ACTA ABP BIOCHIMICA POLONICA

Vol. 61, No 4/2014
693–698
on-line at: www.actabp.pl

Regular paper

Possible computational filter to detect proteins associated to influenza A subtype H1N1

Carlos Polanco[1]✉, Thomas Buhse[2], Jorge Alberto Castañón-González[1] and José Lino Samaniego[1]

[1]Facultad de Ciencias de la Salud, Universidad Anáhuac, Col. Lomas Anáhuac, Huixquilucan Estado de México, México; [2]Centro de Investigaciones Químicas, Universidad Autónoma del Estado de Morelos, Cuernavaca Morelos, México

The design of drugs with bioinformatics methods to identify proteins and peptides with a specific toxic action is increasingly recurrent. Here, we identify toxic proteins towards the influenza A virus subtype H1N1 located at the UniProt database. Our quantitative structure-activity relationship (QSAR) approach is based on the analysis of the linear peptide sequence with the so-called Polarity Index Method that shows an efficiency of 90% for proteins from the Uniprot Database. This method was exhaustively verified with the APD2, CPPsite, Uniprot, and AmyPDB databases as well as with the set of antibacterial peptides studied by del Rio *et al*. and Oldfield *et al.*

Key words: Polarity Index Method, influenza A virus subtype H1N1, drug design, QSAR method

Received: 21 May, 2013; revised: 04 April, 2014; accepted: 12 June, 2014; available on-line: 07 November, 2014

INTRODUCTION

Pandemic of the influenza A virus subtype H1N1 occurred in Mexico in 2009. It refers to the 1918 flu pandemic outbreak in the USA and is commonly known as a Spanish flu. Spanish influenza pandemic caused the death of approximately 3.7% of the Earth population (between 50 and 100 million people). Most deceases took place during the first 25 weeks of the outbreak (USCB, 2013). Comparing both the pandemic processes, the Mexican pandemic was not lethal because the virus was weak and the means of transmission were bird-human. However, it is only a matter of time until the inevitably occurs and a lethal strain arises in humans. One particular point is that this type of influenza virus (Chowell *et al.,* 2011), which can be quickly spread around the world, is a variety of an influenza virus with the gene segments common to birds, pigs and human flu strains. Considering that pigs (Wenjun *et al*., 2009) are susceptible to share the bird and human influenza virus allowing a redistribution of gene segments, we can assume that this virus can rapidly mutate in humans. One of the distinctive signs of the influenza A virus subtype H1N1 pandemic is, in particular, that the virus is easily transmissible among humans. A future pandemic threat can be combated also by predicting the location of the outbreak carrying out a fast count of the people infected using the predictor algorithms (Nishiura, 2011) and other predictive models or by developing new drugs. Some of them are based on proteins and peptides with toxic action towards influenza A subtype H1N1 and are detected by the bioinformatics algorithms. In this sense, each scientific and technological research should be oriented to the field of Proteomics as well as to the design of the efficient computational-mathematical algorithms which are able to identify and predict peptides and proteins with toxic action on this particular type of virus. These techniques may help to avoid the impractical but very necessary technique of trial and error involved in the chemical synthesis of new peptides and proteins. Although nature is recognized as a main source of proteins with toxic action against the influenza A virus subtype H1N1, recent research efforts have been directed to the production of synthetic and hybrid proteins. One of the procedures is to generate proteins replacing and/or removing constitutive amino acids from the natural proteins known for their anti-influenza A H1N1 action (Barik, 2012; Tsai *et al.,* 2012), reducing their size simultaneously keeping or increasing their toxicity. Another technique is to join two peptides or strains of proteins that individually do not have this property but combined together become highly toxic (Mohamed *et al.,* 2009). Altering a peptide to quantify its toxic action in a laboratory through traditional methods of trial and error would take a combination of possibilities beyond any practicability as the number of peptides built from 7 amino acid peptide is $20^7 = 1.28 \times 10^9$. Therefore, new techniques to build proteins against influenza A subtype H1N1 are based on mathematical-computational methods simulating peptide alterations as well as evaluating and qualifying them to determine if a peptide complies with the criteria required. These methods are highly complex in their mathematical-computational design and execution. They simulate the characteristics necessary to evaluate all possible combinatorial. In this work we describe the quantitative structure-activity relationship (QSAR) approach called Polarity Index Method by taking a single physicochemical property, namely the polarity, to identify efficiently the influenza A proteins subtype H1N1 from the UniProt database (Magrane, 2011) accessed on March 19, 2014. This method was previously applied to detect bacteria and selective cationic amphipathic antibacterial peptides (SCAAP) (Polanco & Samaniego, 2009; Polanco *et al*., 2012), taking the existent 20 proteic amino acid

