0.98
	/*	IGKFLHSAKKFGKAF	0.47	11.30	0.26	0.01	1.41
	/*	GKFLHSAKKFGKAFV	0.48	11.30	0.26	0.00	0.60
	/*	KFLHSAKKFGKAFVG	0.50	11.30	0.26	0.02	0.62
		FLHSAKKFGKAFVGE	0.52	10.50	0.13	0.05	0.62
		LHSAKKFGKAFVGEI	0.52	10.50	0.13	0.07	0.77
		HSAKKFGKAFVGEIM	0.55	10.50	0.13	0.12	0.51
		SAKKFGKAFVGEIMN	0.59	10.50	0.13	0.18	0.68
		AKKFGKAFVGEIMNS	0.64	10.50	0.13	0.19	0.34
16	/*	GIGKFLHSAKKFGKAF	0.48	11.30	0.25	0.03	0.99
	/*	IGKFLHSAKKFGKAFV	0.48	11.30	0.25	0.03	1.44
	/*	GKFLHSAKKFGKAFVG	0.49	11.30	0.25	0.02	0.61
	/*	KFLHSAKKFGKAFVGE	0.51	10.80	0.18	0.05	0.62
		FLHSAKKFGKAFVGEI	0.52	10.50	0.12	0.09	0.64
		LHSAKKFGKAFVGEIM	0.53	10.50	0.12	0.11	0.73

(continuación)

Cuadro B.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		HSAKKFGKAFVGEIMN	0.57	10.50	0.12	0.11	0.50
		SAKKFGKAFVGEIMNS	0.61	10.50	0.12	0.16	0.68
17	/*	GIGKFLHSAKKFGKAFV	0.50	11.30	0.23	0.05	1.01
	/*	IGKFLHSAKKFGKAFVG	0.50	11.30	0.23	0.06	1.44
	/*	GKFLHSAKKFGKAFVGE	0.50	10.80	0.17	0.05	0.61
	/*	KFLHSAKKFGKAFVGEI	0.52	10.80	0.17	0.09	0.64
		FLHSAKKFGKAFVGEIM	0.54	10.50	0.11	0.13	0.61
		LHSAKKFGKAFVGEIMN	0.55	10.50	0.11	0.10	0.71
		HSAKKFGKAFVGEIMNS	0.58	10.50	0.11	0.11	0.52
18	/*	GIGKFLHSAKKFGKAFVG	0.51	11.30	0.22	0.07	1.02
	/*	IGKFLHSAKKFGKAFVGE	0.51	10.80	0.16	0.08	1.41
	/*	GKFLHSAKKFGKAFVGEI	0.51	10.80	0.16	0.09	0.63
	/*	KFLHSAKKFGKAFVGEIM	0.53	10.80	0.16	0.12	0.61
		FLHSAKKFGKAFVGEIMN	0.55	10.50	0.11	0.12	0.60
		LHSAKKFGKAFVGEIMNS	0.57	10.50	0.11	0.10	0.71
19	/*	GIGKFLHSAKKFGKAFVGE	0.52	10.80	0.15	0.10	1.01
	/*	IGKFLHSAKKFGKAFVGEI	0.51	10.80	0.15	0.12	1.40
	/*	GKFLHSAKKFGKAFVGEIM	0.52	10.80	0.15	0.12	0.60
	/*	KFLHSAKKFGKAFVGEIMN	0.55	10.80	0.15	0.12	0.60
		FLHSAKKFGKAFVGEIMNS	0.57	10.50	0.10	0.12	0.61
20	/*	GIGKFLHSAKKFGKAFVGEI	0.52	10.80	0.15	0.13	1.01
	/*	IGKFLHSAKKFGKAFVGEIM	0.52	10.80	0.15	0.14	1.34
	/*	GKFLHSAKKFGKAFVGEIMN	0.54	10.80	0.15	0.11	0.59
	/*	KFLHSAKKFGKAFVGEIMNS	0.56	10.80	0.15	0.11	0.61
21	*	GIGKFLHSAKKFGKAFVGEIM	0.53	10.80	0.14	0.15	0.97
	*	IGKFLHSAKKFGKAFVGEIMN	0.54	10.80	0.14	0.14	1.29
	*	GKFLHSAKKFGKAFVGEIMNS	0.55	10.80	0.14	0.11	0.60
22	*	GIGKFLHSAKKFGKAFVGEIMN	0.54	10.80	0.13	0.15	0.95
	*	IGKFLHSAKKFGKAFVGEIMNS	0.55	10.80	0.13	0.13	1.26
23	*	GIGKFLHSAKKFGKAFVGEIMNS	0.55	10.80	0.13	0.14	0.94

L: longitud del péptido, Tipo: [*,/], Péptido: péptido evaluado, MH: momento hidrofóbico $[0.4, 0.6]$ (véase la Sección 2.2.1), PI: punto isoeléctrico $[10.8, 11.8]$ (véase la Sección 2.2.1), C: carga neta promedio $[-0.2, 0.5]$ (véase la Ecuación 2.1), CU: algoritmo modificado de Uversky $[-0.2, 0.5]$ (véase la Ecuación 2.2), AG: AGADIR $[0.0, 10.0]$ (véase la Sección 2.2.1). APAP debe cumplir: MH, PI y AG. Los aceptados estan marcados con [*]. APAP–I debe cumplir: MH, PI y C > CU. Los aceptados estan marcados con [/]. Los péptidos que están calificados con [*] y [/] y están presentes en todas las longitudes están sombreados en color ■. Los péptidos seleccionados para prueba negativa están sombreados en color ■. Los péptidos que están calificados con [*] y [/] y sólo están presentes en las primeras longitudes están sombreados en color ■.

Apéndice C

Candidatos PASAC hallados en `cecropina A`.

L	Tipo	Péptido	MH	PI	C	CU	AG
8	/*	KWKLFKKI	0.49	11.30	0.50	0.031	1.66
		WKLFKKIE	0.62	10.50	0.25	0.00	0.44
		KLFKKIEK	0.69	10.80	0.37	-0.08	0.22
		LFKKIEKV	0.67	10.50	0.25	-0.06	0.86
		FKKIEKVG	0.57	10.50	0.25	-0.08	0.38
		KKIEKVGQ	0.52	10.50	0.25	-0.12	0.15
		KIEKVGQN	0.45	9.50	0.12	-0.08	0.12
		IEKVGQNI	0.63	6.45	0.00	-0.04	0.20
		EKVGQNIR	0.53	9.69	0.12	-0.12	0.10
		KVGQNIRD	0.54	9.69	0.12	-0.16	0.08
		VGQNIRDG	0.52	6.30	0.00	-0.16	0.19
		GQNIRDGI	0.57	6.30	0.00	-0.17	0.13
		QNIRDGII	0.59	6.30	0.00	-0.09	0.14
		NIRDGIIK	0.60	9.69	0.12	0.04	0.10
		IRDGIIKA	0.52	9.69	0.12	0.29	0.11
		RDGIIKAG	0.34	9.69	0.12	0.45	0.28
		DGIIKAGP	0.37	6.30	0.00	0.45	0.63
		GIIKAGPA	0.32	9.69	0.12	0.35	0.13
		IIKAGPAV	0.34	9.69	0.12	0.33	0.17
		IKAGPAVA	0.23	9.69	0.12	0.31	0.00
		KAGPAVAV	0.11	9.69	0.12	0.63	0.01
		AGPAVAVV	0.20	6.05	0.00	1.74	0.05
		GPAVAVVG	0.30	6.05	0.00	2.55	0.09
		PAVAVVGQ	0.36	6.05	0.00	2.61	0.24
		AVAVVGQA	0.26	6.05	0.00	2.17	0.19
		VAVVGQAT	0.28	6.05	0.00	1.03	0.12
		AVVGQATQ	0.36	6.05	0.00	0.26	0.04
		VVGQATQI	0.41	6.05	0.00	0.11	0.05
		VGQATQIA	0.34	6.05	0.00	0.08	0.16
		GQATQIAK	0.28	9.69	0.12	0.03	0.45
9	/*	KWKLFKKIE	0.55	10.80	0.33	0.00	1.32
		WKLFKKIEK	0.65	10.80	0.33	-0.07	0.84
		KLFKKIEKV	0.68	10.80	0.33	-0.07	0.28
		LFKKIEKVG	0.62	10.50	0.22	-0.02	0.95
		FKKIEKVGQ	0.54	10.50	0.22	-0.11	0.37
		KKIEKVGQN	0.48	10.50	0.22	-0.14	0.16
		KIEKVGQNI	0.54	9.50	0.11	-0.03	0.14
		IEKVGQNIR	0.58	9.69	0.11	-0.07	0.23
		EKVGQNIRD	0.54	6.50	0.00	-0.17	0.12
		KVGQNIRDG	0.53	9.69	0.11	-0.17	0.14
		VGQNIRDGI	0.54	6.30	0.00	-0.11	0.19
		GQNIRDGII	0.58	6.30	0.00	-0.11	0.16
		QNIRDGIIK	0.60	9.69	0.11	-0.04	0.22
		NIRDGIIKA	0.56	9.69	0.11	0.13	0.15

(continuación)

Cuadro C.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		IRDGIIKAG	0.43	9.69	0.11	0.38	0.28
		RDGIIKAGP	0.35	9.69	0.11	0.38	0.45
		DGIIKAGPA	0.34	6.30	0.00	0.31	0.72
		GIIKAGPAV	0.33	9.69	0.11	0.42	0.13
		IIKAGPAVA	0.29	9.69	0.11	0.40	0.16
		IKAGPAVAV	0.17	9.69	0.11	0.50	0.01
		KAGPAVAVV	0.15	9.69	0.11	1.07	0.02
		AGPAVAVVG	0.25	6.05	0.00	2.07	0.08
		GPAVAVVGQ	0.33	6.05	0.00	2.11	0.13
		PAVAVVGQA	0.31	6.05	0.00	2.15	0.26
		AVAVVGQAT	0.27	6.05	0.00	1.51	0.20
		VAVVGQATQ	0.32	6.05	0.00	0.54	0.13
		AVVGQATQI	0.38	6.05	0.00	0.20	0.06
		VVGQATQIA	0.37	6.05	0.00	0.17	0.11
		VGQATQIAK	0.31	9.69	0.11	0.06	0.47
10	/*	KWKLFKKIEK	0.60	11.00	0.40	-0.06	1.93
		WKLFKKIEKV	0.66	10.80	0.30	-0.06	0.98
		KLFKKIEKVG	0.64	10.80	0.30	-0.04	0.30
		LFKKIEKVGQ	0.58	10.50	0.20	-0.05	0.93
		FKKIEKVGQN	0.51	10.50	0.20	-0.13	0.38
		KKIEKVGQNI	0.53	10.50	0.20	-0.09	0.17
		KIEKVGQNIR	0.54	10.75	0.20	-0.07	0.15
		IEKVGQNIRD	0.57	6.50	0.00	-0.12	0.24
		EKVGQNIRDG	0.53	6.50	0.00	-0.18	0.18
		KVGQNIRDGI	0.54	9.69	0.10	-0.13	0.14
		VGQNIRDGII	0.56	6.30	0.00	-0.07	0.23
		GQNIRDGIIK	0.59	9.69	0.10	-0.07	0.23
		QNIRDGIIKA	0.57	9.69	0.10	0.03	0.27
		NIRDGIIKAG	0.49	9.69	0.10	0.21	0.31
		IRDGIIKAGP	0.41	9.69	0.10	0.33	0.43
		RDGIIKAGPA	0.34	9.69	0.10	0.28	0.44
		DGIIKAGPAV	0.34	6.30	0.00	0.37	0.73
		GIIKAGPAVA	0.30	9.69	0.10	0.47	0.13
		IIKAGPAVAV	0.23	9.69	0.10	0.55	0.16
		IKAGPAVAVV	0.18	9.69	0.10	0.83	0.02
		KAGPAVAVVG	0.20	9.69	0.10	1.36	0.05
		AGPAVAVVGQ	0.29	6.05	0.00	1.81	0.11
		GPAVAVVGQA	0.31	6.05	0.00	1.84	0.14
		PAVAVVGQAT	0.30	6.05	0.00	1.60	0.27
		AVAVVGQATQ	0.30	6.05	0.00	0.88	0.21
		VAVVGQATQI	0.35	6.05	0.00	0.42	0.14
		AVVGQATQIA	0.37	6.05	0.00	0.24	0.11
		VVGQATQIAK	0.34	9.69	0.10	0.14	0.36
11		KWKLFKKIEKV	0.62	11.00	0.36	-0.05	2.00
		WKLFKKIEKVG	0.64	10.80	0.27	-0.03	0.98
		KLFKKIEKVGQ	0.61	10.80	0.27	-0.06	0.30
		LFKKIEKVGQN	0.55	10.50	0.18	-0.08	0.94
		FKKIEKVGQNI	0.54	10.50	0.18	-0.09	0.39
		KKIEKVGQNIR	0.53	11.05	0.27	-0.11	0.18
		KIEKVGQNIRD	0.54	9.50	0.09	-0.11	0.16
		IEKVGQNIRDG	0.56	6.50	0.00	-0.14	0.30
		EKVGQNIRDGI	0.54	6.50	0.00	-0.14	0.18
		KVGQNIRDGII	0.56	9.69	0.09	-0.09	0.16
		VGQNIRDGIIK	0.57	9.69	0.09	-0.04	0.31
		GQNIRDGIIKA	0.57	9.69	0.09	-0.01	0.28
		QNIRDGIIKAG	0.51	9.69	0.09	0.09	0.44
		NIRDGIIKAGP	0.46	9.69	0.09	0.20	0.47
		IRDGIIKAGPA	0.39	9.69	0.09	0.26	0.43
		RDGIIKAGPAV	0.34	9.69	0.09	0.34	0.44
		DGIIKAGPAVA	0.31	6.30	0.00	0.42	0.71
		GIIKAGPAVAV	0.25	9.69	0.09	0.60	0.13
		IIKAGPAVAVV	0.22	9.69	0.09	0.84	0.16

(continuación)

Cuadro C.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		IKAGPAVAVVG	0.21	9.69	0.09	1.07	0.04
		KAGPAVAVVGQ	0.24	9.69	0.09	1.28	0.07
		AGPAVAVVGQA	0.28	6.05	0.00	1.65	0.13
		GPAVAVVGQAT	0.30	6.05	0.00	1.45	0.15
		PAVAVVGQATQ	0.31	6.05	0.00	1.00	0.27
		AVAVVGQATQI	0.33	6.05	0.00	0.68	0.21
		VAVVGQATQIA	0.34	6.05	0.00	0.42	0.17
		AVVGQATQIAK	0.34	9.69	0.09	0.20	0.34
12		KWKLFKKIEKVG	0.61	11.00	0.33	-0.03	2.00
		WKLFKKIEKVGQ	0.61	10.80	0.25	-0.05	1.00
	/*	KLFKKIEKVGQN	0.58	10.80	0.25	-0.08	0.31
		LFKKIEKVGQNI	0.57	10.50	0.16	-0.05	0.95
		FKKIEKVGQNIR	0.54	11.05	0.25	-0.10	0.40
		KKIEKVGQNIRD	0.54	10.50	0.16	-0.14	0.19
		KIEKVGQNIRDG	0.54	9.50	0.08	-0.12	0.21
		IEKVGQNIRDGI	0.56	6.50	0.00	-0.10	0.31
		EKVGQNIRDGII	0.55	6.50	0.00	-0.11	0.21
		KVGQNIRDGIIK	0.57	10.75	0.16	-0.06	0.22
		VGQNIRDGIIKA	0.56	9.69	0.08	0.01	0.37
		GQNIRDGIIKAG	0.53	9.69	0.08	0.04	0.45
		QNIRDGIIKAGP	0.49	9.69	0.08	0.10	0.62
		NIRDGIIKAGPA	0.43	9.69	0.08	0.17	0.48
		IRDGIIKAGPAV	0.38	9.69	0.08	0.31	0.44
		RDGIIKAGPAVA	0.32	9.69	0.08	0.38	0.44
		DGIIKAGPAVAV	0.27	6.30	0.00	0.53	0.70
		GIIKAGPAVAVV	0.24	9.69	0.08	0.85	0.14
		IIKAGPAVAVVG	0.24	9.69	0.08	1.05	0.20
		IKAGPAVAVVGQ	0.24	9.69	0.08	1.04	0.06
		KAGPAVAVVGQA	0.25	9.69	0.08	1.22	0.08
		AGPAVAVVGQAT	0.28	6.05	0.00	1.35	0.14
		GPAVAVVGQATQ	0.31	6.05	0.00	0.97	0.15
		PAVAVVGQATQI	0.33	6.05	0.00	0.79	0.29
		AVAVVGQATQIA	0.33	6.05	0.00	0.65	0.25
		VAVVGQATQIAK	0.33	9.69	0.08	0.35	0.39
13	/*	KWKLFKKIEKVGQ	0.59	11.00	0.30	-0.05	2.01
	/*	WKLFKKIEKVGQN	0.58	10.80	0.23	-0.07	0.99
	/*	KLFKKIEKVGQNI	0.59	10.80	0.23	-0.06	0.31
		LFKKIEKVGQNIR	0.56	11.05	0.23	-0.07	1.00
		FKKIEKVGQNIRD	0.54	10.50	0.15	-0.13	0.42
		KKIEKVGQNIRDG	0.53	10.50	0.15	-0.15	0.23
		KIEKVGQNIRDGI	0.54	9.50	0.07	-0.09	0.22
		IEKVGQNIRDGII	0.56	6.50	0.00	-0.08	0.33
		EKVGQNIRDGIIK	0.56	9.50	0.07	-0.08	0.27
		KVGQNIRDGIIKA	0.56	10.75	0.15	-0.01	0.26
		VGQNIRDGIIKAG	0.52	9.69	0.07	0.06	0.54
		GQNIRDGIIKAGP	0.50	9.69	0.07	0.05	0.63
		QNIRDGIIKAGPA	0.46	9.69	0.07	0.08	0.62
		NIRDGIIKAGPAV	0.41	9.69	0.07	0.21	0.49
		IRDGIIKAGPAVA	0.35	9.69	0.07	0.35	0.44
		RDGIIKAGPAVAV	0.28	9.69	0.07	0.48	0.44
		DGIIKAGPAVAVV	0.26	6.30	0.00	0.75	0.69
		GIIKAGPAVAVVG	0.25	9.69	0.07	1.03	0.17
		IIKAGPAVAVVGQ	0.26	9.69	0.07	1.03	0.20
		IKAGPAVAVVGQA	0.24	9.69	0.07	1.02	0.07
		KAGPAVAVVGQAT	0.25	9.69	0.07	1.05	0.08
		AGPAVAVVGQATQ	0.29	6.05	0.00	0.95	0.14
		GPAVAVVGQATQI	0.33	6.05	0.00	0.79	0.17
		PAVAVVGQATQIA	0.33	6.00	0.00	0.75	0.32
		AVAVVGQATQIAK	0.32	9.69	0.07	0.53	0.45
14	/*	KWKLFKKIEKVGQN	0.57	11.00	0.28	-0.06	1.94
	/*	WKLFKKIEKVGQNI	0.59	10.80	0.21	-0.05	0.98
	/*	KLFKKIEKVGQNIR	0.58	11.25	0.28	-0.07	0.33

(continuación)

