Detecting HP Pattern–Based Grammars to Synthesize Proteins: Inferring Sequence–Structure–Function Relationship

Giuseppe Nicosia, Eva Sciacca
University of Catania
Mathematics and Computer Science Department
Viale A. Doria 6, I 95125 Catania (CT), Italy
{nicosia/sciacca}@dmi.unict.it

Luca Zammataro
Institute for Cancer Research and Treatment
Strada Prov. 142 Km. 3.95,
I 10060 Candiolo (TO), Italy
luca.zammataro@ircc.it

Abstract

The detection of protein characters that could reveal how protein chains are constituted, is an important step to understand the main functions of specific classes of proteins. We made use of the concept of "HP Pattern–Based" grammars to study the connection between protein chains and protein functions. In order to consider the structure of the proteins the HP models were used. Amino acid sequences were treated as a formal language, and it was built a set of HP Pattern–Based grammars to describe this language by means the `Teiresias` *pattern discovery tool.*

First, this methodology was tested on the class of Antimicrobial peptides (AmPs). The deduced derivation rules of HP Pattern–Based Grammars were validated by the regular grammar designed by [11] which was used to create new, unnatural, AmPs sequences. Then, our approach was applied to characterize a function of the Pleckstrin Homology domain(PH Domain) which represents an important three dimensional domain which bind to phosphoinositides. Nowadays, interactions among PH domain amino acids and inositol phosphate are not well characterized. For the first time, by means of an HP Pattern–Based grammar, we highlight that this binding function can be described in terms of hydrophocity patterns.

Our approach points out some fundamental aspects regarding the relationship between sequence, structure and function of proteins.

1. Introduction

The discovery of sequence similarity in the primary structure of proteins or genes usually corresponds to residues conserved during evolution due to an important structural or functional role. This kind of analysis of biological sequences is a crucial task to synthesize new artificial protein sequences with therapeutic properties.

Our aim was to find a set of derivation rules of a grammar for specific classes of proteins in order to construct new protein chains with the properties of the considered class. To achieve our goal we have treated amino acid sequences as a formal language and built a set of regular grammars to describe this language. In order to face with the structure of the proteins we used hydrophobic-hydrophilic model (HP model) [10][14][7], in which amino acids are subdivided into two classes Hydrophobic (H) and Hydrophilic (P). We translated these sequences of amino acids into sequences of H and P. To find a set of regular grammars, which describes the HP Pattern–Based Grammar, we used the `Teiresias` pattern discovery tool [13], which demonstrated high–quality performances in the discovery of rigid patterns (motifs) in biological sequences. For every pattern we identified the derivation rules which bind every H and P of the pattern with the amino acids in the sequences. We will call the set of these kind of derivation rules "HP Pattern–Based Grammar."

First of all, we tested our methodology on the class of Antimicrobial peptides (AmPs). An attempt to understand protein characters which could contribute to reveal how protein chains are constructed is to examine all possible combinatorial sets of three, four, and five amino acids: triplets, quartets, and pentats, collectively called "constituent sequences"[12]. The derivation rules of our HP Pattern–Based Grammar were used to build the constituent sequences of AmPs. Then to test the correctness of this grammar we validated it by the regular grammar designed in [11] and which was used to create new, unnatural AmPs sequences.

As a test case we computed an HP Pattern–Based Grammar to characterize a binding function of the Pleckstrin Homology domain(PH Domain). For biological reasons, an important three dimensional domain is represented by PH domain, a 100 amino acid domain which is the major pro-

tein kinase C substrate of platelets, and which was found in several proteins as serine/threonine kinases, GTPase-activating proteins, phospholipases and cytoskeletal proteins and in many factors involved in signal transduction [18]. The PH domain consists of an up and down beta barrel of seven antiparallel beta strands and a C terminal amphiphilic alpha helix.

Nowadays, interactions among PH domain amino acids and inositol phosphate are not well characterized. A clear role of PH Domain in inositol phosphate interaction could shed in light the importance of this domain in signal transduction. For the first time, by means of HP Pattern–Based grammar, we highlight that this function can be described in terms of hydrophocity patterns.

2. HP Pattern–Based Grammar

Our aim was to find *HP Pattern–Based Grammars* for specific classes of proteins in order to synthesize new protein chains with the properties of the considered classes.

To achieve our goal, every protein sequence S_i belonging to a specific class of proteins S has been represented in a formal language as $S_i \in \Sigma^+$ where Σ is the alphabet, i.e., the set of all amino acids.