✉e-mail: polanco@unam.mx

classification differentiated by the side chain R and divided into four different categories according to their polarity profiles (Kawashima & Kanehisa, 2000). The Polarity Index Method uses this classification only to identify the characteristic template of the influenza A protein subtype H1N1 group, which was exhaustively tested with 7 databases and was proven to be highly efficient. Our work shows the efficiency of a computational mathematical method that identifies with a high level of precision influenza A subtype H1N1 proteins but does not intend to carry out any experimental verification on the peptides.

MATERIAL AND METHODS

The Polarity Index Method has already been published to identify efficiently selective antibacterial peptides from the APD2 database (Polanco *et al.*, 2012). For this reason, we mention only the necessary modifications for the identification of influenza A subtype H1N1 proteins. Later, we present a detailed example to clarify its mechanism (Section 2.8).

Polarity Index Method updates

The method essentially measures the polar profile of the peptide in a comprehensive manner by taking into account 16 polar interactions from the four polarity groups P+, P–, N, and NP (Polanco *et al.*, 2012). Its metric considers reading of the linear sequence of the amino acids of the peptide or protein. In order to perform a comprehensive test, we considered all groups of peptides and proteins that have been studied so far. First of all, we calibrated with the peptides found in the Uniprot database and verified our approach with the following databases: the entire set of antimicrobial peptides from APD2 database (Wang & Wang, 2009), the set of cells penetrating the endocytic pathway of peptides and the non-endocytic pathway from the CPPsite database (Gautam *et al.*, 2012), the set of influenza proteins and human neuronal proteins from the Uniprot database (Magrane, 2011), the amyloid peptides from the AmyPDB database (Pawlicki *et al.*, 2008), the set of selective antibacterial peptides studied by del Rio and coworkers (2001), and the set of natively unfolded proteins and natively folded proteins, studied by Oldfield *et al.* (2005).

Modifications

The $\mathbf{P}[i,j]$ matrix in the source program (Polanco *et al.*, 2012) is substituted with the profile of incidents for the corresponding set of influenza A subtype H1N1 proteins. Its worth noting that in this case it was necessary to obtain nine $\mathbf{P}[i,j]$ matrices, because we obtained the same number of sub-classifications (Sections APD2 database preparation – SCAAP Database preparation). Once the $\mathbf{P}[i,j]$ matrix is concluded for each sub-group, it is normalized to unity. In the same way, the $\mathbf{Q}[i,j]$ matrix contains the profile of incidents for the sequence in study.

The Polarity Index Method selected as influenza A subtype H1N1 proteins candidates whose $\mathbf{P}[i,j] + \mathbf{Q}[i,j]$ vector space complied with different rules. The rules mentioned (Table 1) are a result of observing that polar interactions are more frequent than others, today already working in a fully automated version to avoid producing this step manually. Those peptides that meet 4 or 5 rules mentioned in the Table 1, the polarity index method be regarded as peptides associated with influenza A type H1N1. E.g. the rule 1, "Polar interaction 8 is not present in the 12th position" means that the polar 8 interaction [P-, NP] can occur on any of the 16 possible positions, but not in the 12th position. In case of rule 3 "Polar interaction 12 is present in the 1th position" means that the interaction 12 [N, NP] must be present in the first position only.

Multiple and unique action

Peptide sets with unique toxic action are those peptides with verified experimental action over one pathogenic agent, whereas multiple action peptide sets are formed with those peptides with toxic action over two or more pathogens that are over-represented.