Cuadro C.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		LFKKIEKVGQNIRD	0.56	10.50	0.14	-0.10	1.01
		FKKIEKVGQNIRDG	0.54	10.50	0.14	-0.14	0.44
		KKIEKVGQNIRDGI	0.54	10.50	0.14	-0.12	0.23
		KIEKVGQNIRDGII	0.55	9.50	0.07	-0.07	0.23
		IEKVGQNIRDGIIK	0.57	9.50	0.07	-0.06	0.38
		EKVGQNIRDGIIKA	0.55	9.50	0.07	-0.03	0.31
		KVGQNIRDGIIKAG	0.53	10.75	0.14	0.02	0.39
		VGQNIRDGIIKAGP	0.50	9.69	0.07	0.06	0.73
		GQNIRDGIIKAGPA	0.47	9.69	0.07	0.04	0.62
		QNIRDGIIKAGPAV	0.44	9.69	0.07	0.12	0.62
		NIRDGIIKAGPAVA	0.39	9.69	0.07	0.25	0.48
		IRDGIIKAGPAVAV	0.32	9.69	0.07	0.44	0.44
		RDGIIKAGPAVAVV	0.27	9.69	0.07	0.66	0.44
		DGIIKAGPAVAVVG	0.27	6.30	0.00	0.91	0.76
		GIIKAGPAVAVVGQ	0.27	9.69	0.07	1.02	0.17
		IIKAGPAVAVVGQA	0.26	9.69	0.07	1.01	0.20
		IKAGPAVAVVGQAT	0.25	9.69	0.07	0.90	0.08
		KAGPAVAVVGQATQ	0.27	9.69	0.07	0.77	0.09
		AGPAVAVVGQATQI	0.31	6.05	0.00	0.79	0.16
		GPAVAVVGQATQIA	0.33	6.05	0.00	0.75	0.20
		PAVAVVGQATQIAK	0.33	9.69	0.07	0.62	0.50
15	/*	KWKLFKKIEKVGQNI	0.58	11.00	0.26	-0.05	1.88
	/*	WKLFKKIEKVGQNIR	0.58	11.25	0.26	-0.07	1.01
	/*	KLFKKIEKVGQNIRD	0.58	10.80	0.20	-0.10	0.34
		LFKKIEKVGQNIRDG	0.55	10.50	0.13	-0.11	1.01
		FKKIEKVGQNIRDGI	0.54	10.50	0.13	-0.12	0.44
		KKIEKVGQNIRDGII	0.54	10.50	0.13	-0.10	0.24
		KIEKVGQNIRDGIIK	0.55	10.50	0.13	-0.05	0.27
		IEKVGQNIRDGIIKA	0.56	9.50	0.06	-0.02	0.41
		EKVGQNIRDGIIKAG	0.53	9.50	0.06	0.00	0.43
		KVGQNIRDGIIKAGP	0.51	10.75	0.13	0.03	0.53
		VGQNIRDGIIKAGPA	0.48	9.69	0.06	0.06	0.73
		GQNIRDGIIKAGPAV	0.46	9.69	0.06	0.08	0.63
		QNIRDGIIKAGPAVA	0.41	9.69	0.06	0.16	0.62
		NIRDGIIKAGPAVAV	0.35	9.69	0.06	0.32	0.49
		IRDGIIKAGPAVAVV	0.30	9.69	0.06	0.59	0.44
		RDGIIKAGPAVAVVG	0.28	9.69	0.06	0.80	0.48
		DGIIKAGPAVAVVGQ	0.28	6.30	0.00	0.90	0.72
		GIIKAGPAVAVVGQA	0.27	9.69	0.06	1.00	0.17
		IIKAGPAVAVVGQAT	0.26	9.69	0.06	0.90	0.20
		IKAGPAVAVVGQATQ	0.26	9.69	0.06	0.69	0.08
		KAGPAVAVVGQATQI	0.28	9.69	0.06	0.66	0.10
		AGPAVAVVGQATQIA	0.31	6.05	0.00	0.76	0.19
		GPAVAVVGQATQIAK	0.32	9.69	0.06	0.64	0.37
16	/*	KWKLFKKIEKVGQNIR	0.57	11.40	0.31	-0.06	1.88
	/*	WKLFKKIEKVGQNIRD	0.58	10.80	0.18	-0.09	1.00
	/*	KLFKKIEKVGQNIRDG	0.57	10.80	0.18	-0.11	0.36
		LFKKIEKVGQNIRDGI	0.56	10.50	0.12	-0.09	0.98
		FKKIEKVGQNIRDGII	0.55	10.50	0.12	-0.10	0.44
	/*	KKIEKVGQNIRDGIIK	0.55	10.80	0.18	-0.08	0.28
		KIEKVGQNIRDGIIKA	0.55	10.50	0.12	-0.02	0.30
		IEKVGQNIRDGIIKAG	0.54	9.50	0.06	0.01	0.52
		EKVGQNIRDGIIKAGP	0.51	9.50	0.06	0.01	0.58
		KVGQNIRDGIIKAGPA	0.49	10.75	0.12	0.03	0.54
		VGQNIRDGIIKAGPAV	0.46	9.69	0.06	0.09	0.74
		GQNIRDGIIKAGPAVA	0.43	9.69	0.06	0.11	0.63
		QNIRDGIIKAGPAVAV	0.38	9.69	0.06	0.22	0.62
		NIRDGIIKAGPAVAVV	0.34	9.69	0.06	0.45	0.49
		IRDGIIKAGPAVAVVG	0.30	9.69	0.06	0.72	0.49
		RDGIIKAGPAVAVVGQ	0.29	9.69	0.06	0.81	0.47
		DGIIKAGPAVAVVGQA	0.28	6.30	0.00	0.90	0.69
		GIIKAGPAVAVVGQAT	0.27	9.69	0.06	0.91	0.17

(continuación)

Cuadro C.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		IIKAGPAVAVVGQATQ	0.27	9.69	0.06	0.70	0.19
		IKAGPAVAVVGQATQI	0.28	9.69	0.06	0.60	0.10
		KAGPAVAVVGQATQIA	0.29	9.69	0.06	0.64	0.13
		AGPAVAVVGQATQIAK	0.31	9.69	0.06	0.65	0.34
17	/*	KWKLFKKIEKVGQNIRD	0.57	11.00	0.23	-0.08	1.84
	/*	WKLFKKIEKVGQNIRDG	0.57	10.80	0.17	-0.10	0.99
	/*	KLFKKIEKVGQNIRDGI	0.57	10.80	0.17	-0.09	0.36
	/*	LFKKIEKVGQNIRDGII	0.56	10.50	0.11	-0.07	0.95
	/*	FKKIEKVGQNIRDGIIK	0.55	10.80	0.17	-0.08	0.46
	/*	KKIEKVGQNIRDGIIKA	0.55	10.80	0.17	-0.05	0.31
		KIEKVGQNIRDGIIKAG	0.53	10.50	0.11	0.01	0.41
		IEKVGQNIRDGIIKAGP	0.52	9.50	0.05	0.02	0.65
		EKVGQNIRDGIIKAGPA	0.49	9.50	0.05	0.01	0.58
		KVGQNIRDGIIKAGPAV	0.47	10.75	0.11	0.06	0.55
		VGQNIRDGIIKAGPAVA	0.44	9.69	0.05	0.12	0.73
		GQNIRDGIIKAGPAVAV	0.40	9.69	0.05	0.17	0.64
		QNIRDGIIKAGPAVAVV	0.36	9.69	0.05	0.33	0.62
		NIRDGIIKAGPAVAVVG	0.33	9.69	0.05	0.56	0.53
		IRDGIIKAGPAVAVVGQ	0.31	9.69	0.05	0.73	0.47
		RDGIIKAGPAVAVVGQA	0.28	9.69	0.0	0.81	0.45
		DGIIKAGPAVAVVGQAT	0.28	6.30	0.00	0.82	0.65
		GIIKAGPAVAVVGQATQ	0.28	9.69	0.05	0.72	0.17
		IIKAGPAVAVVGQATQI	0.28	9.69	0.05	0.62	0.20
		IKAGPAVAVVGQATQIA	0.28	9.69	0.05	0.59	0.12
		KAGPAVAVVGQATQIAK	0.29	10.75	0.11	0.56	0.28
18	/*	KWKLFKKIEKVGQNIRDG	0.57	11.00	0.22	-0.09	1.78
	/*	WKLFKKIEKVGQNIRDGI	0.57	10.80	0.16	-0.09	0.96
	/*	KLFKKIEKVGQNIRDGII	0.57	10.80	0.16	-0.08	0.36
	/*	LFKKIEKVGQNIRDGIIK	0.56	10.80	0.16	-0.06	0.95
	/*	FKKIEKVGQNIRDGIIKA	0.55	10.80	0.16	-0.05	0.47
	/*	KKIEKVGQNIRDGIIKAG	0.53	10.80	0.16	-0.01	0.40
		KIEKVGQNIRDGIIKAGP	0.52	10.50	0.11	0.01	0.53
		IEKVGQNIRDGIIKAGPA	0.50	9.50	0.05	0.02	0.65
		EKVGQNIRDGIIKAGPAV	0.48	9.50	0.05	0.03	0.59
		KVGQNIRDGIIKAGPAVA	0.45	10.75	0.11	0.09	0.55
		VGQNIRDGIIKAGPAVAV	0.41	9.69	0.05	0.17	0.74
		GQNIRDGIIKAGPAVAVV	0.38	9.69	0.05	0.26	0.63
		QNIRDGIIKAGPAVAVVG	0.36	9.69	0.05	0.42	0.66
		NIRDGIIKAGPAVAVVGQ	0.34	9.69	0.05	0.58	0.52
		IRDGIIKAGPAVAVVGQA	0.30	9.69	0.05	0.74	0.46
		RDGIIKAGPAVAVVGQAT	0.28	9.69	0.05	0.75	0.43
		DGIIKAGPAVAVVGQATQ	0.28	6.30	0.00	0.67	0.62
		GIIKAGPAVAVVGQATQI	0.29	9.69	0.05	0.64	0.18
		IIKAGPAVAVVGQATQIA	0.29	9.69	0.05	0.61	0.22
		IKAGPAVAVVGQATQIAK	0.28	10.75	0.11	0.52	0.26
19	/*	KWKLFKKIEKVGQNIRDGI	0.57	11.00	0.21	-0.08	1.71
	/*	WKLFKKIEKVGQNIRDGII	0.57	10.80	0.15	-0.07	0.93
	/*	KLFKKIEKVGQNIRDGIIK	0.57	11.00	0.21	-0.06	0.38
	/*	LFKKIEKVGQNIRDGIIKA	0.56	10.80	0.15	-0.03	0.94
	/*	FKKIEKVGQNIRDGIIKAG	0.53	10.80	0.15	-0.02	0.55
	/*	KKIEKVGQNIRDGIIKAGP	0.52	10.80	0.15	-0.00	0.51
		KIEKVGQNIRDGIIKAGPA	0.50	10.50	0.10	0.01	0.53
		IEKVGQNIRDGIIKAGPAV	0.49	9.50	0.05	0.04	0.65
		EKVGQNIRDGIIKAGPAVA	0.46	9.50	0.05	0.06	0.59
		KVGQNIRDGIIKAGPAVAV	0.42	10.75	0.10	0.13	0.55
		VGQNIRDGIIKAGPAVAVV	0.39	9.69	0.05	0.25	0.73
		GQNIRDGIIKAGPAVAVVG	0.37	9.69	0.05	0.34	0.67
		QNIRDGIIKAGPAVAVVGQ	0.36	9.69	0.05	0.44	0.64
		NIRDGIIKAGPAVAVVGQA	0.33	9.69	0.05	0.60	0.50
		IRDGIIKAGPAVAVVGQAT	0.30	9.69	0.05	0.69	0.44
		RDGIIKAGPAVAVVGQATQ	0.29	9.69	0.05	0.62	0.42
		DGIIKAGPAVAVVGQATQI	0.29	6.30	0.00	0.60	0.60

(continuación)

Cuadro C.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		GIIKAGPAVAVVGQATQIA	0.29	9.69	0.05	0.63	0.20
		IIKAGPAVAVVGQATQIAK	0.29	10.75	0.10	0.54	0.34
20	/*	KWKLFKKIEKVGQNIRDGII	0.57	11.00	0.20	-0.07	1.65
	/*	WKLFKKIEKVGQNIRDGIIK	0.58	11.00	0.20	-0.06	0.93
	/*	KLFKKIEKVGQNIRDGIIKA	0.57	11.00	0.20	-0.03	0.40
	/*	LFKKIEKVGQNIRDGIIKAG	0.54	10.80	0.15	-0.00	0.99
	/*	FKKIEKVGQNIRDGIIKAGP	0.52	10.80	0.15	-0.01	0.65
	/*	KKIEKVGQNIRDGIIKAGPA	0.50	10.80	0.15	-0.00	0.52
		KIEKVGQNIRDGIIKAGPAV	0.49	10.50	0.10	0.04	0.54
		IEKVGQNIRDGIIKAGPAVA	0.47	9.50	0.05	0.07	0.65
		EKVGQNIRDGIIKAGPAVAV	0.43	9.50	0.05	0.10	0.59
		KVGQNIRDGIIKAGPAVAVV	0.40	10.75	0.10	0.21	0.55
		VGQNIRDGIIKAGPAVAVVG	0.39	9.69	0.05	0.33	0.77
		GQNIRDGIIKAGPAVAVVGQ	0.37	9.69	0.05	0.36	0.65
		QNIRDGIIKAGPAVAVVGQA	0.35	9.69	0.05	0.46	0.61
		NIRDGIIKAGPAVAVVGQAT	0.33	9.69	0.05	0.57	0.48
		IRDGIIKAGPAVAVVGQATQ	0.31	9.69	0.05	0.58	0.42
		RDGIIKAGPAVAVVGQATQI	0.30	9.69	0.05	0.56	0.41
		DGIIKAGPAVAVVGQATQIA	0.30	6.30	0.00	0.59	0.60
		GIIKAGPAVAVVGQATQIAK	0.29	10.75	0.10	0.56	0.32
21	/*	KWKLFKKIEKVGQNIRDGIIK	0.57	11.10	0.23	-0.05	1.61
	/*	WKLFKKIEKVGQNIRDGIIKA	0.57	11.00	0.19	-0.03	0.92
	/*	KLFKKIEKVGQNIRDGIIKAG	0.55	11.00	0.19	-0.01	0.48
	/*	LFKKIEKVGQNIRDGIIKAGP	0.53	10.80	0.14	0.00	1.06
	/*	FKKIEKVGQNIRDGIIKAGPA	0.51	10.80	0.14	-0.01	0.65
	/*	KKIEKVGQNIRDGIIKAGPAV	0.49	10.80	0.14	0.01	0.53
		KIEKVGQNIRDGIIKAGPAVA	0.47	10.50	0.09	0.06	0.54
		IEKVGQNIRDGIIKAGPAVAV	0.44	9.50	0.04	0.11	0.65
		EKVGQNIRDGIIKAGPAVAVV	0.41	9.50	0.04	0.17	0.59
		KVGQNIRDGIIKAGPAVAVVG	0.40	10.75	0.09	0.27	0.59
		VGQNIRDGIIKAGPAVAVVGQ	0.38	9.69	0.04	0.35	0.74
		GQNIRDGIIKAGPAVAVVGQA	0.37	9.69	0.04	0.38	0.63
		QNIRDGIIKAGPAVAVVGQAT	0.34	9.69	0.04	0.45	0.59
		NIRDGIIKAGPAVAVVGQATQ	0.33	9.69	0.04	0.48	0.46
		IRDGIIKAGPAVAVVGQATQI	0.31	9.69	0.04	0.53	0.42
		RDGIIKAGPAVAVVGQATQIA	0.30	9.69	0.04	0.55	0.41
22	/*	KWKLFKKIEKVGQNIRDGIIKA	0.57	11.10	0.22	-0.03	1.56
	/*	WKLFKKIEKVGQNIRDGIIKAG	0.56	11.00	0.18	-0.01	0.96
	/*	KLFKKIEKVGQNIRDGIIKAGP	0.54	11.00	0.18	-0.00	0.57
	/*	LFKKIEKVGQNIRDGIIKAGPA	0.52	10.80	0.13	0.00	1.03
	/*	FKKIEKVGQNIRDGIIKAGPAV	0.49	10.80	0.13	0.00	0.65
	/*	KKIEKVGQNIRDGIIKAGPAVA	0.47	10.80	0.13	0.03	0.53
		KIEKVGQNIRDGIIKAGPAVAV	0.44	10.50	0.09	0.10	0.55
		IEKVGQNIRDGIIKAGPAVAVV	0.43	9.50	0.04	0.17	0.65
		EKVGQNIRDGIIKAGPAVAVVG	0.41	9.50	0.04	0.23	0.63
		KVGQNIRDGIIKAGPAVAVVGQ	0.39	10.75	0.09	0.30	0.57
		VGQNIRDGIIKAGPAVAVVGQA	0.38	9.69	0.04	0.37	0.72
		GQNIRDGIIKAGPAVAVVGQAT	0.36	9.69	0.04	0.37	0.60
		QNIRDGIIKAGPAVAVVGQATQ	0.35	9.69	0.04	0.38	0.57
		NIRDGIIKAGPAVAVVGQATQI	0.33	9.69	0.04	0.44	0.45
		IRDGIIKAGPAVAVVGQATQIA	0.32	9.69	0.04	0.52	0.42
		RDGIIKAGPAVAVVGQATQIAK	0.30	10.75	0.09	0.51	0.51
23	/*	KWKLFKKIEKVGQNIRDGIIKAG	0.55	11.10	0.21	-0.01	1.57
	/*	WKLFKKIEKVGQNIRDGIIKAGP	0.55	11.00	0.17	-0.00	1.03
	/*	KLFKKIEKVGQNIRDGIIKAGPA	0.53	11.00	0.17	-0.00	0.57
	/*	LFKKIEKVGQNIRDGIIKAGPAV	0.50	10.80	0.13	0.02	1.02
	/*	FKKIEKVGQNIRDGIIKAGPAVA	0.48	10.80	0.13	0.02	0.65
	/*	KKIEKVGQNIRDGIIKAGPAVAV	0.45	10.80	0.13	0.06	0.53
		KIEKVGQNIRDGIIKAGPAVAVV	0.43	10.50	0.08	0.16	0.55
		IEKVGQNIRDGIIKAGPAVAVVG	0.42	9.50	0.04	0.22	0.68
		EKVGQNIRDGIIKAGPAVAVVGQ	0.40	9.50	0.04	0.25	0.61
		KVGQNIRDGIIKAGPAVAVVGQA	0.39	10.75	0.08	0.32	0.56

(continuación)

Cuadro C.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		VGQNIRDGIIKAGPAVAVVGQAT	0.37	9.69	0.04	0.36	0.69
		GQNIRDGIIKAGPAVAVVGQATQ	0.36	9.69	0.04	0.32	0.58
		QNIRDGIIKAGPAVAVVGQATQI	0.35	9.69	0.04	0.35	0.56
		NIRDGIIKAGPAVAVVGQATQIA	0.33	9.69	0.04	0.44	0.46
		IRDGIIKAGPAVAVVGQATQIAK	0.31	10.75	0.08	0.48	0.51
24	/*	KWKLFKKIEKVGQNIRDGIIKAGP	0.54	11.10	0.2	-0.00	1.60
	/*	WKLFKKIEKVGQNIRDGIIKAGPA	0.53	11.00	0.16	-0.00	1.01
	/*	KLFKKIEKVGQNIRDGIIKAGPAV	0.52	11.00	0.16	0.01	0.57
	/*	LFKKIEKVGQNIRDGIIKAGPAVA	0.49	10.80	0.12	0.04	1.00
	/*	FKKIEKVGQNIRDGIIKAGPAVAV	0.46	10.80	0.12	0.06	0.65
	/*	KKIEKVGQNIRDGIIKAGPAVAVV	0.43	10.80	0.12	0.12	0.53
		KIEKVGQNIRDGIIKAGPAVAVVG	0.42	10.50	0.08	0.21	0.58
		IEKVGQNIRDGIIKAGPAVAVVGQ	0.42	9.50	0.04	0.25	0.66
		EKVGQNIRDGIIKAGPAVAVVGQA	0.39	9.50	0.04	0.27	0.59
		KVGQNIRDGIIKAGPAVAVVGQAT	0.38	10.75	0.08	0.31	0.54
		VGQNIRDGIIKAGPAVAVVGQATQ	0.37	9.69	0.04	0.31	0.67
		GQNIRDGIIKAGPAVAVVGQATQI	0.36	9.69	0.04	0.30	0.57
		QNIRDGIIKAGPAVAVVGQATQIA	0.35	9.69	0.04	0.36	0.55
		NIRDGIIKAGPAVAVVGQATQIAK	0.33	10.75	0.08	0.41	0.54
25	/*	KWKLFKKIEKVGQNIRDGIIKAGPA	0.53	11.10	0.20	-0.00	1.56
	/*	WKLFKKIEKVGQNIRDGIIKAGPAV	0.52	11.00	0.16	0.01	0.99
	/*	KLFKKIEKVGQNIRDGIIKAGPAVA	0.50	11.00	0.16	0.03	0.57
	/*	LFKKIEKVGQNIRDGIIKAGPAVAV	0.47	10.80	0.12	0.07	0.98
	/*	FKKIEKVGQNIRDGIIKAGPAVAVV	0.44	10.80	0.12	0.10	0.64
	/*	KKIEKVGQNIRDGIIKAGPAVAVVG	0.43	10.80	0.12	0.16	0.57
		KIEKVGQNIRDGIIKAGPAVAVVGQ	0.42	10.50	0.08	0.23	0.57
		IEKVGQNIRDGIIKAGPAVAVVGQA	0.41	9.50	0.04	0.26	0.64
		EKVGQNIRDGIIKAGPAVAVVGQAT	0.39	9.50	0.04	0.27	0.57
		KVGQNIRDGIIKAGPAVAVVGQATQ	0.38	10.75	0.08	0.27	0.52
		VGQNIRDGIIKAGPAVAVVGQATQI	0.37	9.69	0.04	0.30	0.65
		GQNIRDGIIKAGPAVAVVGQATQIA	0.36	9.69	0.04	0.31	0.57
		QNIRDGIIKAGPAVAVVGQATQIAK	0.34	10.75	0.08	0.34	0.63
26	/*	KWKLFKKIEKVGQNIRDGIIKAGPAV	0.52	11.10	0.19	0.01	1.52
	/*	WKLFKKIEKVGQNIRDGIIKAGPAVA	0.51	11.00	0.15	0.03	0.97
	/*	KLFKKIEKVGQNIRDGIIKAGPAVAV	0.48	11.00	0.15	0.06	0.58
	/*	LFKKIEKVGQNIRDGIIKAGPAVAVV	0.45	10.80	0.11	0.11	0.97
	*	FKKIEKVGQNIRDGIIKAGPAVAVVG	0.43	10.80	0.11	0.15	0.67
	*	KKIEKVGQNIRDGIIKAGPAVAVVGQ	0.42	10.80	0.11	0.18	0.55
		KIEKVGQNIRDGIIKAGPAVAVVGQA	0.41	10.50	0.07	0.25	0.55
		IEKVGQNIRDGIIKAGPAVAVVGQAT	0.40	9.50	0.03	0.27	0.62
		EKVGQNIRDGIIKAGPAVAVVGQATQ	0.39	9.50	0.03	0.23	0.56
		KVGQNIRDGIIKAGPAVAVVGQATQI	0.38	10.75	0.07	0.26	0.51
		VGQNIRDGIIKAGPAVAVVGQATQIA	0.37	9.69	0.03	0.30	0.65
		GQNIRDGIIKAGPAVAVVGQATQIAK	0.36	10.75	0.07	0.29	0.64
27	*	KWKLFKKIEKVGQNIRDGIIKAGPAVA	0.51	11.10	0.18	0.02	1.48
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAV	0.49	11.00	0.14	0.05	0.96
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVV	0.47	11.00	0.14	0.10	0.57
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVG	0.45	10.80	0.11	0.16	0.98
	*	FKKIEKVGQNIRDGIIKAGPAVAVVGQ	0.43	10.80	0.11	0.17	0.65
	*	KKIEKVGQNIRDGIIKAGPAVAVVGQA	0.42	10.80	0.11	0.20	0.54
		KIEKVGQNIRDGIIKAGPAVAVVGQAT	0.40	10.50	0.07	0.24	0.54
		IEKVGQNIRDGIIKAGPAVAVVGQATQ	0.40	9.50	0.03	0.23	0.60
		EKVGQNIRDGIIKAGPAVAVVGQATQI	0.39	9.50	0.03	0.22	0.55
		KVGQNIRDGIIKAGPAVAVVGQATQIA	0.38	10.75	0.07	0.26	0.51
		VGQNIRDGIIKAGPAVAVVGQATQIAK	0.36	10.75	0.07	0.29	0.71
28	*	KWKLFKKIEKVGQNIRDGIIKAGPAVAV	0.49	11.10	0.17	0.05	1.45
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVV	0.47	11.00	0.14	0.10	0.94
	/*	KLFKKIEKVGQNIRDGIIKAGPAVAVVG	0.46	11.00	0.14	0.14	0.60
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVGQ	0.44	10.80	0.10	0.17	0.95
	*	FKKIEKVGQNIRDGIIKAGPAVAVVGQA	0.42	10.80	0.10	0.18	0.63
	*	KKIEKVGQNIRDGIIKAGPAVAVVGQAT	0.41	10.80	0.10	0.20	0.52
		KIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.40	10.50	0.07	0.22	0.52

