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 $\alpha$-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:

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

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

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

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 **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

$$\begin{aligned}\mathbb{L}(h_1,\ldots,h_n \mid o_1,\ldots,o_n) &\propto \mathbf{P}(h_1,\ldots,h_n \mid o_1,\ldots,o_n)\\ &= \prod_{i=1}^{n}\mathbf{P}(o_i \mid h_i).\prod_{i=1}^{n}\mathbf{P}(h_i \mid h_{i-1})\end{aligned} \qquad (3)$$

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 $\Omega = \{\pi, \mathrm{A},\mathrm{B}\}$:

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 $\mathrm{P}_0$.

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 $\Omega$, is given by:

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

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 $\mathrm{P}_0$.

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

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

$\mathbf{P}(O,H|\Omega)$ (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 $\Omega = \{\mathrm{P}_0,\mathrm{A},\mathrm{B}\}$.

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 $\mathbf{P}_0$.**

Absolute frequency of natural and synthetic antibacterial peptides which act exclusively against bacteria, fungi, viruses and mammalian cancer cells **(set A)** to vector $\mathrm{P}_0$. 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).

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

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 $\text{Index}_A$.**

Absolute frequency of natural and synthetic antibacterial peptides which act exclusively against bacteria **(set B)** to vector $\text{Index}_A$. 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 $\mathbf{P}(o_i|h_{i?IndexA})$ to be candidate SCAAPs if ($o_i$ = *true*) the $\mathbf{P}(o_i = true|h_{i?IndexA})$ = 0.95, and its complement ($o_i$ = *false*) $\mathbf{P}(o_i = false|h_{i?IndexA})$ = 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)

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

$$\sigma_W^2=\frac{n_1n_2(n_1+n_2+1)}{12} \quad (11)$$

Hence $c_1$ and $c_2$ were obtained substituting $\mu_W$ and $\sigma_W$ in Eqns. (12) and (13)

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

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

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<br>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-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 post-genomic 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 $\alpha$-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 $\alpha$-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, Engströn 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.