APD2 database preparation

3636 peptides were taken from the antimicrobial peptide database (APD2 database) (Wang & Wang, 2009) and classified by their *multiple* action as follows: 149 Gram– ONLY, 1711 Gram+/Gram– ONLY, 315 Gram+ ONLY, 141 cancer cells, 744 fungi, 21 insects, 244 mammalian cells, and 47 parasites; and 1059 were classified by their *unique* action as follows: 111 Gram– ONLY, 213 Gram+ ONLY, 518 Gram+/Gram– ONLY, 20 cancer cells, 88 fungi, 2 insects, 11 mammalian cells, and 9 parasites.

CPPsite database data preparation

115 cell-penetrating peptides were classified from the CPPsite database (Gautam *et al.,* 2012) by their uptake mechanism as follows: 93 non-endocytic pathway, and 22 endocytic pathway. Those peptides with different penetration mechanisms included in the CPPsite database were not considered.

Natively unfolded and folded proteins data preparation

148 proteins, of which 51 natively *unfolded* proteins and 97 natively *folded* proteins, were selected from the Supplementary information from Oldfield *et al.* (Oldfield *et al.,* 2005).

UniProt database preparation

Proteins extracted from the Uniprot database (Magrane, 2011): (i) set of proteins associated with influenza A type H1N1 in nine subgroups: 33 HA, 33 M1, 16 M2, 27 NA, 58 NP, 29 NS1, 24 PA, 49 PB1, and 1 PB2 proteins, and (ii) 3616 proteins which expressed in neurons, and located in every living organism studied. In that set we found 755 human revised proteins expressed in neurons, and 2879 non-human revised proteins expressed in neurons.

AmyPDB database data preparation

We analyzed 15 of 1705 proteins originally classified in several amyloid protein families stored in the AmyPDB database (Pawlicki *et al.*, 2008) and restricted to: (i) Amyloid formed *in vivo* (the precursor protein, or a specific sub-segment, forms fibrils in human), and (ii) Amyloid formed *in vitro* (the polypeptide forms fibrils under experimental conditions).

SCAAP Database preparation

30 Selective Cationic Amphipathic Antibacterial Peptides (SCAAP) were used in Table 2 and Table 2A from del Rio and coworkers (2001).

Table 1. **P**[*i,j*] Polarity matrix

	P+	P–	N	NP
P+	0.0149871381	0.0158818923	0.0158818923	0.0438429713
P–	0.0152667491	0.0098982221	0.0530142039	0.0267867129
N	0.0515043065	0.0282966103	0.1578123271	0.1626775563

Number of incidences of proteins expressed in terms of their relative frequencies (see example Section 2.10).

Table 2. Polarity Index Method (test)

Linear position	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
P[*i,j*] + Q[*i,j*] vector of study.	$Q_{(1,1)}$	$Q_{(1,2)}$	$Q_{(1,3)}$	$Q_{(1,4)}$	$Q_{(2,1)}$	$Q_{(2,2)}$	$Q_{(2,3)}$	$Q_{(2,4)}$	$Q_{(3,1)}$	$Q_{(3,2)}$	$Q_{(3,3)}$	$Q_{(3,4)}$	$Q_{(4,1)}$	$Q_{(4,2)}$	$Q_{(4,3)}$	$Q_{(4,4)}$
Rule # 1. Polar interaction 8 is **not** present in the 12th position	√	√	√	√	√	√	√	√	√	√	√	×	√	√	√	√
Rule # 2. Polar interaction 6 is present 16st position	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	√
Rule # 3. Polar interaction 12 is present in the 1th position	√	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Rule # 4. Polar interaction 11 is present in the 1st position	√	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Rule # 5. Polar interaction 10 is not present in the 1th position	×	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

Polarity index method identification rule. (√): The polar interaction is present in the position. (×): The polar interaction is **not** present in the position.

Test plan

The discriminative efficiency of the polarity index method is measured by calculating three aspects: (i) the number of hits in the identification of the specific group; (ii) the percentage of errors in the identification of the other groups. In this sense, the method must be efficient in identifying the group and simultaneously rejecting those peptides or proteins which are not a part of this group, and (iii) graphing the relative frequency of each polar interaction of all subgroups of the proteins (Sections APD2 database preparation — SCAAP Database preparation), associated with influenza A subtype H1N1, extracted from Uniprot database (Magrane, 2011).