(continuación)

Cuadro C.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		IEKVGQNIRDGIIKAGPAVAVVGQATQI	0.40	9.50	0.03	0.22	0.59
		EKVGQNIRDGIIKAGPAVAVVGQATQIA	0.38	9.50	0.03	0.23	0.54
		KVGQNIRDGIIKAGPAVAVVGQATQIAK	0.37	11.05	0.10	0.25	0.58
29	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVV	0.47	11.10	0.17	0.09	1.42
	/*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVG	0.47	11.00	0.13	0.13	0.96
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQ	0.45	11.00	0.13	0.16	0.59
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVGQA	0.43	10.80	0.10	0.19	0.92
	*	FKKIEKVGQNIRDGIIKAGPAVAVVGQAT	0.42	10.80	0.10	0.18	0.62
	*	KKIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.41	10.80	0.10	0.18	0.51
		KIEKVGQNIRDGIIKAGPAVAVVGQATQI	0.40	10.50	0.06	0.21	0.51
		IEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.40	9.50	0.03	0.23	0.59
		EKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.38	10.50	0.06	0.22	0.61
30	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVG	0.47	11.10	0.16	0.13	1.41
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVGQ	0.46	11.00	0.13	0.15	0.93
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQA	0.44	11.00	0.13	0.17	0.57
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVGQAT	0.43	10.80	0.10	0.19	0.90
	*	FKKIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.41	10.80	0.10	0.16	0.60
	*	KKIEKVGQNIRDGIIKAGPAVAVVGQATQI	0.41	10.80	0.10	0.17	0.50
		KIEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.40	10.50	0.06	0.22	0.51
		IEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.39	10.50	0.06	0.22	0.65
31	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQ	0.46	11.10	0.16	0.14	1.37
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVGQA	0.45	11.00	0.12	0.16	0.91
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQAT	0.44	11.00	0.12	0.17	0.56
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.42	10.80	0.09	0.17	0.87
	*	FKKIEKVGQNIRDGIIKAGPAVAVVGQATQI	0.41	10.80	0.09	0.16	0.59
	*	KKIEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.40	10.80	0.09	0.18	0.50
		KIEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.39	10.80	0.09	0.21	0.57
32	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQA	0.45	11.10	0.15	0.16	1.33
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVGQAT	0.45	11.00	0.12	0.17	0.88
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.43	11.00	0.12	0.15	0.55
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.43	11.00	0.12	0.15	0.55
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVGQATQI	0.42	10.80	0.09	0.16	0.85
	*	FKKIEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.41	10.80	0.09	0.17	0.59
	*	KKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.40	11.00	0.12	0.17	0.56
33	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQAT	0.45	11.10	0.15	0.16	1.29
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.44	11.00	0.12	0.15	0.86
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQI	0.43	11.00	0.12	0.15	0.54
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.42	10.80	0.09	0.17	0.84
	*	FKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.41	11.00	0.12	0.16	0.64
34	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQ	0.44	11.10	0.14	0.14	1.26
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQI	0.44	11.00	0.11	0.14	0.84
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.43	11.00	0.11	0.16	0.54
	*	LFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.41	11.00	0.11	0.17	0.88
35	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQI	0.44	11.10	0.14	0.14	1.23
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.44	11.00	0.11	0.15	0.83
	*	KLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.42	11.10	0.14	0.15	0.59
36	*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIA	0.44	11.10	0.13	0.15	1.21
	*	WKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.43	11.10	0.13	0.15	0.87
37	/*	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK	0.43	11.19	0.16	0.14	1.23

L: longitud del péptido, Tipo: [*,/], Péptido: péptido evaluado, MH: momento hidrofóbico $[0.4, 0.6]$ (véase la Sección 2.2.1), PI: punto isoeléctrico $[10.8, 11.8]$ (véase la Sección 2.2.1), C: carga neta promedio $[-0.2, 0.5]$ (véase la Ecuación 2.1), CU: algoritmo modificado de Uversky $[-0.2, 0.5]$ (véase la Ecuación 2.2), AG: AGADIR $[0.0, 10.0]$ (véase la Sección 2.2.1). APAP debe cumplir: MH, PI y AG. Los aceptados estan marcados con [*]. APAP–I debe cumplir: MH, PI y C > CU. Los aceptados estan marcados con [/]. Los péptidos que están calificados con [*] y [/] y están presentes en todas las longitudes están

sombreados en color ■. Los péptidos seleccionados para prueba negativa están sombreados en color ■. Los péptidos que están calificados con [*] y [/] y sólo están presentes en las primeras longitudes están sombreados en color ■.

Apéndice D

Candidatos PASAC hallados en `melitina`.

L	Tipo	Péptido	MH	PI	C	CU	AG
8		GIGAVLKV	0.48	9.65	0.12	1.83	0.41
		IGAVLKVL	0.55	9.65	0.12	2.05	0.99
		GAVLKVLT	0.51	9.65	0.12	1.48	0.25
		AVLKVLTT	0.50	9.65	0.12	1.28	0.27
		VLKVLTTG	0.37	9.65	0.12	1.04	0.19
		LKVLTTGL	0.42	9.65	0.12	0.72	0.11
		KVLTTGLP	0.42	9.65	0.12	0.46	0.03
		VLTTGLPA	0.37	6.09	0.00	0.47	0.02
		LTTGLPAL	0.47	6.09	0.00	0.51	0.01
		TTGLPALI	0.43	6.09	0.00	0.77	0.00
		TGLPALIS	0.53	6.09	0.00	1.16	0.01
		GLPALISW	0.49	6.09	0.00	1.54	0.04
		LPALISWI	0.56	6.09	0.00	1.85	0.09
		PALISWIK	0.54	9.65	0.12	1.18	0.44
	*	ALISWIKR	0.56	11.64	0.25	0.54	0.68
	/*	LISWIKRK	0.45	11.80	0.37	0.12	1.71
		ISWIKRKR	0.30	12.55	0.50	-0.16	1.41
		SWIKRKRQ	0.31	12.55	0.50	-0.47	1.04
		WIKRKRQQ	0.30	12.55	0.50	-0.74	0.90
9		GIGAVLKVL	0.52	9.65	0.11	2.00	0.57
		IGAVLKVLT	0.53	9.65	0.11	1.81	1.21
		GAVLKVLTT	0.50	9.65	0.11	1.15	0.35
		AVLKVLTTG	0.44	9.65	0.11	1.18	0.30
		VLKVLTTGL	0.40	9.65	0.11	0.98	0.22
		LKVLTTGLP	0.42	9.65	0.11	0.56	0.11
		KVLTTGLPA	0.40	9.65	0.11	0.44	0.03
		VLTTGLPAL	0.42	6.09	0.00	0.57	0.02
		LTTGLPALI	0.45	6.09	0.00	0.77	0.01
		TTGLPALIS	0.48	6.09	0.00	0.84	0.01
		TGLPALISW	0.51	6.09	0.00	1.20	0.03
		GLPALISWI	0.52	6.09	0.00	1.68	0.05
		LPALISWIK	0.55	9.65	0.11	1.43	0.22
	*	PALISWIKR	0.55	11.64	0.22	0.64	1.55
	/*	ALISWIKRK	0.51	11.80	0.33	0.24	1.80
		LISWIKRKR	0.37	12.55	0.44	-0.01	2.96
		ISWIKRKRQ	0.30	12.55	0.44	-0.29	1.70
		SWIKRKRQQ	0.31	12.55	0.44	-0.59	1.66
10		GIGAVLKVLT	0.51	9.65	0.10	1.81	0.73
		IGAVLKVLTT	0.52	9.65	0.10	1.44	1.48
		GAVLKVLTTG	0.46	9.65	0.10	1.10	0.38
		AVLKVLTTGL	0.43	9.65	0.10	1.11	0.34
		VLKVLTTGLP	0.40	9.65	0.10	0.77	0.22
		LKVLTTGLPA	0.40	9.65	0.10	0.52	0.11
		KVLTTGLPAL	0.42	9.65	0.10	0.52	0.03

(continuación)

Cuadro D.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		VLTTGLPALI	0.42	6.09	0.00	0.78	0.02
		LTTGLPALIS	0.48	6.09	0.00	0.83	0.02
		TTGLPALISW	0.48	6.09	0.00	0.92	0.03
		TGLPALISWI	0.53	6.09	0.00	1.35	0.05
		GLPALISWIK	0.53	9.65	0.10	1.38	0.14
	*	LPALISWIKR	0.55	11.64	0.20	0.86	0.76
	*	PALISWIKRK	0.52	11.80	0.30	0.34	3.94
		ALISWIKRKR	0.44	12.55	0.40	0.07	3.08
		LISWIKRKRQ	0.35	12.55	0.40	-0.14	3.37
		ISWIKRKRQQ	0.30	12.55	0.40	-0.40	2.47
11		GIGAVLKVLTT	0.51	9.65	0.09	1.48	0.91
		IGAVLKVLTTG	0.48	9.65	0.09	1.34	1.55
		GAVLKVLTTGL	0.45	9.65	0.09	1.05	0.42
		AVLKVLTTGLP	0.43	9.65	0.09	0.89	0.34
		VLKVLTTGLPA	0.40	9.65	0.09	0.69	0.21
		LKVLTTGLPAL	0.42	9.65	0.09	0.59	0.10
		KVLTTGLPALI	0.42	9.65	0.09	0.70	0.03
		VLTTGLPALIS	0.45	6.09	0.00	0.83	0.02
		LTTGLPALISW	0.48	6.09	0.00	0.90	0.04
		TTGLPALISWI	0.50	6.09	0.00	1.06	0.04
		TGLPALISWIK	0.53	9.65	0.09	1.17	0.13
	*	GLPALISWIKR	0.54	11.64	0.18	0.89	0.51
	*	LPALISWIKRK	0.53	11.80	0.27	0.50	1.98
		PALISWIKRKR	0.46	12.55	0.36	0.15	6.17
		ALISWIKRKRQ	0.40	12.55	0.36	-0.05	3.52
		LISWIKRKRQQ	0.34	12.55	0.36	-0.23	4.40
12		GIGAVLKVLTTG	0.48	9.65	0.08	1.39	0.98
		IGAVLKVLTTGL	0.47	9.65	0.08	1.26	1.65
		GAVLKVLTTGLP	0.44	9.65	0.08	0.87	0.43
		AVLKVLTTGLPA	0.42	9.65	0.08	0.80	0.34
		VLKVLTTGLPAL	0.41	9.65	0.08	0.74	0.21
		LKVLTTGLPALI	0.42	9.65	0.08	0.74	0.11
		KVLTTGLPALIS	0.44	9.65	0.08	0.76	0.03
		KVLTTGLPALIS	0.44	9.65	0.08	0.76	0.03
		VLTTGLPALISW	0.46	6.09	0.00	0.89	0.04
		LTTGLPALISWI	0.50	6.09	0.00	1.02	0.05
		TTGLPALISWIK	0.51	9.65	0.08	0.96	0.12
	*	TGLPALISWIKR	0.54	11.64	0.16	0.81	0.47
	*	GLPALISWIKRK	0.52	11.80	0.25	0.56	1.38
		LPALISWIKRKR	0.48	12.55	0.33	0.27	3.31
		PALISWIKRKRQ	0.43	12.55	0.33	0.00	6.91
		ALISWIKRKRQQ	0.38	12.55	0.33	-0.14	4.57
13		GIGAVLKVLTTGL	0.47	9.65	0.07	1.31	1.06
		IGAVLKVLTTGLP	0.46	9.65	0.07	1.05	1.72
		GAVLKVLTTGLPA	0.43	9.65	0.07	0.79	0.43
		AVLKVLTTGLPAL	0.43	9.65	0.07	0.83	0.34
		VLKVLTTGLPALI	0.41	9.65	0.07	0.88	0.22
		LKVLTTGLPALIS	0.44	9.65	0.07	0.79	0.11
		KVLTTGLPALISW	0.45	9.65	0.07	0.81	0.05
		VLTTGLPALISWI	0.47	6.09	0.00	1.00	0.05
		LKVLTTGLPALIS	0.44	9.65	0.07	0.79	0.11
		KVLTTGLPALISW	0.45	9.65	0.07	0.81	0.05
		VLTTGLPALISWI	0.47	6.09	0.00	1.00	0.05
		LTTGLPALISWIK	0.50	9.65	0.07	0.94	0.12
	*	TTGLPALISWIKR	0.52	11.64	0.15	0.70	0.43
	*	TGLPALISWIKRK	0.52	11.80	0.23	0.54	1.27
		GLPALISWIKRKR	0.48	12.55	0.30	0.33	2.40
		LPALISWIKRKRQ	0.45	12.55	0.30	0.09	3.81
		PALISWIKRKRQQ	0.41	12.55	0.30	-0.08	8.52
14		GIGAVLKVLTTGLP	0.46	9.65	0.07	1.11	1.11
		IGAVLKVLTTGLPA	0.45	9.65	0.07	0.95	1.69
		GAVLKVLTTGLPAL	0.44	9.65	0.07	0.82	0.42

(continuación)

Cuadro D.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
		AVLKVLTTGLPALI	0.43	9.65	0.07	0.96	0.35
		VLKVLTTGLPALIS	0.43	9.65	0.07	0.91	0.22
		LKVLTTGLPALISW	0.45	9.65	0.07	0.83	0.12
		KVLTTGLPALISWI	0.47	9.65	0.07	0.91	0.06
		VLTTGLPALISWIK	0.48	9.65	0.07	0.93	0.11
	*	LTTGLPALISWIKR	0.51	11.64	0.14	0.70	0.41
	*	TTGLPALISWIKRK	0.51	11.80	0.21	0.48	1.17
		TGLPALISWIKRKR	0.49	12.55	0.28	0.33	2.22
		GLPALISWIKRKRQ	0.46	12.55	0.28	0.14	2.81
		LPALISWIKRKRQQ	0.43	12.55	0.28	-0.01	4.90
15		GIGAVLKVLTTGLPA	0.45	9.65	0.06	1.02	1.10
		IGAVLKVLTTGLPAL	0.45	9.65	0.06	0.97	1.64
		GAVLKVLTTGLPALI	0.44	9.65	0.06	0.94	0.43
		AVLKVLTTGLPALIS	0.44	9.65	0.06	0.98	0.34
		VLKVLTTGLPALISW	0.44	9.65	0.06	0.95	0.22
		LKVLTTGLPALISWI	0.46	9.65	0.06	0.93	0.12
	*	KVLTTGLPALISWIK	0.48	10.80	0.13	0.86	0.12
	*	VLTTGLPALISWIKR	0.49	11.64	0.13	0.72	0.38
	*	LTTGLPALISWIKRK	0.50	11.80	0.20	0.51	1.11
		TTGLPALISWIKRKR	0.48	12.55	0.26	0.31	2.07
		TGLPALISWIKRKRQ	0.47	12.55	0.26	0.15	2.62
		GLPALISWIKRKRQQ	0.44	12.55	0.26	0.03	3.71
16		GIGAVLKVLTTGLPAL	0.45	9.65	0.06	1.02	1.07
		IGAVLKVLTTGLPALI	0.45	9.65	0.06	1.07	1.63
		GAVLKVLTTGLPALIS	0.45	9.65	0.06	0.96	0.42
		AVLKVLTTGLPALISW	0.44	9.65	0.06	1.01	0.34
		VLKVLTTGLPALISWI	0.45	9.65	0.06	1.03	0.22
	*	LKVLTTGLPALISWIK	0.47	10.80	0.12	0.87	0.17
	*	KVLTTGLPALISWIKR	0.49	11.80	0.18	0.68	0.36
	*	VLTTGLPALISWIKRK	0.49	11.80	0.18	0.53	1.03
		LTTGLPALISWIKRKR	0.48	12.55	0.25	0.34	1.95
		TTGLPALISWIKRKRQ	0.46	12.55	0.25	0.15	2.45
		TGLPALISWIKRKRQQ	0.45	12.55	0.25	0.05	3.47
17		GIGAVLKVLTTGLPALI	0.45	9.65	0.05	1.12	1.08
		IGAVLKVLTTGLPALIS	0.46	9.65	0.05	1.08	1.56
		GAVLKVLTTGLPALISW	0.45	9.65	0.05	0.99	0.41
		AVLKVLTTGLPALISWI	0.46	9.65	0.05	1.08	0.33
	*	VLKVLTTGLPALISWIK	0.46	10.80	0.11	0.97	0.26
	*	LKVLTTGLPALISWIKR	0.48	11.80	0.17	0.70	0.40
	*	KVLTTGLPALISWIKRK	0.48	11.95	0.23	0.51	0.97
		VLTTGLPALISWIKRKR	0.47	12.55	0.23	0.37	1.82
		LTTGLPALISWIKRKRQ	0.46	12.55	0.23	0.18	2.32
		TTGLPALISWIKRKRQQ	0.45	12.55	0.23	0.05	3.27
18		GIGAVLKVLTTGLPALIS	0.46	9.65	0.05	1.12	1.04
		IGAVLKVLTTGLPALISW	0.46	9.65	0.05	1.10	1.52
		GAVLKVLTTGLPALISWI	0.46	9.65	0.05	1.05	0.40
	/*	AVLKVLTTGLPALISWIK	0.46	10.80	0.11	1.02	0.36
	*	VLKVLTTGLPALISWIKR	0.47	11.80	0.16	0.79	0.47
		LKVLTTGLPALISWIKRK	0.48	11.95	0.22	0.54	0.97
		KVLTTGLPALISWIKRKR	0.46	12.55	0.27	0.37	1.72
		VLTTGLPALISWIKRKRQ	0.45	12.55	0.22	0.21	2.18
		LTTGLPALISWIKRKRQQ	0.45	12.55	0.22	0.08	3.10
19		GIGAVLKVLTTGLPALISW	0.46	9.65	0.05	1.14	1.02
		IGAVLKVLTTGLPALISWI	0.47	9.65	0.05	1.16	1.45
	/*	GAVLKVLTTGLPALISWIK	0.47	10.80	0.10	1.00	0.43
	*	AVLKVLTTGLPALISWIKR	0.47	11.80	0.15	0.84	0.55
		VLKVLTTGLPALISWIKRK	0.47	11.95	0.21	0.62	1.00
		LKVLTTGLPALISWIKRKR	0.46	12.55	0.26	0.40	1.67
		KVLTTGLPALISWIKRKRQ	0.45	12.55	0.26	0.22	2.06
		VLTTGLPALISWIKRKRQQ	0.44	12.55	0.21	0.11	2.91
20		GIGAVLKVLTTGLPALISWI	0.47	9.65	0.05	1.20	0.98
	*	IGAVLKVLTTGLPALISWIK	0.47	10.80	0.10	1.10	1.42