In order to face with the structure of the proteins, we used the HP model. In fact a major contribution to the free energy of the native conformation of a protein is due to interactions between hydrophobic amino acids that tend to form a core in the spatial structure shielded from the surrounding solvent by hydrophilic amino acids. In the model the amino acid sequence of a protein is abstracted as a sequence of hydrophobic and hydrophilic amino acids. Even though some amino acids cannot be classified clearly as being either hydrophobic or hydrophilic, the model disregards this fact to achieve simplicity.

Therefore, every sequence of amino acids S_i was "translated" into the binary sequence of H and P : $S_i^{HP} \in \{H, P\}^+$. Let S^{HP} be the set of translated sequences in S. The translation was carried out by means of three different HP models: Kyte-Doolittle [10], Rose [14], Hopp-Woods [7] (see table 1).

We deduced the patterns of S^{HP} using TEIRESIAS ALGORITHM [13] setting the patterns length L and the maximum number of non-wild card. For every pattern of length L $C_{HP}^L = c_{HP_1} \dots c_{HP_L}$ of the sequences S_i^{HP} we identified the derivation rules which bind every H and P (i.e. $c_{HP_l}, l = 1 \dots L$ when it is a non wild-card) of the pattern C_{HP}^L with the amino acids aa_l in the correspondent patterns $C^L = c_1 \dots c_L$ of the sequences S_i.

Let $|J|$ be the total number of amino acids aa_l that corresponds to every c_l. These amino acids aa_j are coupled with their corresponding frequency f_j in which they appear within the patterns C^L of the sequences S_i correspondent

Amino-acid	Kyte-Doolittle [10]	Hopp-Woods [7]	Rose et al. [14]
A	H	H	P
C	H	H	H
D	P	P	P
E	P	P	P
F	H	H	H
G	P	H	P
H	P	H	H
I	H	H	H
K	P	P	P
L	H	H	H
M	H	H	H
N	P	P	P
P	P	P	P
Q	P	P	P
R	P	P	P
S	P	P	P
T	P	H	P
V	H	H	H
W	P	H	H
Y	P	H	H

Table 1. Amino acids Hydropathy index

to the HP sequence S_i^{HP} which contains the pattern C_{HP}^l. Every derivation rule is in the form:

$$\mathbf{H}|\mathbf{P} \mapsto (aa_1, f_1)| \dots |(aa_{|J|}, f_{|J|}) \quad (1)$$

Let K be a subset of J. There were considered only the couples of amino acids and their corresponding frequencies $(aa_k, f_k) \in K \subseteq J$ which satisfy the following condition:

$$f_k > f_h : \ (aa_h, f_h) \in J \setminus K \ \wedge \ \sum_{k \in K} f_k > \textit{threshold} \quad (2)$$

This condition allow us to discard the less frequent amino acids and to consider only the most frequent, where the sum of the frequencies of the considered amino acids is at least equal to a chosen cut–off value *threshold*.

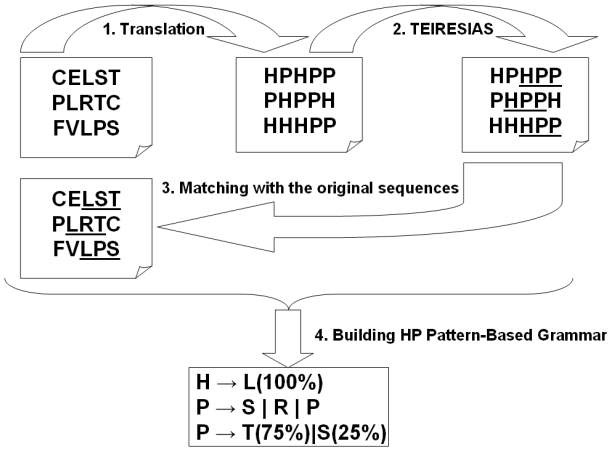

Figure 1. Exemplification of the proposed approach to build HP-Pattern Based Grammar

The figure 1 show a simplified exemplification of the main steps of the methodology on three short amino acid sequences: CELST, PLRTC and FVLPS. In this example we found an HP Pattern–Based Grammar of length 3.

The algorithm 1 shows the pseudo code of the proposed approach.