Example

Although this method has been already published (Polanco *et al.*, 2012), we provide here a detailed description of an illustrative example in order to clarify the used algorithm. Our aim is to get to know if the protein MSLLTEVET YVLSIIP SGPLKAEIAQRLEDVFA GKNTDLEVLM EWLKTRPILSPLTK GILGFVFTLTVPSERGLQRRRFV QNALNG NGDPNNMDKAVKLYRKLK REITFHGAKEISLSYSAGALASCMGLIYNRM GAVTTEVAFGLVCATCEQIADSQHR SHRQMVTTTNPLIRHENRMVLAS TTAKAMEQMAGSSEQAA EAMEVASQ ARQMVQAMRTIGTH PSSSAGLKNDLLENQAYQKRMGVQ MQRFK, is in accordance with the polarity index method. To answer this question it is necessary to follow the following five steps:

The above sequence is converted to its numeric equivalent according to the following rule of equivalence: The amino acids: H, K, and R are replaced by the number "1"; the amino acids: D, and E are replaced by the number "2"; the amino acids: C, G, N, Q, S, T, and Y are replaced by the number "3"; finally the amino acids: A, F, I, L, M, P, V, and W are replaced by the number "4". Note that the four numerical equivalents {1, 2, 3, and 4} correspond to the four polar groups: [P+], [P-], [N], and [NP] and are listed in the same order. The numeric equivalence of the aforementioned sequence is: 434432423344 344433441 424431 4224443133242444244131 4443443134434 443434432134311144334433332 4334214414 311411243413412 4343334344433434433 1 43443324 443443433234423311311344 3333444112314444 33 34144234433323 442442443341 3443441343314333434 13 244234343311 434343141.

Read the resulting numerical sequence, from left to right, moving one position at a time. Each pair is considered as an element (*i,j*), in this case the first pair is (*i,j*) = (4,3), the second pair is (*i,j*) = (3,4),respectively, and same strategy should be applied further, until the last pair (*i,j*) = (4,1). Please note that the pairs (*i,j*) correspond to a square matrix of order 4 which we named forward **Q**[*i,j*] matrix and where the element *i* represents the row and *j* represents the column of **Q**[*i,j*] matrix.

Count the occurrences of every (*i,j*) pair in the **Q**[*i,j*] matrix. In this way the **Q**[*i,j*] matrix represents the occurrences of the numerical sequence.

The matrix **Q**[*i,j*] is weighted, and added to **P**[*i,j*] matrix (Table 1), i.e. **Q**[*i,j*] = **Q**[*i,j*] + **P**[*i,j*]. Finally, the **Q**[*i,j*] matrix is linearized. As a result, the **Q**[*i,j*] matrix becomes a vector of 16 elements, i.e. {16, 15, 12, 11, 9, 8, 4, 13, 14, 3, 10, 1, 7, 2, 5, 6}.

The vector is compared with the rules in Table 2. In this example, all the rules are accepted, and therefore, this protein is considered as an influenza A protein subtype H1N1 candidate.

Table 3. Percentages of polarity matches.