(continuación)

Cuadro D.1 (continuación)

L	Tipo	Péptido	MH	PI	C	CU	AG
	*	GAVLKVLTTGLPALISWIKR	0.47	11.80	0.15	0.84	0.60
		AVLKVLTTGLPALISWIKRK	0.47	11.95	0.20	0.68	1.05
		VLKVLTTGLPALISWIKRKR	0.45	12.55	0.25	0.47	1.67
		LKVLTTGLPALISWIKRKRQ	0.45	12.55	0.25	0.25	2.00
		KVLTTGLPALISWIKRKRQQ	0.44	12.55	0.25	0.12	2.76
21	*	GIGAVLKVLTTGLPALISWIK	0.47	10.80	0.09	1.13	0.97
	*	IGAVLKVLTTGLPALISWIKR	0.48	11.80	0.14	0.93	1.53
	*	GAVLKVLTTGLPALISWIKRK	0.47	11.95	0.19	0.68	1.07
		AVLKVLTTGLPALISWIKRKR	0.46	12.55	0.23	0.52	1.68
		VLKVLTTGLPALISWIKRKRQ	0.44	12.55	0.23	0.31	1.97
		LKVLTTGLPALISWIKRKRQQ	0.44	12.55	0.23	0.15	2.66
22	*	GIGAVLKVLTTGLPALISWIKR	0.48	11.80	0.13	0.97	1.10
		IGAVLKVLTTGLPALISWIKRK	0.48	11.95	0.18	0.76	1.91
		GAVLKVLTTGLPALISWIKRKR	0.46	12.55	0.22	0.53	1.66
		AVLKVLTTGLPALISWIKRKRQ	0.45	12.55	0.22	0.35	1.96
		VLKVLTTGLPALISWIKRKRQQ	0.43	12.55	0.22	0.20	2.61
23		GIGAVLKVLTTGLPALISWIKRK	0.48	11.95	0.17	0.80	1.49
		IGAVLKVLTTGLPALISWIKRKR	0.47	12.55	0.21	0.60	2.41
		GAVLKVLTTGLPALISWIKRKRQ	0.45	12.55	0.21	0.37	1.94
		AVLKVLTTGLPALISWIKRKRQQ	0.44	12.55	0.21	0.24	2.57
24		GIGAVLKVLTTGLPALISWIKRKR	0.47	12.55	0.20	0.64	2.00
		IGAVLKVLTTGLPALISWIKRKRQ	0.46	12.55	0.20	0.43	2.63
		GAVLKVLTTGLPALISWIKRKRQQ	0.44	12.55	0.20	0.25	2.52
25		GIGAVLKVLTTGLPALISWIKRKRQ	0.46	12.55	0.20	0.47	2.23
		IGAVLKVLTTGLPALISWIKRKRQQ	0.45	12.55	0.20	0.31	3.13
26		GIGAVLKVLTTGLPALISWIKRKRQQ	0.45	12.55	0.19	0.34	2.73

L: longitud del péptido, Tipo: [*,/], Péptido: péptido evaluado, MH: momento hidrofóbico $[0.4, 0.6]$ (véase la Sección 2.2.1), PI: punto isoeléctrico $[10.8, 11.8]$ (véase la Sección 2.2.1), C: carga neta promedio $[-0.2, 0.5]$ (véase la Ecuación 2.1), CU: algoritmo modificado de Uversky $[-0.2, 0.5]$ (véase la Ecuación 2.2), AG: AGADIR $[0.0, 10.0]$ (véase la Sección 2.2.1). APAP debe cumplir: MH, PI y AG. Los aceptados estan marcados con [*]. APAP–I debe cumplir: MH, PI y C > CU. Los aceptados estan marcados con [/]. Los péptidos que están calificados con [*] y [/] y están presentes en todas las longitudes están sombreados en color ■. Los péptidos seleccionados para prueba negativa están sombreados en color ■. Los péptidos que están calificados con [*] y [/] y sólo están presentes en las primeras longitudes están sombreados en color ■.

Apéndice E

Candidato PASAC en `gambicina`.

L	Tipo	Péptido	MH	PI	C	CU	AG
85		MKQQTVFVLLALLLVSASCVDALVYVYAKTCSTCR SLGARNCGYGSLGSKKYVSCDGATAIRNCDDCRRR FGTCQDRYITECFIG	0.31	9.90	0.05	0.17	7.16

L: longitud del péptido, Tipo: [*,/], Péptido: péptido evaluado, MH: momento hidrofóbico $[0.4, 0.6]$, PI: punto isoeléctrico $[10.8, 11.8]$, C: carga neta promedio $[-0.2, 0.5]$, CU: algoritmo modificado de Uversky $[-0.2, 0.5]$, AG: AGADIR $[0.0, 10.0]$. APAP debe cumplir: MH, PI y AG. Los aceptados estan marcados con *. APAP–II debe cumplir: MH, PI y C > CU. Los aceptados estan marcados con /. Los péptidos que están calificados tanto con * como por / y están presentes en todas las longitudes están sombreados en color ■. Los péptidos seleccionados para prueba negativa están sombreados en color ■. Los Los péptidos que están calificados tanto con * como por / y sólo están presentes en las primeras longitudes están sombreados en color ■.

Apéndice F

Detection of selective cationic amphipatic antibacterial peptides by Hidden Markov Models.

Artículo publicado en Acta Biochimica Polonica

Acta Biochimica Polonica

Vol. 56 No. 1/2009, 167–176

on-line at: www.actabp.pl

Regular paper

Detection of selective cationic amphipatic antibacterial peptides by Hidden Markov models

Carlos Polanco[1]✉ and Jose L. Samaniego[2]

[1]*Instituto de Fisiología Celular, and* [2]*Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior s/n Ciudad Universitaria Delegación Coyoacán, México*

Received: 30 January, 2009; revised: 03 March, 2009; accepted: 11 March, 2009
available on-line: 16 March, 2009

Antibacterial peptides are researched mainly for the potential benefit they have in a variety of socially relevant diseases, used by the host to protect itself from different types of pathogenic bacteria. We used the mathematical-computational method known as Hidden Markov models (HMMs) in targeting a subset of antibacterial peptides named Selective Cationic Amphipatic Antibacterial Peptides (SCAAPs). The main difference in the implementation of HMMs was focused on the detection of SCAAP using principally five physical-chemical properties for each candidate SCAAPs, instead of using the statistical information about the amino acids which form a peptide. By this method a cluster of antibacterial peptides was detected and as a result the following were found: 9 SCAAPs, 6 synthetic antibacterial peptides that belong to a subregion of Cecropin A and Magainin 2, and 19 peptides from the Cecropin A family. A scoring function was developed using HMMs as its core, uniquely employing information accessible from the databases.

Keywords: antibacterial peptides, Hidden Markov models

BACKGROUND

The increasing number of pathogens resistant to conventional antibiotics and the rising cost of production of the latter have led to the search for new drugs. One option for the development of these drugs is the production of antibacterial peptides found in nature, for these are the first defence line of living beings.

Antibacterial peptides have a wide variety of applications, from their use as antimicrobials to their use, after adaptations, as anticarcinogens (Ellerby *et al.*, 1999; Del Río *et al*., 2001) to human obesity control aids (Kolonin *et al*., 2004). It has also been observed that antibacterial peptides do not necessarily act exclusively against just bacteria. An example of a large non-specific antibacterial 85-peptide is gambicin: MKQQTVFVLLALLLVSASCVDALVYVYAKTC-STCRSLGARNCGYGSLGSKKYVSCDGATAIRNCD-DCRRRFGTCQDRYITECFIG-NH_2, which shows activity against bacteria and fungi (Vizioli *et al*., 2001).

The Selective Cationic Amphipatic Antibacterial Peptides (SCAAPs) are a recent and promising alternative for discovering new drugs effective in treating bacterial infections. They are characterized by being less than 60 amino acids in length, not adopting an α-helicoidal structure in neutral pH water solution and having a *therapeutic index* higher than 75 (Del Río *et al*., 2001). The therapeutic index of a peptide is defined (Ellerby *et al.*, 1999; Del Río *et al*., 2001) as the ratio between the minimum inhibitory concentrations observed against mammalian and bacterial cells: the higher the value, the more specific the peptide for bacterial-like membranes. In other words, SCAAPs display strong lytic activity against bacteria, but have no toxicity against normal eukaryotic cells such as erythrocytes (Shin *et al*., 2000).

Computer-based approaches may accelerate the discovery of new SCAAPs. However, detection of SCAAPs among every possible antibacterial peptide is not feasible either computationally or by biological assays. Their variation is 20^n where $n \in N$ is the

✉Corresponding author: Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Circuito Exterior s/n Ciudad Universitaria Delegación Coyoacán CP 04510, D.F., México; e-mail: polanco@unam.mx

length of peptide. For instance, an improved version of our program APAP (Del Río *et al*., 2001) executed on a cluster of 100 CPUs can not evaluate more than 20^{13} sequences of length 13 aa; it takes more than 10 months of processing time in a single PC (not shown). APAP-I, as well as APAP, evaluates the following physical-chemical properties for each peptide: isoelectric point (IP), average helical hydrophobic moment (HM), mean hydrophobicity (MH), mean net charge (MC) and AGADIR (helix/coil transition algorithm). APAP-I is 396000 times more efficient than the program APAP because it was designed to run on a high performance computing platform, and oriented to evaluate short peptides (8–11 aa). Thus, identification of new SCAAPs by searching the full space of peptide sequences may not be practical.

An alternative approach would be to search for new SCAAPs in sequences likely to have antibacterial activity. In this regard, it is possible to search for SCAAPs in peptides obtained from venoms (Conde *et al*., 2000) or to identify sequence patterns present in known antibacterial peptides. To identify such patterns, Hidden Markov Models (HMMs) provide a theory for profile methods (Resch, 2004; Prado-Prado *et al*., 2007a; 2007b). These HMMs may be used to predict new antibacterial peptides based on numeric indices of the peptide.

This type of study is known in the literature as Quantitative Structure-Activity Relationships (QSAR) or more generic Quantitative Structure-Property Relationships (QSPR) models. In fact, not only HMMs but other types of Markov models have been largely used to seek QSAR (quantitative structure-activity relationships)/QSPR (quantitative structure-property relationships) (González-Díaz *et al*., 2007f). For instance, the MARCH-INSIDE approach (Markov Chains Invariants for Network Simulation and Design) introduced by González-Díaz and coworkers makes use of Markov Chains theory to infer QSAR/QSPR models at different structural levels. Applications range from QSAR models of low-molecular-weight drugs (Santana *et al*., 2006; Cruz-Monteagudo *et al*., 2007; González-Díaz *et al*., 2007b; 2008b; Prado-Prado *et al*., 2008), to QSAR/QSPR models for protein and nucleic acid sequences (Aguero *et al*., 2008a; 2008b), protein 3D structure (González-Díaz *et al*., 2007a; 2007c; 2007d), RNA secondary structures (González-Díaz *et al*., 2003b; 2005; 2007e), viral surfaces (González-Díaz *et al*., 2003a) and of course peptides (Ramos de Armas *et al*., 2005).

The idea has been extended to include also Quantitative Proteome-Property Relationship (QPPR) models that personalize predictions of drug cardiotoxicity (González-Díaz *et al*., 2008a; 2008b; 2008c), or human prostate cancer (Ferino *et al*., 2008; González-Díaz *et al*., 2009), based on protein composition of Blood Proteomes. These Markov methods use different types of transition probabilities described by atom-atom, nucleotide-nucleotide, amino acid-amino acid, or even protein-protein matrices. Two recent in-depth reviews of the field were published recently (González-Díaz *et al*., 2008a; 2008c).

This article presents an approximation by Hidden Markov Models to detect SCAAPs based on physical-chemical similarity. As previously described (Del Río *et al*., 2001) the advantage of HMMs for this purpose is that they may identify patterns not obvious from iterative approaches such as APAP. This in turn may accelerate the discovery of new SCAAPs.

HMMs were implemented by using four sets of antibacterial peptides and one set of proteins:

Set A: 59 natural and synthetic antibacterial peptides extracted from (**set C**), which act exclusively against bacteria, fungi, viruses and mammalian cancer cells, with 3D structure determined by NMR spectroscopy or X-ray diffraction (NCBI, September, 2007).

Set B: 28 natural and synthetic antibacterial peptides extracted from (**set C**), which act exclusively against bacteria, with their 3D structure were detected by NMR spectroscopy or X-rays (NCBI, September, 2007).

Set C: 500 natural and synthetic antibacterial peptides which have a non-specific action against bacteria. The method used to predict the 3D structure is not relevant (NCBI, September, 2007).

Set D: 3 natural and synthetic antibacterial peptides extracted from (**set C**): Gambicin; Mellitin and Temporin H (XXA, frog) (NCBI, September, 2007).

Set E: 391836 natural and synthetic proteins detected in nature (Uniprot, August, 2008).

A *stochastic process* is a mathematical model for any phenomenon evolving or varying in time (or space etc.) subject to random influences (e.g., the stock market price of a commodity observed in time, the distribution of colors or shades in a noisy picture observed in an unordered two-dimensional lattice etc.).

Markov Models. Introduction

The condition prediction H at the time $t \in N$ is concerned with hypothesizing what the condition H will be at the time $t+1$, based on the observations of the condition H in the past (Resch, 2004).

We collected the relative frequency on the condition h_i (on time i) depending on what the condition H was like one day earlier h_{i-1}, the day before that h_{i-2}, and so forth.

The conditional probability is

$$\mathbf{P}\{h_n \mid h_{n-1}\} = \mathbf{P}\{h_n \mid h_{n-1}, h_{n-2}, \ldots, h_1\}$$

However, the larger the value of i is, the more observations we must collect. For n states of

the condition H the number of past histories will be $|H|^{n-1}$.

If we take the Markov assumption, we would have the probability of an observation at time i depend on h_{-1}. So we can express the probability of a sequence $\{h_1,...,h_n\}$ using this assumption:

$$\mathbf{P}\{h_1,\ldots,h_n\} = \mathbf{P}\{h_1\}\prod_{i=2}^{n}\mathbf{P}\{h_i \mid h_{i-1}\} \quad (1)$$

As a consequence of the Markov assumption, the number of past histories is reduced to $h_n \times h_{n-1}$.

HMMs. Mathematical description

If A, B are two events, then we define the probability of A given B as

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \quad (2)$$

One can work in the *mathematical ideal world* with the probability $\bar{\mathbf{P}}$ to achieve various mathematical objectives, and then reinterpret these results back in the *real world* with a measure change back to $\mathbf{P}$ *via* the inverse Radon-Nikodym derivative.

If circumstances only allow us to obtain the condition H based on another condition O, the condition H is hidden from us. We evaluate the conditional probability $\mathbf{P}(h_i|o_i)$ according to Eqn. (2).

$$\mathbf{P}(h_i \mid o_i) = \frac{\mathbf{P}(o_i \mid h_i)\mathbf{P}(h_i)}{\mathbf{P}(o_i)}$$

If we assume that, for all i the H_i, O_i, are independent of all o_j, h_j, for all $i \neq j$, Eqn. (1) can be rewritten as

$$\mathsf{L}(h_1,\ldots,h_n \mid o_1,\ldots,o_n) \propto \mathbf{P}(h_1,\ldots,h_n \mid o_1,\ldots,o_n) = \prod_{i=1}^{n}\mathbf{P}(o_i \mid h_i).\prod_{i=1}^{n}\mathbf{P}(h_i \mid h_{i-1}) \quad (3)$$

Eqn. (3) is known as a measure of the probability and is referred to as the *likelihood* function L.

The expectation maximization (EM) algorithm reestimates the parameters of the model.

Many of the density functions are exponential in nature; it is therefore easier to compute the EM of a likelihood function by finding the maximum of the *natural ln* of L, known as the *ln-likelihood* function:

$$l(h_i|o_i) = \ln(\mathrm{L}(h_i|o_i))$$

due to the monotonicity of the *ln* function.

HMMs. Terminology

HMMs are specified by the set of states $S = \{s_1,s_2,...,s_n\}$, corresponding to the possible condition H, and the parameter set $\Omega = \{\pi, A,B\}$:

The **initial probabilities** $\pi_i = \mathbf{P}(h_i = s_i)$ are probabilities of s_i being the first state of a state sequence h_i. They are collected in the vector P_0.

The **transition probabilities** are the probabilities that go from state i to state j: $a_{i,j} = \mathbf{P}(h_n = s_j) | h_{n-1} = s_i)$. They are collected in matrix A.

The **emission probabilities** characterize the likelihood of a discrete observation $o_n \in \{v_i,...,v_n\}: b_{i,k} = \mathbf{P}(o_n = v_k | h_n = s_i)$, and the probabilities to observe v_k if the current state is $h_n = s_i$. The numbers $b_{i,k}$ are gathered in matrix B.

The likelihood of $O = \{o_1,...,o_n\}$ along the path $H = \{h_1,...,h_n\}$ determined from HMMs with parameters Ω, is given by:

$$\mathsf{L}(h_i \mid o_i) \propto \mathbf{P}(O,H \mid \Omega) = \prod_{i=1}^{n}\mathbf{P}(O \mid H,\Omega)\prod_{i=1}^{8}\mathbf{P}(H \mid \Omega) \quad (4)$$

where the probabilities $\mathbf{P}(O|H,\Omega)$ and $\mathbf{P}(H|\Omega)$ are expressed in terms of matrices A, B (Eqns. 5 and 6) and the vector P_0.

$$\mathbf{P}(O \mid H,\Omega) = \prod_{i=1}^{n}\mathbf{P}(H,\Omega) = b_{h_1,o_1}\, b_{h_2,o_2} \ldots b_{h_n,o_n} \quad (5)$$

$$\mathbf{P}(H \mid \Omega) = \pi_{h_1}\prod_{i=1}^{8} a_{h_i,h_{i+1}} = \mathsf{P}_{0h_i}\, a_{h_1,h_2}\, a_{h_2,h_3} \ldots a_{h_7,h_8} \quad (6)$$

$\mathbf{P}(O,H|\Omega)$ (Eqn. 4) is known as the joint *likelihood* of an observation sequence and it is equivalent to Eqn. (1).

HMMs. Implementation

The set of states S corresponding to the twenty different amino acids from which every antibacterial peptide is formed: S = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}, and the parameter set was formed by $\Omega = \{P_0,A,B\}$.

The **vector** p_0 contains $\frac{1}{n}\sum_{i=1}^{n}\mathsf{p}_{0i}$, where n is the length of the peptide to be tested, and p_{0i} is the relative frequency distribution of amino acids from the same peptide, derived from the absolute frequency distribution from natural and synthetic antibacterial peptides from **(set A)** (Table 1). Their 3D structure was detected by NMR spectroscopy or X-rays dif-

Table 1. Elements of vector P_0.

Absolute frequency of natural and synthetic antibacterial peptides which act exclusively against bacteria, fungi, viruses and mammalian cancer cells **(set A)** to vector P_0. The letters in the table refer to the 20 amino acids (one-letter code), and the numbers represent the corresponding frequency of that amino acid in the set.

A	C	D	E	F	G	H	I	K	L	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
103	132	23	32	61	182	39	129	146	101	9	52	67	49	135	87	53	85	31	57

Table 2. Elements of matrix A.

Absolute frequency distribution of all amino acids taken of pairs (contiguously), from **(set C)**. Every letter is equivalent to each amino acid, in this manner, the occurrence of pair of amino acids (A_{ci}; A_{cj}) is built with the amino acid from row ($_i$) and the amino acid from column ($_j$).