Algorithm 1 HP-Pattern Based Grammar Finder pseudo-code

1: input FASTA Sequences **S**, pattern_length, HP model
2: $\mathbf{S^{HP}}$:=translate(**S**, HP model)
3: $\mathbf{Pattern^{HP}}$=TEIRESIAS($\mathbf{S^{HP}}$, pattern_length)
// Find the set of all HP patterns
4: **for** $\mathbf{Pattern_i^{HP}} \in \mathbf{Pattern^{HP}}$ **do**
5: $\mathbf{Pattern} := find\_match(\mathbf{Pattern_i^{HP}}, \mathbf{S}, \mathbf{S^{HP}})$
// this procedure finds the set of amino acid patterns
// in the set of original sequences **S** which match the
// corresponding HP pattern in the set of translated
// sequences $\mathbf{S^{HP}}$
6: $\mathbf{frequency} := compute\_frequencies(\mathbf{Pattern})$;
// this procedure calculates the frequencies, amino
// acid by amino acid, within the set of amino acid
// patterns.
7: **print** $\mathbf{Pattern_i^{HP}} \mapsto$ **Pattern**, (**frequency**)
8: **end for**

3. Validation stage using AmPs Test Banch

Antimicrobial peptides (AmPs) are small proteins that are used by the innate immune system to combat bacterial infection in multicellular eukaryotes[19], they are found in diverse contexts including frog skin, scorpion venom and human sweat. There is mounting evidence that these peptides are less susceptible to bacterial resistance than traditional antibiotics and could form the basis for a new class of therapeutic agents[5].

3.1. Method

To validate the correctness of the HP Patter-Based grammars we used as a test banch 526 well-characterized eukaryotic AmPs sequences from the Antimicrobial Peptide Database (APD) [17].

We examined all possible combinatorial sets of three, four, and five amino acids, collectively called "*constituent sequences*"[12], considering sequences of three, four and five amino acids (called triplets, quartets and pentats).

In this validation stage we deduced the constituent sequences of the AmPs setting the maximum number of literals in the patterns and the number of non wild-card of Teiresias algorithm to three, four and five. Then we extracted all possible combinatorial sets of three amino acids in the form of H and P, $8\ (=2^3)$, all possible combinatorial sets of four amino acids in the form of H and P, $16\ (=2^4)$, and all possible combinatorial sets of five amino acids in the form of H and P, $32\ (=2^5)$.

We validated and compared our resulting grammar set with the set of regular grammars designed by [11] which was used to create new, unnatural AmPs sequences.

This set of 684 regular grammars can be represented as a formal language as $G_i \in \{\Sigma^*[\Sigma^*]^*\}^*$. These grammars describe a common arrangement of amino acids, for example, the frog AmP brevinin- 1E contains the amino acid sequence fragment PKIFCKITRK, which matches the grammar P[KAYS] [ILN] [FGI]C [KPSA] [IV] [TS] [RKC] [KR] (the bracketed expression [KAYS] indicates that, at the second position in the grammar, lysine, alanine, tyrosine or serine is equally acceptable).

Every G_i was "translated" into the sequence $G_i^{HP} \in \{H, P, .\}^*$, every amino acid was converted by means of the three HP models (using the table 1) and the sites between '[' and ']' were classified as:

H if the number of hydrophobic amino acids inside the site was more than the number of hydrophilic amino acids,

P if the number of hydrophilic amino acid inside the site was more than the number of hydrophobic amino acids,

. if the number of hydrophobic amino acids inside the site was equal to the number of hydrophilic amino acids

By this translation we deduced derivation rules (in the form of the formula 1) and compared them with our built derivation rules.

3.2. Results

Let us call **M** the set of the HP Pattern-Based Grammars built directly by means of AmP sequences (i.e. our model) and the given data **D** the set of HP Pattern-Based Grammars built by means of the 684 grammars designed by [11] (i.e. the reference data).

We performed a comparison between this two sets showing the percentage of correctly predicted PCP which is the *sensitivity* and the percentage of correctly non–predicted PCN which is the *specificity*:

- $PCP = 100 * \frac{TP}{TP+FN}$, and
- $PCN = 100 * \frac{TN}{TN+FP}$.

where TP, TN, FP, FN stands for : true positive, true negative, false positive and false negative.

In particular TP, TN, FP and FN quantities were computed as follow:

TP the number of amino acids on the right side of the derivation rules present in both **D** and **M**

TN the number of amino acids on the right side of the derivation rules not present in both **D** and **M**

FP the number of amino acids on the right side of the derivation rules present in **M** but not in **D**

FN the number of amino acids on the right side of the derivation rules present in **D** but not in **M**

The results using the three different HP models in table 1 showed high specificity and high sensitivity.

The chosen $threshold$ in the formula 2 was 70%. The figure 2 shows the validation of the quartets and pentats within AmPs sequences using Kyte-Doolittle HP Model.