	*	B+	B-	B+/-	Fu	Pa	Ca	Ma	In	Am	Sc	Ce	Cne	Un	Fo	Hu	Nh	HA	NA	NP	M1	M2	NS1	PA	PB1	PB2
HA	U	4	1	1	0	0	0	0	0	0	0	0	0	2	2	2	2	91	15	0	0	0	0	0	0	0
HA	M	4	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NA	U	4	2	3	5	0	0	0	0	0	0	0	2	0	0	1	2	3	93	0	0	0	0	0	0	0
NA	M	5	2	4	6	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NP	U	2	1	2	0	0	0	0	0	7	3	0	3	0	6	7	8	0	0	100	0	0	0	0	4	0
NP	M	2	1	1	1	2	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M1	U	0	2	0	0	0	0	0	0	7	0	0	1	0	2	4	4	0	0	0	91	0	0	0	0	0
M1	M	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M2	U	10	2	5	1	0	0	18	0	7	0	0	5	2	5	7	6	0	0	0	0	100	14	0	0	0
M2	M	10	1	5	4	4	12	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NS1	U	7	2	5	1	0	0	9	0	0	0	0	8	0	2	3	2	0	0	0	0	0	83	4	0	0
NS1	M	7	1	4	2	2	3	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PA	U	1	1	1	1	0	0	0	0	0	0	0	2	0	1	4	4	0	0	0	0	0	0	96	0	0
PA	M	2	1	1	1	0	1	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PB1	U	5	3	1	0	0	0	18	0	13	0	0	2	0	3	6	7	0	0	5	0	0	0	0	94	0
PB1	M	7	5	4	4	2	3	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PB2	U	1	1	1	0	0	0	0	0	0	0	0	1	0	4	3	1	0	0	0	0	0	0	4	0	100
PB2	M	3	1	2	1	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Percentages (hits/total rounded integer part) found by polarity index method pointed to: nine subgroups of proteins associated to influenza A type H1N1: **HA**, **NA**, **NP**, **M1**, **M2**, **NS1**, **PA**, **PB1**, y **PB2** from Uniprot database (Magrane, 2011). **B+**: GRAM+ bacteria, **B-**: GRAM– bacteria, **B+/-**: GRAM+ and GRAM- bacteria, **Fu**: Fungi, **Pa**: Parasites, **Ca**: Cancer cells, **Ma**: Mammalian cells, and **In**: Insects from APD2 database (Wang & Wang, 2009). **Am**: Amyloidosis proteins from AmyPDB database (*Pawlicki et al., 2008).* **Sc***: Selective antibacterial peptides from del Rio et al. (del Rio et al., 2001).* **Ce***: Cells penetrating peptides endocytic pathway proteins, and* **Cne***: Cells penetrating peptides non-endocytic pathway proteins* from CPPsite database (Gautam *et al.,* 2012). **Un**: Natively unfolded proteins, and **Fo**: Natively folded proteins studied by Oldfield *et al.* (Oldfield *et al.,* 2005). **Hu**: Human neuronal proteins, and **Nh**: Non human neuronal proteins from Uniprot database (Magrane, 2011). **U**: Unique action: Peptides with pathogenic action against only one group. **M**: Multiple action: Peptides with pathogenic action against two or more groups (Section Test plan).

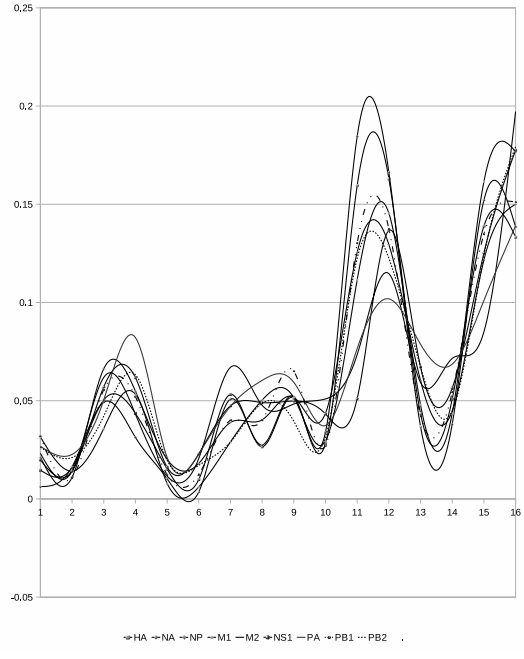

Figure 1. Comparison of the polar profile, from the nine subgroups of influenza proteins: HA, NA, NP, M1, M2, NS1, PA, PB1, AND PB2.
The 18 columns on the x-axis correspond to 16 amino acids of vector incidences (Section Test plan).

RESULTS

(i) The Polarity Index Method made a discriminating and positive identification of the nine subgroups of proteins associated with influenza A subtype H1N1 extracted from the UniProt database (90%, double-blind test) and shows an almost discriminative score with the remaining eight sub-classifications containing APD2, AmyPDB, Uniprot (Human and non human proteins), and CPPsite database, and the sets from del Rio *et al.*, and Oldfield *et al.* (Table 3).