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	165	38	11	29	28	134	17	85	205	11	66	40	15	24	34	76	50	64	9	2
C	50	87	15	7	34	32	24	30	85	43	4	17	41	7	142	41	30	43	4	54
D	24	17	8	3	15	10	2	32	25	27	6	2	4	7	4	16	20	28	9	9
E	14	8	5	8	6	20	15	9	48	24	4	7	4	8	44	19	8	13	3	3
F	27	69	17	4	23	48	14	25	62	109	1	14	33	18	49	34	9	26	3	10
G	103	51	19	39	61	10	25	137	164	185	15	39	74	55	102	62	64	89	33	53
H	14	18	5	11	15	18	19	17	15	29	6	3	10	4	25	19	14	57	0	4
I	95	40	13	25	43	143	31	53	108	69	10	31	53	21	59	68	28	45	6	19
L	105	55	34	25	60	155	24	55	143	129	9	23	108	26	65	80	22	51	28	10
M	15	4	6	5	2	11	0	6	12	18	0	8	1	5	12	6	2	7	1	2
N	30	17	6	8	27	37	10	14	29	36	11	9	23	4	36	12	23	36	3	7
P	43	16	8	7	45	47	16	79	45	36	6	21	55	17	69	30	18	64	13	17
Q	44	23	3	4	12	47	12	27	24	7	5	12	21	20	16	8	16	16	4	2
R	40	62	32	17	45	94	23	60	73	69	7	48	97	30	118	35	20	54	20	21
S	63	67	13	16	20	78	21	43	74	52	14	15	12	17	33	31	28	52	10	15
T	51	79	8	5	17	34	5	38	29	47	8	4	14	15	44	11	13	38	4	13
V	107	59	14	13	36	133	13	49	63	103	2	22	47	11	46	47	32	60	8	12
W	15	8	5	8	5	16	3	9	31	22	3	14	9	9	5	5	2	4	4	0
Y	10	51	3	2	6	30	1	13	22	20	1	11	9	5	39	14	19	11	0	8

fraction, and was taken from the database BBCM (NCBI, September, 2007).

The **matrix A** represents the relative frequency of all 400 possible pairs of amino acids. These pairs were taken in two directions: ($a_{i,j}$,$a_{i+1,j}$) and ($a_{i-1,j}$,$a_{i,j}$), for specific *j*. The matrix was built from natural and synthetic antibacterial peptides which have non-specific action against bacteria **(set C)**; the method used to predict the 3D structure is not relevant (Table 2). These peptides were taken from the database BBCM (NCBI, September, 2007).

Every pair of amino acids from the peptide to be tested was extracted from **matrix A**.

The **matrix B** exhibits the conditional probability of the peptide to be tested as the result of two conditions: first, the calculation of each natural and synthetic antibacterial peptide by program APAP-I (this program evaluated if the peptide is or is not a candidate SCAAP); second, if the $\text{Index}_A \geq 0.08$.

Index A (Eqn. 7) is formed by the relative frequency distribution of amino acid A_i from the peptide to be tested, derived from the absolute frequency distribution from natural and synthetic antibacterial peptides which act exclusively against bacteria **(set B)** (Table 3). (NCBI, September, 2007).

$$\textbf{Index}_{\mathbf{A}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{A}_i, i \in [1,n] \qquad (7)$$

The program APAP-I was used to evaluate if a peptide to be tested from **(set C)** was a candidate SCAAP or not, with the evaluation of different physical-chemical properties. APAP-I is formed by two subprograms:

APAP-IA which evaluated the isoelectric point IP, helical hydrophobic moment HM and AGADIR.

APAP-I-B which evaluated the isoelectric point IP, helical hydrophobic moment HM, mean hydrophobicity MH and mean net charge MC.

The physical-chemical properties in acceptable ranges were:

Isoelectric point (IP) (Del Rio *et al*., 2001). This is the pH at which a particular peptide carries no net electrical charge. The value range considered was from 10.8 to 11.8.

Helical hydrophobic moment (HM) (Eisenberg *et al*., 1982). This is a sum of the hydrophobicities of the side chains of a helix of *n* amino acids. The length of a vector corresponding to the hydrophobicities is the numerical hydrophobicity associated to the kind of side chain, and its direction is determined by the orientation of the side chain according to the helix axis. A large value of HM means that the helix is amphiphilic perpendicular to its axis. The value range considered was from 0.4 to 0.6.

Table 3. Elements of vector Index_A.

Absolute frequency of natural and synthetic antibacterial peptides which act exclusively against bacteria **(set B)** to vector Index_A. The letters in the table refer to the 20 amino acids (one-letter code), and the numbers represent the corresponding frequency of that amino acid in the set.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
49	68	10	16	35	106	22	71	112	52	2	30	33	19	60	42	24	45	17	27

Mean hydrophobicity (MH) (Del Río *et al*., 2001). This is the mean of the hydrophobicities of the amino acids normalized to 1 over all amino acids of the peptide. The algorithm was given by the technical department of the Swiss Institute of Bioinformatics (Swiss). The value range considered was from 0.35 to 0.55.

Mean net charge (MC) (Del Río *et al*., 2001). This is determined by Eqn. (8). The algorithm was given by Uversky (Uversky, 2000; Uversky *et al*., 2002).

$$\mathrm{MC}(R,K,D,E) = \frac{1}{n}(R_i + K_i - D_i - E_i), i \in [1,n] \quad (8)$$

The variables R_i, K_i, D_i and E_i represent the number of times the amino acids arginine (R), lysine (K), aspartic acid (D) and glutamic acid (E) appeared, accepting those peptides whose *MC*(*R,K,D,E*) evaluated with Eqn. (8) are above or equal to the number obtained by Eqn. (9) with the same mean hydrophobicity (MH).

$$\mathrm{MC}(MH) = 45.896MH^4 - 47.528MH^3 + 13.324MH^2 + 2.302MH - 1.291 \quad (9)$$

AGADIR (Lacroix *et al*., 1997; Del Río *et al*., 2001). Predicts the helical behaviour of a peptide. The value range considered was from 0.00 to 10.00.

The **matrix B** shows the conditional probability of $\mathbf{P}(o_i | h_{i?\mathrm{IndexA}})$ to be candidate SCAAPs if (o_i = *true*) the $\mathbf{P}(o_i = true | h_{i?\mathrm{IndexA}}) = 0.95$, and its complement ($o_i$ = *false*) $\mathbf{P}(o_i = false | h_{i?\mathrm{IndexA}}) = 0.05$. These numbers are obtained as a result of many computational assays.

HMMs. Tests

As a **negative test**, the validation of HMMs to detect candidate SCAAPs consisted of testing:

The total number of natural and synthetic antibacterial peptides which had a non-specific action and whose structure could not be determined by either method **(set C)** (i.e. NMR spectroscopy or X-rays) over two sets:

A set of three natural and synthetic antibacterial peptides **(set D)**: Gambicin characterized by non-specific action and no SCAAPs (according to the program APAP-I); Mellitin characterized by toxicity against erythrocytes; Temporin [H XXA, frog] was determined by circular dichroism (CD).

The total number of natural and synthetic proteins that were detected in nature **(set E)** were used to build the matrices A and B, and test the **(set C)**.

HMMs. Statistical analysis

A two-sample rank test by Wilcoxon, Mann and Whitney (Kreyszig, 1979) was made to test over two populations:

Natural and synthetic antibacterial peptides **(set C)** *versus* natural and synthetic antibacterial peptides which act exclusively against bacteria **(set B)**.

Natural and synthetic antibacterial peptides with an exclusive action against bacteria **(set B)** *versus* natural and synthetic antibacterial peptides detected by program APAP-I.

These statistical tests were used to verify the hypothesis that two populations have the same distribution to be a candidate SCAAPs or not. The assumption was that the populations tested correspond to continuous distributions, and to obtain critical values c_1 and c_2, using the fact that if the hypothesis is true, then the random variable *W*, over the populations described is approximately normal with mean and variance (Eqns. 10 and 11)

$$\mu_W = \frac{n_1(n_1 + n_2 + 1)}{2} \quad (10)$$

$$\sigma_W^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \quad (11)$$

Hence c_1 and c_2 were obtained substituting μ_W and σ_W in Eqns. (12) and (13)

$$P(W \le c_1) = \Phi\left(\frac{c_1 - \mu_W}{\sigma_W}\right) = 2.5\% \quad (12)$$

$$P(W \ge c_2) = 1 - \Phi\left(\frac{c_2 - \mu_W}{\sigma_W}\right) = 2.5\% \quad (13)$$

The test was conducted only on the **(sets B** and **C)** because this pair is more similar than the other sets involved **(A, D** and **E)**.

RESULTS

Objective

The use of HMMs for prediction and understanding of antimicrobial peptides has been reported for the last three decades (Andrés & Dimarcq, 2007), particularly the detection of antimicrobial peptides by multivariate linear regression and physical-chemical properties (Hilpert *et al*., 2008).

In this article we use HMMs for the prediction of candidate SCAAPs based on five physical chemical properties: isoelectric point (IP), helical hydrophobic moment (HM), mean hydrophobicity (MH), mean net charge (MC), and AGADIR; and the relative frequency distribution of single and pair amino acids over the sequence of the peptide.

Identification of SCAAPs

We retrieved a cluster of 57 natural and synthetic antibacterial peptides (Table 4) which act ex-

Table 4. Cluster of antibacterial peptides predicted by HMMs and listed in descending order (set C)

NL: Position of the antibacterial peptide on the list. NP: Number which corresponds to the antibacterial peptide according to HMMs. F: Family. If natural SCAAPs were a part of **(set B)**, [*S*]. If Brevinin, [*B*]. If Cathelin, [*Ca*]. If Cecropin, [*C*]. If Moricin, [M]. AP-A: Peptide which was accepted by the program APAP-IA (Section HMMs. Implementation). AP-B: Peptide which was accepted by the program APAP-IB (Section HMMs. Implementation)

NL	NP	F	AP-A	AP-B	Name of the sequence	References
1	454	*C*	+	+	Cecropin-B type 1 precursor (Cecropin-B1) [Contains: Cecropin-B (AalCecB); Cecropin-B amidated isoform]	Sun *et al.*, 1999
2	417		+	+	Parabutoporin	Moerman *et al.*, 2002
3	16	*S;C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P3)	Oh *et al.*, 1999
4	61	*C*	+	+	Cecropin B [Bombyx mori]	Taniai *et al.*, 1995
5	458	*S*	+	+	Cathelin-like protein [Mus musculus]	Popsueva *et al.*, 1996
6	172	*C*	+	+	Hyphancin-3D precursor (Hyphancin-IIID) (Cecropin-A)	
7	15	*S;C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12)	
					Hybrid Peptide Analogue(P4)	Oh *et al.*, 1999
8	58 C		+	+	Cecropin-B	Ku *et al.*, 1982
9	174	*C*	+	+	Hyphancin-3F precursor (Hyphancin-IIIF) (Cecropin-A2)	
10	68		+	+	Defensin NP-3a [Oryctolagus cuniculus]	Linzmeier *et al.*, 1993
11	57	*S*	+	+	Cecropin-A precursor (Cecropin-C)	Gudmundsson *et al.*, 1991
12	425	*C*	+	+	RecName: Full=Cecropin-A	
13	175	*C*	+	+	Hyphancin-3G precursor (Hyphancin-IIIG) (Cecropin-A3)	
14	259	*C*			Cecropin-A1 precursor (Cecropin-A) (AalCecA)	Sun *et al.*, 1999
15	356				Ranatuerin-2Lb	Soraya *et al.*, 2000
16	173	*C*	+	+	Hyphancin-3E precursor (Hyphancin-IIIE) (Cecropin-A1)	
17	176	*Ca*	+	+	Cathelin-like protein [Mus musculus]	Popsueva *et al.*, 1996
18	474				Sentrin/SUMO-speci_c protease [Plasmodium yoelii yoelii str. XNL]	Carlton *et al.*, 2002
19	67		+	+	Defensin NP-3a [Oryctolagus cuniculus]	Linzmeier *et al.*, 1993
20	32	*S;M*			Chain A, Solution Structure of Antibacterial Peptide (Moricin)	Hemmi *et al.*, 2002
21	52				M Moricin [Bombyx mori].	Hemmi *et al.*, 2002
22	9	*S;C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide	Oh *et al.*, 1999
23	74				GK14120 [Drosophila willistoni]	Zimin *et al.*, 2008
24	426	*M*			RecName: Full=Virescein.	
25	106		+	+	Xenopsin precursor protein [Xenopus laevis]	Moore *et al.*, 1991
26	169		+	+	Antibacterial peptide PMAP-37 precursor (Myeloid antibacterial peptide 37)	Tossi *et al.*, 1995 Tossi *et al.*, 1995
27	435	*C*			Cecropin 1 [Musca domestica]	Tossi *et al.*, 1995
28	424	*C*			Cecropin precursor	Tossi *et al.*, 1995
29	75				Sarcotoxin-1B precursor (Sarcotoxin IB)	Kanai *et al.*, 1989
30	355		+	+	Hadrurin	Torres-Larios *et al.*, 2000
31	434	*C*			Cecropin-1 precursor	Rosetto *et al.*, 1993
32	493	*C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P2)	Oh *et al.*, 1999
33	490	*C*	+	+	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12)	
					Hybrid Peptide Analogue(P3)	Oh *et al.*, 1999
34	14	*S;C*	+	+	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P2)	Oh *et al.*, 1999
35	469		+	+	Ribosomal protein L1 [Helicobacter pylori G27]	
36	386	*B*			Brevinin-1SY	Matutte *et al.*, 2000
37	102				Megakaryocyte stimulating factor [Trichomonas vaginalis G3]	Carlton *et al.*, 2002
38	127	*B*			RecName: Full=Brevinin-1	Morikawa *et al.*, 1992
39	405				Maximin-H14 antimicrobial peptide precursor [Bombina maxima]	
40	267				Neutrophil defensin 3 (HANP-3)	Mak *et al.*, 1996
41	265				Neutrophil defensin 1 (HANP-1)	Mak *et al.*, 1996
42	380	*B*			Brevinin 1Pb precursor [Rana pipiens]	Tennessen *et al.*, 2007
43	119	*C*			Cecropin C CG1373-PA [Drosophila melanogaster]	Hoskins *et al.*, 2007
44	56				Bombinin	
45	263				Fabatin precursor [Vicia faba]	
46	262				Fabatin precursor [Vicia faba]	
47	125	*B*			Brevinin-1E	Marenah *et al.*, 2006
48	346				Ponericin-W2	Orivel *et al.*, 2001
49	345				Ponericin-W1	Orivel *et al.*, 2001
50	132				Ceratotoxin A [Ceratitis capitata]	Rosetto *et al.*, 1993
51	239				Gaegurin-6	Park *et al.*, 1995
52	488				Nigrocin-2P precursor [Rana palustris]	
53	137				Ranalexin precursor	Clark *et al.*, 1994
54	392				Temporin-1Ca	Halverson *et al.*, 2000
55	19	*S;Ca*			Cathelin-related peptide SC5 precursor 1 (Antibacterial peptide SMAP-29) (Myeloid antibacterial peptide MAP-29)	Mahoney *et al.*, 1995
56	112				Defensin related cryptdin 4 [Mus musculus]	Strausberg *et al.*, 2002
57	459	*S;Ca*			Cathelin-like protein [Mus musculus]	Popsueva *et al.*, 1996

clusively against bacteria, fungi, viruses and mammalian cancer cells, whose 3D structure was determined by NMR spectroscopy or X-rays from the BBCM protein database (NCBI, September, 2007) **(set C)**. From this set we generated one subset, according to their structure: 28 antibacterial peptides determined by NMR spectroscopy **(set B)**. An HMM profile of the SCAAP family was built from these sets. After calibration, the HMMs were used to search through 500 natural and synthetic antibacterial peptides which have a non-specific action against bacteria (NCBI, September, 2007) **(set C)**; nine hits were found from the search on 500 antibacterial peptides (9, 14, 15, 16, 19, 32, 57, 458 and 459), six synthetic antibacterial peptides were found in Cecropin A and Magainin 2 (3, 9, 14, 15, 490 and 493), 19 peptides were from the Cecropin A family (9, 14, 15, 16, 58, 61, 119, 172, 173, 174, 175, 259, 424, 425, 434, 435, 454, 490 and 493); four peptides were from the Brevinin family (125, 127, 380 and 386), three peptides from the Cathelin family (19, 176 and 459), and two peptides from the Moricin family (32 and 52).

The entire cluster was further analyzed by a search against Swiss-Prot and Translated EMBL protein databases by Smith-Waterman algorithm on GCG/SeqWeb to ensure the identification of these peptides. They are described in Table 4.

Note that the peptide number 32 (position 20 in Table 4) was not accepted by the programs APAP-IA and APAP-IB, but it was accepted by HMMs because of its score.

Negative tests of HMMs

HMMs were tested with:

Three peptides: Gambicin characterized by non-specific action against bacteria, fungi, viruses and mammalian cancer cells; Mellitin characterized by toxicity against erythrocytes; and Temporin H [XXA, frog] determined by circular dichroism (CD). All peptides were accepted by HMMs.

As a full test, we retrieved the complete set of proteins (391 836) from the Uniprot protein database and a new HMM profile was built from these sequences. After calibration, the new HMMs were used with the same set of 500 natural and synthetic antibacterial peptides **(set C)** that we refer to in the identification of SCAAPs in Table 4: No candidate SCAAPs or SCAP family was detected.

Statistical verification of HMMs

In order to verify if a statistical similarity exists between the referred set of peptides involved in the tests, we decided to compare only the more biologically similar sets: the set of 500 natural and synthetic antibacterial peptides which have non-specific action against bacteria **(set C)**, and the set of 28 natural and synthetic antibacterial peptides which act exclusively against bacteria, with their 3D structure detected by NMR spectroscopy or X-ray diffraction **(set B)**.

We ran a Wilcoxon, Mann and Whitney non-parametric test (with $p\text{-}value < 0.05$): the test did not observe any normal correlation between those sets, and consequently it was concluded that no sets had any statistical relation.

DISCUSSION

In this article, we have described the detection of nine SCAAPs by applying a mathematical-computational tool, the HMM search on a predicted peptide database. Compared with the experimental assay search, the HMM is much more sensitive due to its summarizing nature. The key point for a successful HMM search lies in constructing the HMMs profile (a combination of physical-chemical properties and relative frequency distribution of amino acids over the sequence of the peptide). The inclusion of the complete set of proteins from the Uniprot protein database in order to reconfigure HMMs, and the inclusion of three wrong sequences provides more reliability and robustness of this HMM profile.

We recognize some bias with this approach. The major issue is related to the incompleteness of the existing databases. The degree to which the current database is complete is not known, even though our studies are designed to be exhaustive.

While this manuscript was being prepared, a paper was in press that described the detection of short linear cationic antimicrobial peptides using, principally, the nonlinear techniques of support vector machines and artificial neural networks (Hilpert *et al*., 2008). Their methods are more selective and less comprehensive than HMMs described. Thus, these two approaches could be used as complementary tools in identifying novel candidate members of a specific protein family.

Comparative studies

Our HMMs profile was compared with three stochastic methods named HMMER (HMMER), MAST (Bailey & Gribskov, 1998; 2000) and GLAM (Frith *et al*., 2004). These comparisons were concerned with the number of hits each method offers, and the results show that GLAM was superior to the other methods but that HMMs, MAST and HMMER were equally effective.

CONCLUSIONS

The HMMs profile is a mathematical-computational tool for finding potential peptides named Selective Cationic Amphipatic Antibacterial Peptides (SCAAPs) solely by employing information accessible from the databases to provide adequate peptide identification performance. It allows rapid, convenient searches within databases. In summary, HMMs profiles show significant selective efficacy in the detection of SCAAPs, and are a useful model for biological sequence analysis and modeling in the postgenomic era.

Acknowledgements

The authors thank Gabriel del Rio (IFC-UNAM) for critical comments on this article, and Dr. Luis A. Rincon (FC-UNAM) for valuable suggestions to the manuscript. To Instituto de Fisiologia Celular (IFC) and Departamento de Matematicas of the Facultad de Ciencias (FC) for the backing given to the elaboration of this article. To Doctorado en Ciencias Biomedicas and Universidad Nacional Autonoma de Mexico (UNAM) for supporting this research. The authors thank Dagny Valadez and Carlos Warden (FFyL UNAM) for the proof reading of this article.

REFERENCES

Aguero-Chapin G, Antunes A, Ubeira FM, Chou KC, Gonzalez-Diaz H (2008a) Comparative Study of topological indices of macro/supramolecular RNA complex networks. *J Chem Inf Model* **48**: 2265–2277.

Aguero-Chapin G, Gonzalez-Diaz H, de la Riva G, Rodriguez E, Sanchez-Rodriguez A, Podda G *et al* (2008b) MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from *Schizosaccharomyces pombe*, prediction, and experimental assay of a new sequence. *J Chem Inf Model* **48**: 434–448.

Andrés E, Dimarcq JL (2007) Cationic antimicrobial peptides: from innate immunity study to drug development. Update. *Med Mal Infect* **37**: 194–199.

Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**: 48–54.

Bailey TL, Gribskov M (2000) Concerning the accuracy of MAST E-values. *Bioinformatics* **16**: 488–489.

Boulanger N, Brun R, Ehret-Sabatier L, Kunz C, Bulet (2002) Immunopeptides in the defense reactions of Glossina morsitans to bacterial and *Trypanosoma brucei brucei* infections. *Insect Biochem Mol Biol* **32**: 369–375.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, Peterson JD, Pop M, Kosack DS, Shumway MF, Bidwell SL, Shallom SJ, van Aken SE, Riedmuller SB, Feldblyum TV, Cho JK, Quackenbush J, Sedegah M, Shoaibi A, Cummings LM, Florens L, Yates JR, Raine JD, Sinden RE, Harris MA, Cunningham DA, Preiser PR, Bergman LW, Vaidya AB, van Lin LH, Janse CJ, Waters AP, Smith HO, White OR, Salzberg SL, Venter JC, Fraser CM, Hoffman SL, Gardner MJ, Carucci DJ (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**: 512–519.