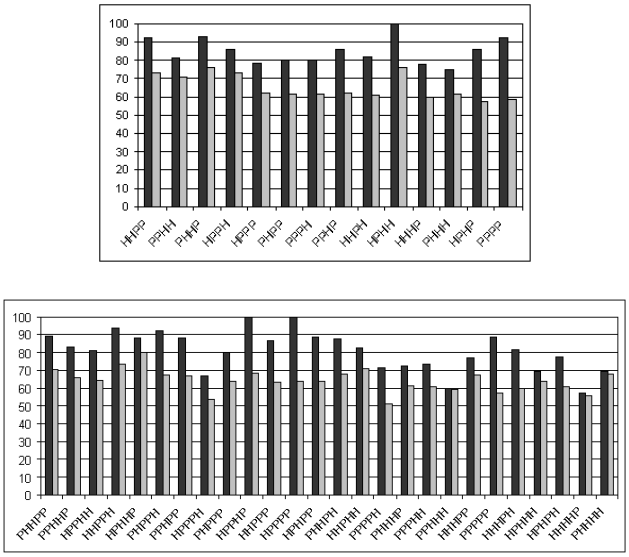

Figure 2. Validation of quartets and pentats within AmPs sequences using Kyte-Doolittle HP Model in terms of PCP(dark grey) and PCN(light grey).

4. The role of PH domains in inositol phosphates binding

PH domains have been identified in more than 500 regulatory proteins, as reported by SMART database [15]. Mainly, eight different proteins structures are known for PH domains: pleckstrin, which was the protein where this domain was firstly detected; spectrin; dynamin; phospholipase C (PLC); Son of Sevenless 1 (SoS1); β-adrenergic receptor kinase (β-Ark); Bruton's tyrosine kinase (Btk), and insulin receptor substrate 1 (IRS-1). PH domain bind to different phosphoinositide polyphosphates and inositol polyphosphates has been systematically examined, revealing a wide range of ligand affinity and specificity. [8]

Pleckstrin, the major protein kinase C substrate of platelets, contains domains of about 100 amino acids at the amino and carboxy termini that have been found in a number of proteins, including serine/threonine kinases, GTPase-activating proteins, phospholipases and cytoskeletal proteins. These conserved sequences, termed pleckstrin-homology (PH) domains, are thought to be involved in signal transduction. But the details of the function and binding partners of the PH domains have not been well characterized [18].

The common PH domain fold, is characterized by a core 7-stranded β-sandwich that is closed off at one splayed corner by the amphipathic C-terminal α helix, and at the other splayed corner by the $\beta1/\beta2$, $\beta3/\beta4$, and $\beta6/\beta7$ loops [4].

Nowadays interactions among PH domain amino acids and inositol phosphate are not well characterized. A clear role of PH Domain in inositol phosphate interaction could shed in light the importance of this domain in *signal transduction*: there are many pathways in which inositol phosphate is implied.

For example human Akt is one of the most important PHD protein, well studied in structure biology and in cell signaling. Human Akt, has an N-terminal pleckstrin homology (PH) domain that binds to the lipid products of phosphoinositide 3-kinase (PI3K), phosphatidylinositol-3,4-bisphosphate [PI(3,4)P2] and phosphatidylinositol-3,4,5-trisphosphate [PI(3,4,5)P3]. This binding happens at plasma membrane level, where human Akt becomes phosphorylated by phosphoinositide-dependent kinase 1 (PDK1). This phosphorylation leads to Akt activation. Akt family member can induce phosphorylations on several factors called Forkhead-related transcription factors. Their role is fundamental in transcriptional of a specific group of genes, including insulin-like growth factor binding protein 1 (IGFBP-1), glucose-6-phosphatase, and phosphoenolpyruvate carboxykinase (PEPCK). When these factors are phosphorylated by Akt family members, their transcriptional responses is inhibited [3].

Many other proteins containing PH Domain, are not directly involved in the biology of phosphoinositides. Most of these proteins are, however, closely associated with membranes in cells. [1]

Some PH molecules can bind inositol phosphates at a characteristic corner of the PH domain which is formed by the $\beta1/\beta2$, $\beta3/\beta4$, and $\beta6/\beta7$ loops. The typical four phosphate groups of inositol phosphate are bound in the central part of the positively charged face of each PH domain [4]. Structural information of this binding can be contained inside the patterns recurrences obtained by HP Pattern–Based

grammars identification.

In this section are presented the results of our proposed approach on a set of 81 sequences of a specific Pleckstrin Homology domain from PDB archive [2]. In our experiments the chosen threshold in the formula 2 was 90%.