The smooth graphics (Fig. 1), which correspond to the nine subgroups of proteins associated to influenza virus subtype A H1N1 (Section APD2 database preparation – SCAAP Database preparation), have no coincidences in their maximum and minimum points for the polar interactions: 3 and 4 ([P+,N], and [P+,NP]), from 6 to 8 ([P-,P-], [P-,P+], and [P-,N]), 11, and 12 ([N,N], and [N,NP]), and 14 ([NP,P-]). The method is sensitive to the number of differences, i.e. the greater is the number of differences, the greater is its efficiency. In this case, the number is very high while usually in this kind of evaluations the number of differences is two or less.

DISCUSSION

The polarity is a measure of the electromagnetic stability of matter, while electronegativity (Matsunaga *et al.,* 2003) is a numeric equivalent, which metric is involved in more than 84% (Thakur *et al.,* 2012) of the bioinformatics algorithms related to understand the toxic action of proteins. We think this metric is not sufficient if it is represented only by a single number. Instead, we have shown that the count of the polar incidences, i.e. 16 in the case of the Polarity Index Method, is much more

comprehensive. We believe that this characteristic explains why this method provides an effective discriminative measure of the influenza A proteins group subtype H1N1, the SCAAP (Polanco *et al*., 2012; Polanco *et al*., 2013). In addition, it is also important to mention that the metric considers only one measure. This means that the algorithm is not complex, allowing its implementation for cluster computing under parallel programming, i.e. in a collaborative programming that allows to run multiple instructions simultaneously. This could help analyze peptide spaces in order to better understand the selection mechanisms of biological systems concerning the amino acids subgroups. The method mentioned here has been defined as a QSAR method (González-Díaz & Uriarte, 2005), although, due to its polarity matrix, we consider this method rather as a Markov model (Rabiner, 1989). It has already been used in a more comprehensive version called hidden Markov model (Rabiner, 1989). However, the main obstacle to consider it as a Markov model is that its polarity matrix does not conform exactly as a Markov matrix, because it is not stochastic. A stochastic Markov matrix is this one in which the lines or columns add up to 1. Nevertheless, we believe that rendering the stochastic matrix will undoubtedly enhance the efficiency of a method. Therefore, by using multiple Markov matrix on a Markov model, called Hierarchical Hidden Markov Model (HHMM) (Wang *et al*., 2013), the new method will have different profiles of the same phenomenon, each of them represented by a Markov matrix and interacting together under a hierarchical weighting. Such Markov model has been used extensively on speech recognition (Lee, 2008). Bioinformatics arose thanks to these kind of algorithms (Hagen, 2000) making it easier to identify similarities in protein strains. For that reason, we are developing a new version of our model with such Markov structure. The effectiveness of the polarity index method in double-blind test reaches 90%, on eight different databases of proteins associated to influenza virus subtype A (H1N1). This level of success is high enough not to consider it as a lucky coincidence, although the reason at the more molecular level remains unknown for us so far. We believe that the polarity is a fundamental property of matter that characterizes the form of how a protein adopts to the lipid-aqueous space so that the amino acid sequence (primary structure) expresses such conformational structure. Until now, we have verified this conjecture in all groups of peptides and proteins that we cite in this work. We have even used that property for modeling prebiotic scenarios. Nevertheless, the closer biochemical reason still remains unknown. However, the present lack of biochemical insight stands in contrast to the efficiency of the Polarity Index Method in the identification of peptides and/or proteins and its usefulness for prospective drug design from a more macroscopical settled modeling approach, or as a first filter in the bioinformatics identification of peptides/proteins.

CONCLUSIONS

The adaptation of the polarity index method to identify the nine subgroups of influenza A subtype H1N1 proteins, and reject eight groups of peptides not associated with influenza, scattered in different databases. It has proven to be an efficient algorithm, measuring the polarity of the protein from its linear sequence.

Availability

The source program, the txt-files of each subgroups of proteins, and xls-files for the smooth graphics, are given as "Supplementary Material".