Clark DP, Durell S, Maloy WL, Zasloff M (1994) Ranalexin. A novel antimicrobial peptide from bullfrog (Rana catesbeiana) skin, structurally related to the bacterial antibiotic, polymyxin. *J Biol Chem* **269**: 10849–10855.

Conde R, Zamudio FZ, Rodríguez MH, Possani LD (2000) Scorpine, an anti-malaria and anti-bacterial agent purified from scorpion venom. *FEBS Lett* **471**: 165–168.

Cruz-Monteagudo M, González-Díaz H, Aguero-Chapin G, Santana L, Borges F, Domínguez RE *et al.* (2007) Computational chemistry development of a unified free energy Markov Model for the distribution of 1300 chemicals to 38 different environmental or biological systems. *J Comput Chem* **28**: 1909–1922.

Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MNDS, Uriartei E, Chou K-C *et al*. (2008a) Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case polymer. *Polymer* **49**: 5575–5587.

Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MN, Uriarte E, Gonzalez-Diaz H (2008b) Quantitative Proteome-Property Relationships (QPPRs). Part 1: finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra. *Bioorg Med Chem* **16**: 9684–9693.

Cruz-Monteagudo M, González-Díaz H, Borges F, Dominguez ER, Cordeiro MN (2008c) 3D-MEDNEs: An alternative *in silico*. Technique for chemical research in toxicology. 2. Quantitative Proteome-Toxicity Relationships (QPTR) based on mass spectrum spiral entropy. *Chem Res Toxicol* **21**: 619–632.

Del Río G, Castro-Obregon S, Rao R, Ellerby MH, Bredesen DE (2001) APAP, a sequence-pattern recognition approach identifies substance P as a potential apoptotic peptide. *FEBS Lett* **494**: 213–219.

Eisenberg D, Weiss RM, Terwilliger TC (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **23**: 371–374.

Ellerby MH, Arap W, Kain R, Andrusiak R, Del Río G, Krajewski S, Lombardo CR, Rao R, Ruoslahti E, Bredesen DE, Pasqualini R (1999) Anti-cancer activity of targeted pro-apoptotic peptides. *Nat Med* **5**: 1032–1038.

ExPASy Proteomics Server. http://www.expasy.ch/sprot

Ferino G, Gonzalez-Diaz H, Delogu G, Podda G, Uriarte E (2008) Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer. *Biochem Biophys Res Commun* **372**: 320–325.

Ferino G, Delogu G, Podda G, Uriarte E, González-Díaz H (2009) Quantitative proteome-disease relationships (QPDRs) in clinical chemistry: prediction of prostate cancer with spectral moments of PSA/MS star networks. In *Clinical Chemistry Research*; Mitchem BHaS, ChL, ed. NY: Nova Science Publisher.

Frith CM, Hansen U, Spouge JL, Weng Z (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res* **32**: 189–200.

González-Díaz H, Molina RR, Uriarte E (2003a) Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropie. *Polymer* **45**: 3845–3853.

González-Díaz H, de Armas RR, Molina R (2003b) Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Psi-RNA

packaging region with drugs. *Bioinformatics* **19**: 2079–2087.

González-Díaz H, Pérez-Bello A, Uriarte E (2005) Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA-QSAR for mycobacterial promoters. *Polymer* **46**: 6461–6473.

González-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E (2007a) A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *J Proteome Res* **6**: 904–908.

Gonzalez-Diaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A (2007c) Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. *Comput Chem* **28**: 1042–108.

González-Díaz H, Pérez-Castillo Y, Podda G, Uriarte E (2007d) Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. *J Comput Chem* **28**: 1990–1995.

González-Díaz H, Aguero-Chapin G, Varona J, Molina R, Delogu G, Santana L *et al*. (2007e) 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. *J Comput Chem* **28**: 1049–1056.

González-Díaz H, Vilar S, Santana L, Uriarte E (2007f) Medicinal chemistry and bioinformatics — current trends in drugs discovery with networks topological indices. *Curr Top Med Chem* **7**: 1025–1039.

Gonzalez-Diaz H, Prado-Prado F, Ubeira FM (2008a) Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr Top Med Chem* **8**: 1676–1690.

González-Díaz, Prado-Prado F (2008b) Unified QSAR and network-based computational chemistry approach to antimicrobials. Part 1: Multispecies activity models for antifungals. *J Comput Chem* **29**: 656–657.

González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008c) Proteomics, networks and connectivity indices *Proteomics* **8**: 750–778.

Gudmundsson GH, Lidholm DA, Asling B, Gan R, Boman HG (1991) The cecropin locus. Cloning and expression of a gene cluster encoding three antibacterial peptides in *Hyalophora cecropia*. *J Biol Chem* **266**: 11510–11517.

Hemmi H, Ishibashi J, Hara S, Yamakawa M (2002) Solution structure of moricin, an antibacterial peptide, isolated from the silkworm *Bombyx mori*. *FEBS Lett* **518**: 33–38.

Hilpert K, Fjell CD, Cherkasov A (2008) Short linear cationic antimicrobial peptides: screening, optimizing, and prediction. *Methods Mol Biol* **494**: 127–159.

HMMER, UBC Bioinformatics Centre. http://bioinformatics.ubc.ca/resources/tools/hmmer

Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, Villasante A, Dimitri P, Karpen GH, Celniker SE (2007) Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science* **316**: 1625–1628.

Goraya J, Wang Y, Li Z, O'Flaherty M, Knoop FC, Platz JE, Conlon JM (2000) Peptides with antimicrobial activity from four different families isolated from the skins of the North American frogs *Rana luteiventris, Rana berlandieri* and *Rana pipiens*. *Eur J Biochem* **267**: 894–900.

Halverson T, Basir YJ, Knoop FC, Conlon JM (2000) Purification and characterization of antimicrobial peptides from the skin of the North American green frog *Rana clamitans*. *Peptides* **21**: 469–476.

Kanai A, Natori S (1989) Cloning of gene cluster for sarcotoxin I, antibacterial proteins of *Sarcophaga peregrine*. *FEBS Lett* **258**: 199–202.

Kreyszig E (1970) *Introductory Mathematical Statistics, Principles and Methods*. John Wiley & Sons. Inc. New York.

Kolonin MG, Saha PK, Chan L, Pasqualini R, Arap W (2004) Reversal of obesity by targeted ablation of adipose tissue. *Nat Med* **10**: 625–632.

Lacroix E, Viguera AR, Serrano L (1997) Elucidating the folding problem of α-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol* **1**: 173–191.

Linzmeier R, Michaelson D, Liu L, Ganz T (1993) The structure of neutrophil defensin genes. *FEBS Lett* **321**: 267–273.

Mahoney MM, Lee AY, Brezinski-Caliguri DJ, Huttner KM (1995) Molecular analysis of the sheep cathelin family reveals a novel antimicrobial peptide. *FEBS Lett* **377**: 519–522.

Mak P, Wójcik K, Thogersen IB, Dubin A (1996) Isolation, antimicrobial activities, and primary structures of hamster neutrophil defensins. *Infect Immun* **64**: 4444–4449.

Marenah L, Flatt PR, Orr DF, Shaw C, Abdel-Wahab YH (2006) Skin secretions of Rana saharica frogs reveal antimicrobial peptides esculentins-1 and -1B and brevinins-1E and -2EC with novel insulin releasing activity. *J Endocrinol* **188**: 1–9.

Matutte B, Storey KB, Knoop FC, Conlon JM (2000) Induction of synthesis of an antimicrobial peptide in the skin of the freeze-tolerant frog, *Rana sylvatica*, in response to environmental stimuli. *FEBS Lett* **483**: 135–138.

Moerman L, Bosteels S, Noppe W, Willems J, Clynen E, Schoofs L, Thevissen K, Tytgat J, Van Eldere J, Van Der Walt J, Verdonck F (2002) Antibacterial and antifungal properties of α-helical, cationic peptides in the venom of scorpions from southern Africa. *Eur J Biochem* **269**: 4799–4810.

Moore KS, Bevins CL, Brasseur MM, Tomassini N, Turner K, Eck H, Zasloff M (1991) Antimicrobial peptides in the stomach of *Xenopus laevis*. *J Biol Chem* **266**: 19851–19857.

Morikawa N, Hagiwara K, Nakajima T (1992) Brevinin-1 and -2, unique antimicrobial peptides from the skin of the frog, *Rana brevipoda* porsa. *Biochem Biophys Res Commun* **189**: 184–190.

NCBI, National Center for Biotechnology Information (NCBI) Protein BLAST. http://www.ncbi.nlm.nih.gov

Oh D, Shin SY, Kang JH, Hahm KS, Kim KL, Kim Y (1999) NMR structural characterization of cecropin A(1-8) - magainin 2(1-12) and cecropin A (1-8) - melittin (1-12) hybrid peptides *J Pept Res* **53**: 578–589.

Orivel J, Redeker V, Le-Caer JP, Krier F, Revol-Junelles AM, Longeon A, Chaffotte A, Dejean A, Rossier J (2001) Ponericins, new antibacterial and insecticidal peptides from the venom of the ant *Pachycondyla goeldii*. *J Biol Chem* **276**: 17823–17829.

Park JM, Jung JE, Lee BJ (1995) Antimicrobial peptides from the skin of a Korean frog, *Rana rugosa*. *Biochem Biophys Res Commun* **205**: 948–954.

Popsueva AE, Zinovjeva MV, Visser JW, Zijlmans JM, Fibbe WE, Belyavsky AV (1996) A novel murine cathelin-like protein expressed in bone marrow. *FEBS Lett* **391**: 5–8.

Prado-Prado FJ, de la Vega OM, Uriarte E, Ubeira FM, Chou KC, Gonzalez-Diaz H (2007a) Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg Med Chem* **17**: 569–575.

Prado-Prado FJ, Gonzalez-Diaz H, Santana L, Uriarte E (2007b) Unified QSAR approach to antimicrobials. Part 2: Predicting activity against more than 90 different species in order to halt antibacterial resistance. *Bioorg Med Chem* **15**: 897–902.

Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC (2008) Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg Med Chem* **16**: 5871–5880.

Ramos de Armas R, González-Díaz H, Molina R, Uriarte E (2005) Stochastic-based descriptors studying biopolymers biological properties: extended MARCH-INSIDE methodology describing antibacterial activity of lactoferricin derivatives. *Biopolymers* **77**: 247–256.

Rosetto M, Manetti AG, Marchini D, Dallai R, Telford JL, Baldari CT (1993) Sequences of two cDNA clones from the medfly Ceratitis capitata encoding antibacterial peptides of the cecropin family. *Gene* **134**: 241–243.

Resch B (2004) Hidden Markov Models. A tutorial for the course computational intelligence. Signal processing and speech communication laboratory. http://www.igi.tugraz.at/lehre/CI/tutorials/HMM/HMM.pdf

Shin SY, Kang JH, Janq SY, Kim Y, Kim KL, Kahm KS (2000) Effects of the hinge region of cecropin A(1-8)-magainin 2(1-12), a synthetic antimicrobial peptide, on liposomes, bacterial and tumor cells. *Biochim Biophys Acta* **1463**: 209–218.

Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak, SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalon DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ, Marra MA (2002) Generation and initial analysis of more than 15000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci USA* **99**: 16899–16903.

Torres-Larios A, Gurrola GB, Zamudio FZ, Possani LD (2000) Hadrurin, a new antimicrobial peptide from the venom of the scorpion *Hadrurus aztecus*. *Eur J Biochem* **267**: 5023–5031.

Tossi A, Scocchi M, Zanetti M, Storici P, Gennaro R (1995) PMAP-37, a novel antibacterial peptide from pig myeloid cells. cDNA cloning, chemical synthesis and activity. *Eur J Biochem* **228**: 941–946.

Qu Z, Steiner H, Engström A, Bennich H, Boman HG (1982) Insect immunity: isolation and structure of cecropins B and D from pupae of the Chinese oak silk moth, *Antheraea pernyi*. *Eur J Biochem* **127**: 219–224.

Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Mendez-Alvarez E (2006) A QSAR model for *in silico* screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *Med Chem* **49**: 1149–1156.

Swiss, European Bioinformatics Institute 2006–2008. EBI is an Outstation of the European Molecular Biology Laboratory. http://www.ebi.ac.uk/swissprot/

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**: 320–322.

Spivak M (1965) *Calculus on Manifolds*. Benjamin. New York.

Tennessen JA, Blouin MS (2007) Selection for antimicrobial peptide diversity in frogs leads to gene duplication and low allelic variation. *J Mol Evol* **65**: 605–615.

Uniprot Swiss-prot ftp://ftp.expasy.org/databases/swiss-prot/release_compressed/<>uniprot_sprot.fasta.gz

Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**: 415–427.

Uversky VN (2002) What does it mean to be natively unfolded? *Eur J Biochem* **269**: 2–12.

Vizioli J, Bulet P, Hoffmann JA, Kafatos FC, Müller HM, Dimopoulos G (2001) Gambicin: a novel immune responsive antimicrobial peptide from the malaria vector *Anopheles gambiae*. *Proc Natl Acad Sci USA* **98**: 12630–12635.

Zimin AV, Smith DR, Sutton G, Yorke JA (2008) Assembly reconciliation. *Bioinformatics* **1**: 142–145.

Índice alfabético

Acta Biochimica Polonica

Vol. 56 No. 1/2009, 167–176

on-line at: www.actabp.pl

Regular paper

Detection of selective cationic amphipatic antibacterial peptides by Hidden Markov models

Carlos Polanco[1]✉ and Jose L. Samaniego[2]

[1]*Instituto de Fisiología Celular, and* [2]*Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior s/n Ciudad Universitaria Delegación Coyoacán, México*

Received: 30 January, 2009; revised: 03 March, 2009; accepted: 11 March, 2009
available on-line: 16 March, 2009

Antibacterial peptides are researched mainly for the potential benefit they have in a variety of socially relevant diseases, used by the host to protect itself from different types of pathogenic bacteria. We used the mathematical-computational method known as Hidden Markov models (HMMs) in targeting a subset of antibacterial peptides named Selective Cationic Amphipatic Antibacterial Peptides (SCAAPs). The main difference in the implementation of HMMs was focused on the detection of SCAAP using principally five physical-chemical properties for each candidate SCAAPs, instead of using the statistical information about the amino acids which form a peptide. By this method a cluster of antibacterial peptides was detected and as a result the following were found: 9 SCAAPs, 6 synthetic antibacterial peptides that belong to a subregion of Cecropin A and Magainin 2, and 19 peptides from the Cecropin A family. A scoring function was developed using HMMs as its core, uniquely employing information accessible from the databases.

Keywords: antibacterial peptides, Hidden Markov models

BACKGROUND

The increasing number of pathogens resistant to conventional antibiotics and the rising cost of production of the latter have led to the search for new drugs. One option for the development of these drugs is the production of antibacterial peptides found in nature, for these are the first defence line of living beings.

Antibacterial peptides have a wide variety of applications, from their use as antimicrobials to their use, after adaptations, as anticarcinogens (Ellerby *et al.*, 1999; Del Río *et al*., 2001) to human obesity control aids (Kolonin *et al*., 2004). It has also been observed that antibacterial peptides do not necessarily act exclusively against just bacteria. An example of a large non-specific antibacterial 85-peptide is gambicin: MKQQTVFVLLALLLVSASCVDALVYVYAKTC-STCRSLGARNCGYGSLGSKKYVSCDGATAIRNCD-DCRRRFGTCQDRYITECFIG-NH_2, which shows activity against bacteria and fungi (Vizioli *et al*., 2001).

The Selective Cationic Amphipatic Antibacterial Peptides (SCAAPs) are a recent and promising alternative for discovering new drugs effective in treating bacterial infections. They are characterized by being less than 60 amino acids in length, not adopting an α-helicoidal structure in neutral pH water solution and having a *therapeutic index* higher than 75 (Del Río *et al*., 2001). The therapeutic index of a peptide is defined (Ellerby *et al.*, 1999; Del Río *et al*., 2001) as the ratio between the minimum inhibitory concentrations observed against mammalian and bacterial cells: the higher the value, the more specific the peptide for bacterial-like membranes. In other words, SCAAPs display strong lytic activity against bacteria, but have no toxicity against normal eukaryotic cells such as erythrocytes (Shin *et al*., 2000).

Computer-based approaches may accelerate the discovery of new SCAAPs. However, detection of SCAAPs among every possible antibacterial peptide is not feasible either computationally or by biological assays. Their variation is 20^n where $n \in N$ is the

✉Corresponding author: Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Circuito Exterior s/n Ciudad Universitaria Delegación Coyoacán CP 04510, D.F., México; e-mail: polanco@unam.mx

length of peptide. For instance, an improved version of our program APAP (Del Río *et al*., 2001) executed on a cluster of 100 CPUs can not evaluate more than 20^{13} sequences of length 13 aa; it takes more than 10 months of processing time in a single PC (not shown). APAP-I, as well as APAP, evaluates the following physical-chemical properties for each peptide: isoelectric point (IP), average helical hydrophobic moment (HM), mean hydrophobicity (MH), mean net charge (MC) and AGADIR (helix/coil transition algorithm). APAP-I is 396000 times more efficient than the program APAP because it was designed to run on a high performance computing platform, and oriented to evaluate short peptides (8–11 aa). Thus, identification of new SCAAPs by searching the full space of peptide sequences may not be practical.

An alternative approach would be to search for new SCAAPs in sequences likely to have antibacterial activity. In this regard, it is possible to search for SCAAPs in peptides obtained from venoms (Conde *et al*., 2000) or to identify sequence patterns present in known antibacterial peptides. To identify such patterns, Hidden Markov Models (HMMs) provide a theory for profile methods (Resch, 2004; Prado-Prado *et al*., 2007a; 2007b). These HMMs may be used to predict new antibacterial peptides based on numeric indices of the peptide.

This type of study is known in the literature as Quantitative Structure-Activity Relationships (QSAR) or more generic Quantitative Structure-Property Relationships (QSPR) models. In fact, not only HMMs but other types of Markov models have been largely used to seek QSAR (quantitative structure-activity relationships)/QSPR (quantitative structure-property relationships) (González-Díaz *et al*., 2007f). For instance, the MARCH-INSIDE approach (Markov Chains Invariants for Network Simulation and Design) introduced by González-Díaz and coworkers makes use of Markov Chains theory to infer QSAR/QSPR models at different structural levels. Applications range from QSAR models of low-molecular-weight drugs (Santana *et al*., 2006; Cruz-Monteagudo *et al*., 2007; González-Díaz *et al*., 2007b; 2008b; Prado-Prado *et al*., 2008), to QSAR/QSPR models for protein and nucleic acid sequences (Aguero *et al*., 2008a; 2008b), protein 3D structure (González-Díaz *et al*., 2007a; 2007c; 2007d), RNA secondary structures (González-Díaz *et al*., 2003b; 2005; 2007e), viral surfaces (González-Díaz *et al*., 2003a) and of course peptides (Ramos de Armas *et al*., 2005).

The idea has been extended to include also Quantitative Proteome-Property Relationship (QPPR) models that personalize predictions of drug cardiotoxicity (González-Díaz *et al*., 2008a; 2008b; 2008c), or human prostate cancer (Ferino *et al*., 2008; González-Díaz *et al*., 2009), based on protein composition of Blood Proteomes. These Markov methods use different types of transition probabilities described by atom-atom, nucleotide-nucleotide, amino acid-amino acid, or even protein-protein matrices. Two recent in-depth reviews of the field were published recently (González-Díaz *et al*., 2008a; 2008c).

This article presents an approximation by Hidden Markov Models to detect SCAAPs based on physical-chemical similarity. As previously described (Del Río *et al*., 2001) the advantage of HMMs for this purpose is that they may identify patterns not obvious from iterative approaches such as APAP. This in turn may accelerate the discovery of new SCAAPs.

HMMs were implemented by using four sets of antibacterial peptides and one set of proteins:

Set A: 59 natural and synthetic antibacterial peptides extracted from (**set C**), which act exclusively against bacteria, fungi, viruses and mammalian cancer cells, with 3D structure determined by NMR spectroscopy or X-ray diffraction (NCBI, September, 2007).

Set B: 28 natural and synthetic antibacterial peptides extracted from (**set C**), which act exclusively against bacteria, with their 3D structure were detected by NMR spectroscopy or X-rays (NCBI, September, 2007).

Set C: 500 natural and synthetic antibacterial peptides which have a non-specific action against bacteria. The method used to predict the 3D structure is not relevant (NCBI, September, 2007).

Set D: 3 natural and synthetic antibacterial peptides extracted from (**set C**): Gambicin; Mellitin and Temporin H (XXA, frog) (NCBI, September, 2007).

Set E: 391836 natural and synthetic proteins detected in nature (Uniprot, August, 2008).

A *stochastic process* is a mathematical model for any phenomenon evolving or varying in time (or space etc.) subject to random influences (e.g., the stock market price of a commodity observed in time, the distribution of colors or shades in a noisy picture observed in an unordered two-dimensional lattice etc.).

Markov Models. Introduction

The condition prediction H at the time $t \in N$ is concerned with hypothesizing what the condition H will be at the time $t+1$, based on the observations of the condition H in the past (Resch, 2004).

We collected the relative frequency on the condition h_i (on time i) depending on what the condition H was like one day earlier h_{i-1}, the day before that h_{i-2}, and so forth.

The conditional probability is

$$\mathbf{P}\{h_n \mid h_{n-1}\} = \mathbf{P}\{h_n \mid h_{n-1}, h_{n-2}, \ldots, h_1\}$$

However, the larger the value of i is, the more observations we must collect. For n states of

the condition H the number of past histories will be $|H|^{n-1}$.

If we take the Markov assumption, we would have the probability of an observation at time i depend on h_{-1}. So we can express the probability of a sequence $\{h_1,...,h_n\}$ using this assumption:

$$\mathbf{P}\{h_1,\ldots,h_n\} = \mathbf{P}\{h_1\}\prod_{i=2}^{n}\mathbf{P}\{h_i \mid h_{i-1}\} \quad (1)$$

As a consequence of the Markov assumption, the number of past histories is reduced to $h_n \times h_{n-1}$.

HMMs. Mathematical description

If A, B are two events, then we define the probability of A given B as

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \quad (2)$$

One can work in the *mathematical ideal world* with the probability $\overline{\mathbf{P}}$ to achieve various mathematical objectives, and then reinterpret these results back in the *real world* with a measure change back to $\mathbf{P}$ *via* the inverse Radon-Nikodym derivative.

If circumstances only allow us to obtain the condition H based on another condition O, the condition H is hidden from us. We evaluate the conditional probability $\mathbf{P}(h_i|o_i)$ according to Eqn. (2).

$$\mathbf{P}(h_i \mid o_i) = \frac{\mathbf{P}(o_i \mid h_i)\mathbf{P}(h_i)}{\mathbf{P}(o_i)}$$

If we assume that, for all i the H_i, O_i, are independent of all o_j, h_j, for all $i \neq j$, Eqn. (1) can be rewritten as

$$\mathrm{L}(h_1,\ldots,h_n \mid o_1,\ldots,o_n) \propto \mathbf{P}(h_1,\ldots,h_n \mid o_1,\ldots,o_n) = \prod_{i=1}^{n}\mathbf{P}(o_i \mid h_i).\prod_{i=1}^{n}\mathbf{P}(h_i \mid h_{i-1}) \quad (3)$$

Eqn. (3) is known as a measure of the probability and is referred to as the *likelihood* function L.

The expectation maximization (EM) algorithm reestimates the parameters of the model.

Many of the density functions are exponential in nature; it is therefore easier to compute the EM of a likelihood function by finding the maximum of the *natural ln* of L, known as the *ln-likelihood* function:

$$l(h_i \mid o_i) = \ln(\mathrm{L}(h_i \mid o_i))$$

due to the monotonicity of the *ln* function.

HMMs. Terminology

HMMs are specified by the set of states $S = \{s_1,s_2,...,s_n\}$, corresponding to the possible condition H, and the parameter set $\Omega = \{\pi, A,B\}$:

The **initial probabilities** $\pi_i = \mathbf{P}(h_i = s_i)$ are probabilities of s_i being the first state of a state sequence h_i. They are collected in the vector P_0.

The **transition probabilities** are the probabilities that go from state i to state j: $a_{i,j} = \mathbf{P}(h_n = s_j) \mid h_{n-1} = s_i)$. They are collected in matrix A.

The **emission probabilities** characterize the likelihood of a discrete observation $o_n \in \{v_i,...,v_n\}: b_{i,k} = \mathbf{P}(o_n = v_k \mid h_n = s_i)$, and the probabilities to observe v_k if the current state is $h_n = s_i$. The numbers $b_{i,k}$ are gathered in matrix B.

The likelihood of $O = \{o_1,...,o_n\}$ along the path $H = \{h_1,...,h_n\}$ determined from HMMs with parameters Ω, is given by:

$$\mathrm{L}(h_i \mid o_i) \propto \mathbf{P}(O,H \mid \Omega) = \prod_{i=1}^{n}\mathbf{P}(O \mid H,\Omega)\prod_{i=1}^{8}\mathbf{P}(H \mid \Omega) \quad (4)$$

where the probabilities $\mathbf{P}(O|H,\Omega)$ and $\mathbf{P}(H|\Omega)$ are expressed in terms of matrices A, B (Eqns. 5 and 6) and the vector P_0.

$$\mathbf{P}(O \mid H,\Omega) = \prod_{i=1}^{n}\mathbf{P}(H,\Omega) = b_{h_1,o_1}\, b_{h_2,o_2} \ldots b_{h_n,o_n} \quad (5)$$

$$\mathbf{P}(H \mid \Omega) = \pi_{h_1}\prod_{i=1}^{8} a_{h_i,h_{i+1}} = \mathrm{P}_{0h_i}\, a_{h_1,h_2}\, a_{h_2,h_3} \ldots a_{h_7,h_8} \quad (6)$$

$\mathbf{P}(O,H|\Omega)$ (Eqn. 4) is known as the joint *likelihood* of an observation sequence and it is equivalent to Eqn. (1).

HMMs. Implementation

The set of states S corresponding to the twenty different amino acids from which every antibacterial peptide is formed: S = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}, and the parameter set was formed by $\Omega = \{P_0,A,B\}$.

The **vector** p_0 contains $\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_{0i}$, where n is the length of the peptide to be tested, and p_{0i} is the relative frequency distribution of amino acids from the same peptide, derived from the absolute frequency distribution from natural and synthetic antibacterial peptides from **(set A)** (Table 1). Their 3D structure was detected by NMR spectroscopy or X-rays dif-

Table 1. Elements of vector P_0.

Absolute frequency of natural and synthetic antibacterial peptides which act exclusively against bacteria, fungi, viruses and mammalian cancer cells **(set A)** to vector P_0. The letters in the table refer to the 20 amino acids (one-letter code), and the numbers represent the corresponding frequency of that amino acid in the set.

A	C	D	E	F	G	H	I	K	L	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
103	132	23	32	61	182	39	129	146	101	9	52	67	49	135	87	53	85	31	57

Table 2. Elements of matrix A.

Absolute frequency distribution of all amino acids taken of pairs (contiguously), from **(set C)**. Every letter is equivalent to each amino acid, in this manner, the occurrence of pair of amino acids (A_{ci}; A_{cj}) is built with the amino acid from row ($_i$) and the amino acid from column ($_j$).

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	165	38	11	29	28	134	17	85	205	11	66	40	15	24	34	76	50	64	9	2
C	50	87	15	7	34	32	24	30	85	43	4	17	41	7	142	41	30	43	4	54
D	24	17	8	3	15	10	2	32	25	27	6	2	4	7	4	16	20	28	9	9
E	14	8	5	8	6	20	15	9	48	24	4	7	4	8	44	19	8	13	3	3
F	27	69	17	4	23	48	14	25	62	109	1	14	33	18	49	34	9	26	3	10
G	103	51	19	39	61	10	25	137	164	185	15	39	74	55	102	62	64	89	33	53
H	14	18	5	11	15	18	19	17	15	29	6	3	10	4	25	19	14	57	0	4
I	95	40	13	25	43	143	31	53	108	69	10	31	53	21	59	68	28	45	6	19
L	105	55	34	25	60	155	24	55	143	129	9	23	108	26	65	80	22	51	28	10
M	15	4	6	5	2	11	0	6	12	18	0	8	1	5	12	6	2	7	1	2
N	30	17	6	8	27	37	10	14	29	36	11	9	23	4	36	12	23	36	3	7
P	43	16	8	7	45	47	16	79	45	36	6	21	55	17	69	30	18	64	13	17
Q	44	23	3	4	12	47	12	27	24	7	5	12	21	20	16	8	16	16	4	2
R	40	62	32	17	45	94	23	60	73	69	7	48	97	30	118	35	20	54	20	21
S	63	67	13	16	20	78	21	43	74	52	14	15	12	17	33	31	28	52	10	15
T	51	79	8	5	17	34	5	38	29	47	8	4	14	15	44	11	13	38	4	13
V	107	59	14	13	36	133	13	49	63	103	2	22	47	11	46	47	32	60	8	12
W	15	8	5	8	5	16	3	9	31	22	3	14	9	9	5	5	2	4	4	0
Y	10	51	3	2	6	30	1	13	22	20	1	11	9	5	39	14	19	11	0	8

fraction, and was taken from the database BBCM (NCBI, September, 2007).

The **matrix A** represents the relative frequency of all 400 possible pairs of amino acids. These pairs were taken in two directions: ($a_{i,j}$,$a_{i+1,j}$) and ($a_{i-1,j}$,$a_{i,j}$), for specific *j*. The matrix was built from natural and synthetic antibacterial peptides which have non-specific action against bacteria **(set C)**; the method used to predict the 3D structure is not relevant (Table 2). These peptides were taken from the database BBCM (NCBI, September, 2007).

Every pair of amino acids from the peptide to be tested was extracted from **matrix A**.

The **matrix B** exhibits the conditional probability of the peptide to be tested as the result of two conditions: first, the calculation of each natural and synthetic antibacterial peptide by program APAP-I (this program evaluated if the peptide is or is not a candidate SCAAP); second, if the $\text{Index}_A \geq 0.08$.

Index A (Eqn. 7) is formed by the relative frequency distribution of amino acid A_i from the peptide to be tested, derived from the absolute frequency distribution from natural and synthetic antibacterial peptides which act exclusively against bacteria **(set B)** (Table 3). (NCBI, September, 2007).

$$\textbf{Index}_{\mathbf{A}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{A}_i, i \in [1,n] \qquad (7)$$

The program APAP-I was used to evaluate if a peptide to be tested from **(set C)** was a candidate SCAAP or not, with the evaluation of different physical-chemical properties. APAP-I is formed by two subprograms:

APAP-IA which evaluated the isoelectric point IP, helical hydrophobic moment HM and AGADIR.

APAP-I-B which evaluated the isoelectric point IP, helical hydrophobic moment HM, mean hydrophobicity MH and mean net charge MC.

The physical-chemical properties in acceptable ranges were:

Isoelectric point (IP) (Del Rio *et al*., 2001). This is the pH at which a particular peptide carries no net electrical charge. The value range considered was from 10.8 to 11.8.

Helical hydrophobic moment (HM) (Eisenberg *et al*., 1982). This is a sum of the hydrophobicities of the side chains of a helix of *n* amino acids. The length of a vector corresponding to the hydrophobicities is the numerical hydrophobicity associated to the kind of side chain, and its direction is determined by the orientation of the side chain according to the helix axis. A large value of HM means that the helix is amphiphilic perpendicular to its axis. The value range considered was from 0.4 to 0.6.

Table 3. Elements of vector Index_A.

Absolute frequency of natural and synthetic antibacterial peptides which act exclusively against bacteria **(set B)** to vector Index_A. The letters in the table refer to the 20 amino acids (one-letter code), and the numbers represent the corresponding frequency of that amino acid in the set.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
49	68	10	16	35	106	22	71	112	52	2	30	33	19	60	42	24	45	17	27

Mean hydrophobicity (MH) (Del Río *et al*., 2001). This is the mean of the hydrophobicities of the amino acids normalized to 1 over all amino acids of the peptide. The algorithm was given by the technical department of the Swiss Institute of Bioinformatics (Swiss). The value range considered was from 0.35 to 0.55.

Mean net charge (MC) (Del Río *et al*., 2001). This is determined by Eqn. (8). The algorithm was given by Uversky (Uversky, 2000; Uversky *et al*., 2002).

$$\mathrm{MC}(R,K,D,E) = \frac{1}{n}(R_i + K_i - D_i - E_i), i \in [1,n] \quad (8)$$

The variables R_i, K_i, D_i and E_i represent the number of times the amino acids arginine (R), lysine (K), aspartic acid (D) and glutamic acid (E) appeared, accepting those peptides whose *MC*(*R,K,D,E*) evaluated with Eqn. (8) are above or equal to the number obtained by Eqn. (9) with the same mean hydrophobicity (MH).

$$\mathrm{MC}(MH) = 45.896MH^4 - 47.528MH^3 + 13.324MH^2 + 2.302MH - 1.291 \quad (9)$$

AGADIR (Lacroix *et al*., 1997; Del Río *et al*., 2001). Predicts the helical behaviour of a peptide. The value range considered was from 0.00 to 10.00.

The **matrix B** shows the conditional probability of $\mathbf{P}(o_i | h_{i?\mathrm{IndexA}})$ to be candidate SCAAPs if (o_i = *true*) the $\mathbf{P}(o_i = true | h_{i?\mathrm{IndexA}}) = 0.95$, and its complement ($o_i$ = *false*) $\mathbf{P}(o_i = false | h_{i?\mathrm{IndexA}}) = 0.05$. These numbers are obtained as a result of many computational assays.

HMMs. Tests

As a **negative test**, the validation of HMMs to detect candidate SCAAPs consisted of testing:

The total number of natural and synthetic antibacterial peptides which had a non-specific action and whose structure could not be determined by either method **(set C)** (i.e. NMR spectroscopy or X-rays) over two sets:

A set of three natural and synthetic antibacterial peptides **(set D)**: Gambicin characterized by non-specific action and no SCAAPs (according to the program APAP-I); Mellitin characterized by toxicity against erythrocytes; Temporin [H XXA, frog] was determined by circular dichroism (CD).

The total number of natural and synthetic proteins that were detected in nature **(set E)** were used to build the matrices A and B, and test the **(set C)**.

HMMs. Statistical analysis

A two-sample rank test by Wilcoxon, Mann and Whitney (Kreyszig, 1979) was made to test over two populations:

Natural and synthetic antibacterial peptides **(set C)** *versus* natural and synthetic antibacterial peptides which act exclusively against bacteria **(set B)**.

Natural and synthetic antibacterial peptides with an exclusive action against bacteria **(set B)** *versus* natural and synthetic antibacterial peptides detected by program APAP-I.

These statistical tests were used to verify the hypothesis that two populations have the same distribution to be a candidate SCAAPs or not. The assumption was that the populations tested correspond to continuous distributions, and to obtain critical values c_1 and c_2, using the fact that if the hypothesis is true, then the random variable *W*, over the populations described is approximately normal with mean and variance (Eqns. 10 and 11)

$$\mu_W = \frac{n_1(n_1 + n_2 + 1)}{2} \quad (10)$$

$$\sigma_W^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \quad (11)$$

Hence c_1 and c_2 were obtained substituting μ_W and σ_W in Eqns. (12) and (13)

$$P(W \le c_1) = \Phi\left(\frac{c_1 - \mu_W}{\sigma_W}\right) = 2.5\% \quad (12)$$

$$P(W \ge c_2) = 1 - \Phi\left(\frac{c_2 - \mu_W}{\sigma_W}\right) = 2.5\% \quad (13)$$

The test was conducted only on the **(sets B** and **C)** because this pair is more similar than the other sets involved **(A, D** and **E)**.

RESULTS

Objective

The use of HMMs for prediction and understanding of antimicrobial peptides has been reported for the last three decades (Andrés & Dimarcq, 2007), particularly the detection of antimicrobial peptides by multivariate linear regression and physical-chemical properties (Hilpert *et al*., 2008).

In this article we use HMMs for the prediction of candidate SCAAPs based on five physical chemical properties: isoelectric point (IP), helical hydrophobic moment (HM), mean hydrophobicity (MH), mean net charge (MC), and AGADIR; and the relative frequency distribution of single and pair amino acids over the sequence of the peptide.

Identification of SCAAPs

We retrieved a cluster of 57 natural and synthetic antibacterial peptides (Table 4) which act ex-

Table 4. Cluster of antibacterial peptides predicted by HMMs and listed in descending order (set C)

NL: Position of the antibacterial peptide on the list. NP: Number which corresponds to the antibacterial peptide according to HMMs. F: Family. If natural SCAAPs were a part of **(set B)**, [*S*]. If Brevinin, [*B*]. If Cathelin, [*Ca*]. If Cecropin, [*C*]. If Moricin, [M]. AP-A: Peptide which was accepted by the program APAP-IA (Section HMMs. Implementation). AP-B: Peptide which was accepted by the program APAP-IB (Section HMMs. Implementation)

NL	NP	F	AP-A	AP-B	Name of the sequence	References
1	454	*C*	+	+	Cecropin-B type 1 precursor (Cecropin-B1) [Contains: Cecropin-B (AalCecB); Cecropin-B amidated isoform]	Sun *et al.*, 1999
2	417		+	+	Parabutoporin	Moerman *et al.*, 2002
3	16	*S;C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P3)	Oh *et al.*, 1999
4	61	*C*	+	+	Cecropin B [Bombyx mori]	Taniai *et al.*, 1995
5	458	*S*	+	+	Cathelin-like protein [Mus musculus]	Popsueva *et al.*, 1996
6	172	*C*	+	+	Hyphancin-3D precursor (Hyphancin-IIID) (Cecropin-A)	
7	15	*S;C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12)	
					Hybrid Peptide Analogue(P4)	Oh *et al.*, 1999
8	58 C		+	+	Cecropin-B	Ku *et al.*, 1982
9	174	*C*	+	+	Hyphancin-3F precursor (Hyphancin-IIIF) (Cecropin-A2)	
10	68		+	+	Defensin NP-3a [Oryctolagus cuniculus]	Linzmeier *et al.*, 1993
11	57	*S*	+	+	Cecropin-A precursor (Cecropin-C)	Gudmundsson *et al.*, 1991
12	425	*C*	+	+	RecName: Full=Cecropin-A	
13	175	*C*	+	+	Hyphancin-3G precursor (Hyphancin-IIIG) (Cecropin-A3)	
14	259	*C*			Cecropin-A1 precursor (Cecropin-A) (AalCecA)	Sun *et al.*, 1999
15	356				Ranatuerin-2Lb	Soraya *et al.*, 2000
16	173	*C*	+	+	Hyphancin-3E precursor (Hyphancin-IIIE) (Cecropin-A1)	
17	176	*Ca*	+	+	Cathelin-like protein [Mus musculus]	Popsueva *et al.*, 1996
18	474				Sentrin/SUMO-speci_c protease [Plasmodium yoelii yoelii str. XNL]	Carlton *et al.*, 2002
19	67		+	+	Defensin NP-3a [Oryctolagus cuniculus]	Linzmeier *et al.*, 1993
20	32	*S;M*			Chain A, Solution Structure of Antibacterial Peptide (Moricin)	Hemmi *et al.*, 2002
21	52				M Moricin [Bombyx mori].	Hemmi *et al.*, 2002
22	9	*S;C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide	Oh *et al.*, 1999
23	74				GK14120 [Drosophila willistoni]	Zimin *et al.*, 2008
24	426	*M*			RecName: Full=Virescein.	
25	106		+	+	Xenopsin precursor protein [Xenopus laevis]	Moore *et al.*, 1991
26	169		+	+	Antibacterial peptide PMAP-37 precursor (Myeloid antibacterial peptide 37)	Tossi *et al.*, 1995 Tossi *et al.*, 1995
27	435	*C*			Cecropin 1 [Musca domestica]	Tossi *et al.*, 1995
28	424	*C*			Cecropin precursor	Tossi *et al.*, 1995
29	75				Sarcotoxin-1B precursor (Sarcotoxin IB)	Kanai *et al.*, 1989
30	355		+	+	Hadrurin	Torres-Larios *et al.*, 2000
31	434	*C*			Cecropin-1 precursor	Rosetto *et al.*, 1993
32	493	*C*	+	+	Chain A, Solution Structure of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P2)	Oh *et al.*, 1999
33	490	*C*	+	+	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12)	
					Hybrid Peptide Analogue(P3)	Oh *et al.*, 1999
34	14	*S;C*	+	+	Chain A, Solution Structure Of Cecropin A(1-8)-Magainin 2(1-12) Hybrid Peptide Analogue(P2)	Oh *et al.*, 1999
35	469		+	+	Ribosomal protein L1 [Helicobacter pylori G27]	
36	386	*B*			Brevinin-1SY	Matutte *et al.*, 2000
37	102				Megakaryocyte stimulating factor [Trichomonas vaginalis G3]	Carlton *et al.*, 2002
38	127	*B*			RecName: Full=Brevinin-1	Morikawa *et al.*, 1992
39	405				Maximin-H14 antimicrobial peptide precursor [Bombina maxima]	
40	267				Neutrophil defensin 3 (HANP-3)	Mak *et al.*, 1996
41	265				Neutrophil defensin 1 (HANP-1)	Mak *et al.*, 1996
42	380	*B*			Brevinin 1Pb precursor [Rana pipiens]	Tennessen *et al.*, 2007
43	119	*C*			Cecropin C CG1373-PA [Drosophila melanogaster]	Hoskins *et al.*, 2007
44	56				Bombinin	
45	263				Fabatin precursor [Vicia faba]	
46	262				Fabatin precursor [Vicia faba]	
47	125	*B*			Brevinin-1E	Marenah *et al.*, 2006
48	346				Ponericin-W2	Orivel *et al.*, 2001
49	345				Ponericin-W1	Orivel *et al.*, 2001
50	132				Ceratotoxin A [Ceratitis capitata]	Rosetto *et al.*, 1993
51	239				Gaegurin-6	Park *et al.*, 1995
52	488				Nigrocin-2P precursor [Rana palustris]	
53	137				Ranalexin precursor	Clark *et al.*, 1994
54	392				Temporin-1Ca	Halverson *et al.*, 2000
55	19	*S;Ca*			Cathelin-related peptide SC5 precursor 1 (Antibacterial peptide SMAP-29) (Myeloid antibacterial peptide MAP-29)	Mahoney *et al.*, 1995
56	112				Defensin related cryptdin 4 [Mus musculus]	Strausberg *et al.*, 2002
57	459	*S;Ca*			Cathelin-like protein [Mus musculus]	Popsueva *et al.*, 1996

clusively against bacteria, fungi, viruses and mammalian cancer cells, whose 3D structure was determined by NMR spectroscopy or X-rays from the BBCM protein database (NCBI, September, 2007) **(set C)**. From this set we generated one subset, according to their structure: 28 antibacterial peptides determined by NMR spectroscopy **(set B)**. An HMM profile of the SCAAP family was built from these sets. After calibration, the HMMs were used to search through 500 natural and synthetic antibacterial peptides which have a non-specific action against bacteria (NCBI, September, 2007) **(set C)**; nine hits were found from the search on 500 antibacterial peptides (9, 14, 15, 16, 19, 32, 57, 458 and 459), six synthetic antibacterial peptides were found in Cecropin A and Magainin 2 (3, 9, 14, 15, 490 and 493), 19 peptides were from the Cecropin A family (9, 14, 15, 16, 58, 61, 119, 172, 173, 174, 175, 259, 424, 425, 434, 435, 454, 490 and 493); four peptides were from the Brevinin family (125, 127, 380 and 386), three peptides from the Cathelin family (19, 176 and 459), and two peptides from the Moricin family (32 and 52).