Conflict of Interests

We declare that we do not have any financial and personal interest with other people or organizations that could inappropriately influence (bias) our work.

Author Contributions

Theoretical conception and design: CP. Computational performance: CP. Data analysis: CP, JLS, TB, and JACG. Chemical analysis: TB. Results discussion: CP, JLS, TB, and JACG.

Acknowledgements

The authors thank the Computer Science Department at the Institute for Nuclear Sciences of the National Autonomous University of Mexico for support. We also gratefully acknowledge financial support received form the Mexican-French bilateral research grant CONACYT (188689) — ANR (12-IS07-0006), and Concepción Celis Juárez whose suggestions and proof-reading have greatly improved the original manuscript.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403−410.

Asakura T, Mattiello JA, Obata K, Asakura K, Reilly MP, Tomassini N, Schwartz E, Ohene-Frempong K (1994) Partially oxygenated sickled cells: sickle-shaped red cells found in circulating blood of patients with sickle cell disease. *Proc Natl Acad Sci USA* **91**: 12589−12593.

Austin FJ, Kawaoka Y, Webster RG (1990) Molecular analysis of the haemagglutinin gene of an avian H1N1 influenza virus. *J Gen Virol* **71**: 2471−2474.

Barik S (2012) New treatments for influenza. *BMC Med* **10**: 104.

Beklemishev AB, Blinov VM, Vasilenko SK, Golovin SIa, Karginov VA (1986) [Primary structure of the full-size DNA copy of the hemagglutinin gene of influenza virus A/Kiev/59/79 (H1N1)]. *Bioorg Khim* **12**: 375−381.

Bolintineanu DS, Kaznessis YN (2011) Computational studies of protegrin antimicrobial peptides: a review. *Peptides* **32**: 188−201.

Both GW, Shi CH, Kilbourne ED (1983) Hemagglutinin of swine influenza virus: a single amino acid change pleiotropically affects viral antigenicity and replication. *Proc Natl Acad Sci USA* **80**: 6996−7000.

Chowell G, Echevarría-Zuno S, Viboud C, Simonsen L, Tamerius J, Miller MA, Borja-Aburto VH (2011) Characterizing the epidemiology of the 2009 influenza A/H1N1 pandemic in Mexico. *PLoS Med* **8**: e1000436

Concannon P, Cummings IW, Salser WA (1984) Nucleotide sequence of the influenza virus A/USSR/90/77 hemagglutinin gene. *J Virol* **49**: 276–278.

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

Gautam A, Singh H, Tyagi A, Chaudhary K, Kumar R, Kapoor P, Raghava GP (2012) CPPsite: a curated database of cell penetrating peptides. *Database (Oxford)* **2012**: bas015, accessed May 20, 2013.

González-Díaz H, Uriarte E (2005) Proteins QSAR with Markov average electrostatic potentials. *Bioorg Med Chem Lett* **15**: 5088–5094.

Hagen JB (2000) The origins of bioinformatics. *Nat Rev Genet* **1**: 231–236.

Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. *Neural Comput* **18**: 1527–1554.

Hiti AL, Davis AR, Nayak DP (1981) Complete sequence analysis shows that the hemagglutinins of the H0 and H2 subtypes of human influenza virus are closely related. *Virology* **111**: 113–124.

Ito T, Couceiro JN, Kelm S, Baum LG, Krauss S, Castrucci MR, Donatelli I, Kida H, Paulson JC, Webster RG, Kawaoka Y (1998) Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *J Virol* **72**: 7367–7373.

Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. *Nucleic Acids Res* **28**: 374.

Lamb RA, Lai CJ, Choppin PW (1981) Sequences of mRNAs derived from genome RNA segment 7 of influenza virus: colinear and interrupted mRNAs code for overlapping proteins. *Proc Natl Acad Sci USA* **78**: 4170–4174.

Lee KS (2008) EMG-based speech recognition using hidden markov models with global control variables. *IEEE Trans Biomed Eng* **55**: 930–940.

Luoh SM, McGregor MW, Hinshaw VS (1992) Hemagglutinin mutations related to antigenic variation in H1 swine influenza viruses. *J Virol* **66**: 1066–1073.