The entire cluster was further analyzed by a search against Swiss-Prot and Translated EMBL protein databases by Smith-Waterman algorithm on GCG/SeqWeb to ensure the identification of these peptides. They are described in Table 4.

Note that the peptide number 32 (position 20 in Table 4) was not accepted by the programs APAP-IA and APAP-IB, but it was accepted by HMMs because of its score.

Negative tests of HMMs

HMMs were tested with:

Three peptides: Gambicin characterized by non-specific action against bacteria, fungi, viruses and mammalian cancer cells; Mellitin characterized by toxicity against erythrocytes; and Temporin H [XXA, frog] determined by circular dichroism (CD). All peptides were accepted by HMMs.

As a full test, we retrieved the complete set of proteins (391 836) from the Uniprot protein database and a new HMM profile was built from these sequences. After calibration, the new HMMs were used with the same set of 500 natural and synthetic antibacterial peptides **(set C)** that we refer to in the identification of SCAAPs in Table 4: No candidate SCAAPs or SCAP family was detected.

Statistical verification of HMMs

In order to verify if a statistical similarity exists between the referred set of peptides involved in the tests, we decided to compare only the more biologically similar sets: the set of 500 natural and synthetic antibacterial peptides which have non-specific action against bacteria **(set C)**, and the set of 28 natural and synthetic antibacterial peptides which act exclusively against bacteria, with their 3D structure detected by NMR spectroscopy or X-ray diffraction **(set B)**.

We ran a Wilcoxon, Mann and Whitney non-parametric test (with $p\text{-}value < 0.05$): the test did not observe any normal correlation between those sets, and consequently it was concluded that no sets had any statistical relation.

DISCUSSION

In this article, we have described the detection of nine SCAAPs by applying a mathematical-computational tool, the HMM search on a predicted peptide database. Compared with the experimental assay search, the HMM is much more sensitive due to its summarizing nature. The key point for a successful HMM search lies in constructing the HMMs profile (a combination of physical-chemical properties and relative frequency distribution of amino acids over the sequence of the peptide). The inclusion of the complete set of proteins from the Uniprot protein database in order to reconfigure HMMs, and the inclusion of three wrong sequences provides more reliability and robustness of this HMM profile.

We recognize some bias with this approach. The major issue is related to the incompleteness of the existing databases. The degree to which the current database is complete is not known, even though our studies are designed to be exhaustive.

While this manuscript was being prepared, a paper was in press that described the detection of short linear cationic antimicrobial peptides using, principally, the nonlinear techniques of support vector machines and artificial neural networks (Hilpert *et al*., 2008). Their methods are more selective and less comprehensive than HMMs described. Thus, these two approaches could be used as complementary tools in identifying novel candidate members of a specific protein family.

Comparative studies

Our HMMs profile was compared with three stochastic methods named HMMER (HMMER), MAST (Bailey & Gribskov, 1998; 2000) and GLAM (Frith *et al*., 2004). These comparisons were concerned with the number of hits each method offers, and the results show that GLAM was superior to the other methods but that HMMs, MAST and HMMER were equally effective.

CONCLUSIONS

The HMMs profile is a mathematical-computational tool for finding potential peptides named Selective Cationic Amphipatic Antibacterial Peptides (SCAAPs) solely by employing information accessible from the databases to provide adequate peptide identification performance. It allows rapid, convenient searches within databases. In summary, HMMs profiles show significant selective efficacy in the detection of SCAAPs, and are a useful model for biological sequence analysis and modeling in the postgenomic era.

Acknowledgements

The authors thank Gabriel del Rio (IFC-UNAM) for critical comments on this article, and Dr. Luis A. Rincon (FC-UNAM) for valuable suggestions to the manuscript. To Instituto de Fisiologia Celular (IFC) and Departamento de Matematicas of the Facultad de Ciencias (FC) for the backing given to the elaboration of this article. To Doctorado en Ciencias Biomedicas and Universidad Nacional Autonoma de Mexico (UNAM) for supporting this research. The authors thank Dagny Valadez and Carlos Warden (FFyL UNAM) for the proof reading of this article.

REFERENCES

Aguero-Chapin G, Antunes A, Ubeira FM, Chou KC, Gonzalez-Diaz H (2008a) Comparative Study of topological indices of macro/supramolecular RNA complex networks. *J Chem Inf Model* **48**: 2265–2277.

Aguero-Chapin G, Gonzalez-Diaz H, de la Riva G, Rodriguez E, Sanchez-Rodriguez A, Podda G *et al* (2008b) MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from *Schizosaccharomyces pombe*, prediction, and experimental assay of a new sequence. *J Chem Inf Model* **48**: 434–448.

Andrés E, Dimarcq JL (2007) Cationic antimicrobial peptides: from innate immunity study to drug development. Update. *Med Mal Infect* **37**: 194–199.

Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**: 48–54.

Bailey TL, Gribskov M (2000) Concerning the accuracy of MAST E-values. *Bioinformatics* **16**: 488–489.

Boulanger N, Brun R, Ehret-Sabatier L, Kunz C, Bulet (2002) Immunopeptides in the defense reactions of Glossina morsitans to bacterial and *Trypanosoma brucei brucei* infections. *Insect Biochem Mol Biol* **32**: 369–375.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, Peterson JD, Pop M, Kosack DS, Shumway MF, Bidwell SL, Shallom SJ, van Aken SE, Riedmuller SB, Feldblyum TV, Cho JK, Quackenbush J, Sedegah M, Shoaibi A, Cummings LM, Florens L, Yates JR, Raine JD, Sinden RE, Harris MA, Cunningham DA, Preiser PR, Bergman LW, Vaidya AB, van Lin LH, Janse CJ, Waters AP, Smith HO, White OR, Salzberg SL, Venter JC, Fraser CM, Hoffman SL, Gardner MJ, Carucci DJ (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**: 512–519.

Clark DP, Durell S, Maloy WL, Zasloff M (1994) Ranalexin. A novel antimicrobial peptide from bullfrog (Rana catesbeiana) skin, structurally related to the bacterial antibiotic, polymyxin. *J Biol Chem* **269**: 10849–10855.

Conde R, Zamudio FZ, Rodríguez MH, Possani LD (2000) Scorpine, an anti-malaria and anti-bacterial agent purified from scorpion venom. *FEBS Lett* **471**: 165–168.

Cruz-Monteagudo M, González-Díaz H, Aguero-Chapin G, Santana L, Borges F, Domínguez RE *et al.* (2007) Computational chemistry development of a unified free energy Markov Model for the distribution of 1300 chemicals to 38 different environmental or biological systems. *J Comput Chem* **28**: 1909–1922.

Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MNDS, Uriartei E, Chou K-C *et al*. (2008a) Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case polymer. *Polymer* **49**: 5575–5587.

Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MN, Uriarte E, Gonzalez-Diaz H (2008b) Quantitative Proteome-Property Relationships (QPPRs). Part 1: finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra. *Bioorg Med Chem* **16**: 9684–9693.

Cruz-Monteagudo M, González-Díaz H, Borges F, Dominguez ER, Cordeiro MN (2008c) 3D-MEDNEs: An alternative *in silico*. Technique for chemical research in toxicology. 2. Quantitative Proteome-Toxicity Relationships (QPTR) based on mass spectrum spiral entropy. *Chem Res Toxicol* **21**: 619–632.

Del Río G, Castro-Obregon S, Rao R, Ellerby MH, Bredesen DE (2001) APAP, a sequence-pattern recognition approach identifies substance P as a potential apoptotic peptide. *FEBS Lett* **494**: 213–219.

Eisenberg D, Weiss RM, Terwilliger TC (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **23**: 371–374.

Ellerby MH, Arap W, Kain R, Andrusiak R, Del Río G, Krajewski S, Lombardo CR, Rao R, Ruoslahti E, Bredesen DE, Pasqualini R (1999) Anti-cancer activity of targeted pro-apoptotic peptides. *Nat Med* **5**: 1032–1038.

ExPASy Proteomics Server. http://www.expasy.ch/sprot

Ferino G, Gonzalez-Diaz H, Delogu G, Podda G, Uriarte E (2008) Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer. *Biochem Biophys Res Commun* **372**: 320–325.

Ferino G, Delogu G, Podda G, Uriarte E, González-Díaz H (2009) Quantitative proteome-disease relationships (QPDRs) in clinical chemistry: prediction of prostate cancer with spectral moments of PSA/MS star networks. In *Clinical Chemistry Research*; Mitchem BHaS, ChL, ed. NY: Nova Science Publisher.

Frith CM, Hansen U, Spouge JL, Weng Z (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res* **32**: 189–200.

González-Díaz H, Molina RR, Uriarte E (2003a) Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropie. *Polymer* **45**: 3845–3853.

González-Díaz H, de Armas RR, Molina R (2003b) Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Psi-RNA

packaging region with drugs. *Bioinformatics* **19**: 2079–2087.

González-Díaz H, Pérez-Bello A, Uriarte E (2005) Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA-QSAR for mycobacterial promoters. *Polymer* **46**: 6461–6473.

González-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E (2007a) A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *J Proteome Res* **6**: 904–908.

Gonzalez-Diaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A (2007c) Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. *Comput Chem* **28**: 1042–108.

González-Díaz H, Pérez-Castillo Y, Podda G, Uriarte E (2007d) Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. *J Comput Chem* **28**: 1990–1995.

González-Díaz H, Aguero-Chapin G, Varona J, Molina R, Delogu G, Santana L *et al*. (2007e) 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. *J Comput Chem* **28**: 1049–1056.

González-Díaz H, Vilar S, Santana L, Uriarte E (2007f) Medicinal chemistry and bioinformatics — current trends in drugs discovery with networks topological indices. *Curr Top Med Chem* **7**: 1025–1039.

Gonzalez-Diaz H, Prado-Prado F, Ubeira FM (2008a) Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr Top Med Chem* **8**: 1676–1690.

González-Díaz, Prado-Prado F (2008b) Unified QSAR and network-based computational chemistry approach to antimicrobials. Part 1: Multispecies activity models for antifungals. *J Comput Chem* **29**: 656–657.

González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008c) Proteomics, networks and connectivity indices *Proteomics* **8**: 750–778.

Gudmundsson GH, Lidholm DA, Asling B, Gan R, Boman HG (1991) The cecropin locus. Cloning and expression of a gene cluster encoding three antibacterial peptides in *Hyalophora cecropia*. *J Biol Chem* **266**: 11510–11517.

Hemmi H, Ishibashi J, Hara S, Yamakawa M (2002) Solution structure of moricin, an antibacterial peptide, isolated from the silkworm *Bombyx mori*. *FEBS Lett* **518**: 33–38.

Hilpert K, Fjell CD, Cherkasov A (2008) Short linear cationic antimicrobial peptides: screening, optimizing, and prediction. *Methods Mol Biol* **494**: 127–159.

HMMER, UBC Bioinformatics Centre. http://bioinformatics.ubc.ca/resources/tools/hmmer

Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, Villasante A, Dimitri P, Karpen GH, Celniker SE (2007) Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science* **316**: 1625–1628.

Goraya J, Wang Y, Li Z, O'Flaherty M, Knoop FC, Platz JE, Conlon JM (2000) Peptides with antimicrobial activity from four different families isolated from the skins of the North American frogs *Rana luteiventris, Rana berlandieri* and *Rana pipiens*. *Eur J Biochem* **267**: 894–900.

Halverson T, Basir YJ, Knoop FC, Conlon JM (2000) Purification and characterization of antimicrobial peptides from the skin of the North American green frog *Rana clamitans*. *Peptides* **21**: 469–476.

Kanai A, Natori S (1989) Cloning of gene cluster for sarcotoxin I, antibacterial proteins of *Sarcophaga peregrine*. *FEBS Lett* **258**: 199–202.

Kreyszig E (1970) *Introductory Mathematical Statistics, Principles and Methods*. John Wiley & Sons. Inc. New York.

Kolonin MG, Saha PK, Chan L, Pasqualini R, Arap W (2004) Reversal of obesity by targeted ablation of adipose tissue. *Nat Med* **10**: 625–632.

Lacroix E, Viguera AR, Serrano L (1997) Elucidating the folding problem of α-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol* **1**: 173–191.

Linzmeier R, Michaelson D, Liu L, Ganz T (1993) The structure of neutrophil defensin genes. *FEBS Lett* **321**: 267–273.

Mahoney MM, Lee AY, Brezinski-Caliguri DJ, Huttner KM (1995) Molecular analysis of the sheep cathelin family reveals a novel antimicrobial peptide. *FEBS Lett* **377**: 519–522.

Mak P, Wójcik K, Thogersen IB, Dubin A (1996) Isolation, antimicrobial activities, and primary structures of hamster neutrophil defensins. *Infect Immun* **64**: 4444–4449.

Marenah L, Flatt PR, Orr DF, Shaw C, Abdel-Wahab YH (2006) Skin secretions of Rana saharica frogs reveal antimicrobial peptides esculentins-1 and -1B and brevinins-1E and -2EC with novel insulin releasing activity. *J Endocrinol* **188**: 1–9.

Matutte B, Storey KB, Knoop FC, Conlon JM (2000) Induction of synthesis of an antimicrobial peptide in the skin of the freeze-tolerant frog, *Rana sylvatica*, in response to environmental stimuli. *FEBS Lett* **483**: 135–138.

Moerman L, Bosteels S, Noppe W, Willems J, Clynen E, Schoofs L, Thevissen K, Tytgat J, Van Eldere J, Van Der Walt J, Verdonck F (2002) Antibacterial and antifungal properties of α-helical, cationic peptides in the venom of scorpions from southern Africa. *Eur J Biochem* **269**: 4799–4810.

Moore KS, Bevins CL, Brasseur MM, Tomassini N, Turner K, Eck H, Zasloff M (1991) Antimicrobial peptides in the stomach of *Xenopus laevis*. *J Biol Chem* **266**: 19851–19857.

Morikawa N, Hagiwara K, Nakajima T (1992) Brevinin-1 and -2, unique antimicrobial peptides from the skin of the frog, *Rana brevipoda* porsa. *Biochem Biophys Res Commun* **189**: 184–190.

NCBI, National Center for Biotechnology Information (NCBI) Protein BLAST. http://www.ncbi.nlm.nih.gov

Oh D, Shin SY, Kang JH, Hahm KS, Kim KL, Kim Y (1999) NMR structural characterization of cecropin A(1-8) - magainin 2(1-12) and cecropin A (1-8) - melittin (1-12) hybrid peptides *J Pept Res* **53**: 578–589.

Orivel J, Redeker V, Le-Caer JP, Krier F, Revol-Junelles AM, Longeon A, Chaffotte A, Dejean A, Rossier J (2001) Ponericins, new antibacterial and insecticidal peptides from the venom of the ant *Pachycondyla goeldii*. *J Biol Chem* **276**: 17823–17829.

Park JM, Jung JE, Lee BJ (1995) Antimicrobial peptides from the skin of a Korean frog, *Rana rugosa*. *Biochem Biophys Res Commun* **205**: 948–954.

Popsueva AE, Zinovjeva MV, Visser JW, Zijlmans JM, Fibbe WE, Belyavsky AV (1996) A novel murine cathelin-like protein expressed in bone marrow. *FEBS Lett* **391**: 5–8.

Prado-Prado FJ, de la Vega OM, Uriarte E, Ubeira FM, Chou KC, Gonzalez-Diaz H (2007a) Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg Med Chem* **17**: 569–575.

Prado-Prado FJ, Gonzalez-Diaz H, Santana L, Uriarte E (2007b) Unified QSAR approach to antimicrobials. Part 2: Predicting activity against more than 90 different species in order to halt antibacterial resistance. *Bioorg Med Chem* **15**: 897–902.

Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC (2008) Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg Med Chem* **16**: 5871–5880.

Ramos de Armas R, González-Díaz H, Molina R, Uriarte E (2005) Stochastic-based descriptors studying biopolymers biological properties: extended MARCH-INSIDE methodology describing antibacterial activity of lactoferricin derivatives. *Biopolymers* **77**: 247–256.

Rosetto M, Manetti AG, Marchini D, Dallai R, Telford JL, Baldari CT (1993) Sequences of two cDNA clones from the medfly Ceratitis capitata encoding antibacterial peptides of the cecropin family. *Gene* **134**: 241–243.

Resch B (2004) Hidden Markov Models. A tutorial for the course computational intelligence. Signal processing and speech communication laboratory. http://www.igi.tugraz.at/lehre/CI/tutorials/HMM/HMM.pdf

Shin SY, Kang JH, Janq SY, Kim Y, Kim KL, Kahm KS (2000) Effects of the hinge region of cecropin A(1-8)-magainin 2(1-12), a synthetic antimicrobial peptide, on liposomes, bacterial and tumor cells. *Biochim Biophys Acta* **1463**: 209–218.

Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak, SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalon DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ, Marra MA (2002) Generation and initial analysis of more than 15000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci USA* **99**: 16899–16903.

Torres-Larios A, Gurrola GB, Zamudio FZ, Possani LD (2000) Hadrurin, a new antimicrobial peptide from the venom of the scorpion *Hadrurus aztecus*. *Eur J Biochem* **267**: 5023–5031.

Tossi A, Scocchi M, Zanetti M, Storici P, Gennaro R (1995) PMAP-37, a novel antibacterial peptide from pig myeloid cells. cDNA cloning, chemical synthesis and activity. *Eur J Biochem* **228**: 941–946.

Qu Z, Steiner H, Engström A, Bennich H, Boman HG (1982) Insect immunity: isolation and structure of cecropins B and D from pupae of the Chinese oak silk moth, *Antheraea pernyi*. *Eur J Biochem* **127**: 219–224.

Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Mendez-Alvarez E (2006) A QSAR model for *in silico* screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *Med Chem* **49**: 1149–1156.

Swiss, European Bioinformatics Institute 2006–2008. EBI is an Outstation of the European Molecular Biology Laboratory. http://www.ebi.ac.uk/swissprot/

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**: 320–322.

Spivak M (1965) *Calculus on Manifolds*. Benjamin. New York.

Tennessen JA, Blouin MS (2007) Selection for antimicrobial peptide diversity in frogs leads to gene duplication and low allelic variation. *J Mol Evol* **65**: 605–615.

Uniprot Swiss-prot ftp://ftp.expasy.org/databases/swiss-prot/release_compressed/<>uniprot_sprot.fasta.gz

Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**: 415–427.

Uversky VN (2002) What does it mean to be natively unfolded? *Eur J Biochem* **269**: 2–12.

Vizioli J, Bulet P, Hoffmann JA, Kafatos FC, Müller HM, Dimopoulos G (2001) Gambicin: a novel immune responsive antimicrobial peptide from the malaria vector *Anopheles gambiae*. *Proc Natl Acad Sci USA* **98**: 12630–12635.

Zimin AV, Smith DR, Sutton G, Yorke JA (2008) Assembly reconciliation. *Bioinformatics* **1**: 142–145.