Magrane M. and the UniProt consortium (2011) *UniProt Knowledgebase: a hub of integrated protein data Database bar009*.

Matsunaga N, Rogers DW, Zavitsas AA (2003) Pauling's electronegativity equation and a new corollary accurately predict bond dissociation enthalpies and enhance current understanding of the nature of the chemical bond. *J Org Chem* **68**: 3158–3172.

Mohamed R, Fatemeh J, Abdul RO, Aini I, Sharifah SH, Khatijah Y (2009) Identification and characterisation of a novel antiviral peptide against avian influenza virus H9N2. *Virol J.* **6**: doi: 10.1186/1743-422X-6-74.

Nishiura H (2011) Prediction of pandemic influenza. *Eur J Epidemiol* **26**: 583–584.

Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK (2005) Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **44**: 1989–2000.

Pawlicki, Le Béchec A, Delamarche C (2008) AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics* **10**: 9:273, accessed Mar 10, 2014.

Polanco C, Samaniego JL, Buhse T, Mosqueira FG, Negron-Mendoza A, Ramos-Bernal S, Castanon-Gonzalez JA (2012) Characterization of selective antibacterial peptides by polarity index. *Int J Pept* **2012**: 585027. doi: 10.1155/2012/585027.

Polanco C, Samaniego JL (2009) Detection of selective cationic amphipathic antibacterial peptides by Hidden Markov models. *Acta Biochim Pol* **56**: 167–176.

Polanco C, Castañón-González JA, Macías A, Samaniego JL, Buhse T, Villanueva-Martínez S (2013) Detection of Severe Respiratory Disease Epidemic Outbreaks by CUSUM-Based Overcrowd-Severe-Respiratory-Disease-Index. *Model Comput Math Methods Med* **2013**: 213206.

Polanco C, Buhse T, Samaniego JL, Castañón-González JA (2013) Detection of selective antibacterial peptides by the Polarity Profile method. *Acta Biochim Pol* **60**: 183−189.

Polanco C, Buhse T, Samaniego JL, Castañón-González JA (2013a) A toy model of prebiotic peptide evolution: the possible role of relative amino acid abundances. *Acta Biochim Pol* **60**: 175–182.

Rabiner LR (1989) A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* **77**.

Rota PA, Shaw MW, Kendal AP (1987) Comparison of the immune response to variant influenza type B hemagglutinins expressed in vaccinia virus. *Virology* **161**: 269–75.

Taubenberger JK, Reid AH, Krafft AE, Bijwaard KE, Fanning TG (1997) Initial genetic characterization of the 1918 "Spanish" influenza virus. *Science* **275**: 1793–1796.

Thakur N, Qureshi A, Kumar M (2012) AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res* **40**: W199–204.

U.S. Census Bureau (USCB) International Programs. International Data Base. Revised: August 28, 2012 Version: Data:12.0625 Code:12.0321 https://www.census.gov/popul ation/international/data/worldpop/table_history.php accesed June 22, 2013.

Wang G, Li X, Wang Z (2009) APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res* **37**: D933–D937.

Wang X, Zang M, Xiao G (2013) Epigenetic change detection and pattern recognition *via* Bayesian hierarchical hidden Markov models. *Stat Med* **32**: 2292–307.

Wenjun M Kahn RE, Juergen Richt A (2009) The pig as a mixing vessel for influenza viruses: Human and veterinary implications. *J Mol Genet Med* **3**: 158–166.

Winter G, Fields S, Brownlee GG (1981) Nucleotide sequence of the haemagglutinin gene of a human influenza virus H1 subtype. *Nature* **292**: 72–75.

Winter G, Fields S (1980) Cloning of influenza cDNA ino M13: the sequence of the RNA segment encoding the A/PR/8/34 matrix protein. *Nucleic Acids Res* **8**: 1965–1974.

Yamnikova SS, Mandler J, Bekh-Ochir ZH, Dachtzeren P, Ludwig S, Lvov DK, Scholtissek C (1993) A reassortant H1N1 influenza A virus caused fatal epizootics among camels in Mongolia. *Virology* **197**: 558–563.