In the crystallography of Dapp1, which is considered one of the most known PH domain, the *binding pocket* is characterized by the following primary sequence:

KxxGxVKTxxxR

which is one of the typical PH sequences implied in inositol phosphates binding. This sequence was obtained by 1FAO crystallography, which represents the structure of the pleckstrin homology domain from Dapp1/phish in complex with inositol 1,3,4,5- tetrakisphosphate [4] (see figure 3).

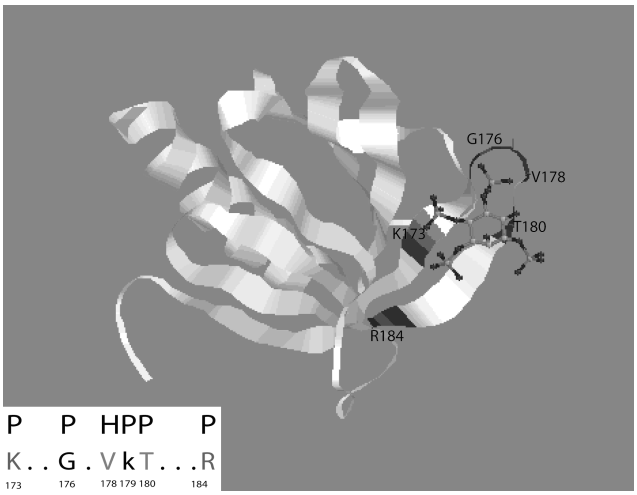

Figure 3. PH pocket of Dapp1: a representation of the interaction between PH domain and inositol phosphate.

This binding pocket expressed by the HP model is

PxxPxHPPxxxP

and it is one of the pattern discovered by our Kyte–Doolittle HP Pattern–Based Grammar shown in figure 4. This grammar was constructed setting the parameters of the Teiresias algorithm as follow: the maximum number of literals in the patterns was set to 12 and the number of non wild-card to 6.

Since all PH domain proteins do not bind the inositol, our HP Pattern–Based Grammar has been validated on a group of 21 proteins, which are known as inositol binding PH proteins. Once obtained the HP grammars, we confronted them with the 17 crystallographies containing the binding (see figure 5). A list of the PH domains crystallographies binding inositols is the following: 1BTN, 1FAO, 1FB8, 1FGY,

$\mathbf{P} \mapsto E(19\%)|S(13\%)|\mathbf{K\ (11\%)}|G(9\%)|Q(9\%)|R(8\%)|D(7\%)|N(7\%)|P(4\%)|Y(4\%)$
x
x
$\mathbf{P} \mapsto E(17\%)|K(16\%)|\mathbf{G\ (13\%)}|R(8\%)|Q(7\%)|H(7\%)|P(5\%)|T(5\%)|Y(5\%)|D(4\%)$
x
$\mathbf{H} \mapsto \mathbf{V\ (27\%)}|L(23\%)|A(12\%)|F(11\%)|I(10\%)|C(8\%)$
$\mathbf{P} \mapsto \mathbf{K\ (18\%)}|G(11\%)|R(10\%)|E(10\%)|Q(9\%)|Y(8\%)|D(6\%)|P(6\%)|T(5\%)|S(5\%)$
$\mathbf{P} \mapsto S(13\%)|K(13\%)|R(13\%)|E(11\%)|Y(8\%)|G(7\%)|\mathbf{T\ (7\%)}|H(6\%)|D(4\%)|N(4\%)|P(4\%)$
x
x
x
$\mathbf{P} \mapsto S(13\%)|\mathbf{R\ (13\%)}|E(12\%)|K(9\%)|Q(8\%)|H(7\%)|G(6\%)|P(6\%)|D(6\%)|N(5\%)|Y(4\%)$

Figure 4. HP Pattern Based Grammar obtained by means of Kyte-Doolittle HP model which characterize the binding pocket KxxGxVKTxxxR of Dapp1.

1FGZ, 1FHW, 1FHX, 1H10, 1MAI, 1U27, 1U29, 1U2B, 1UNQ, 2P0D, 2P0F, 2P0H, 2UVM.

The analysis has shown that the HP pattern describes the essential information for the description of the inositol binding pocket in 18 over 21 proteins. Here, all HP chemical descriptions seems to be implied in interactions with inositol atoms, and at the same time, they contribute to the stability and conservation of folding of all the proteins. The alignment in figure 5, has been obtained by means of software SSM (Secondary Structure Matching) [9] for the secondary and tertiary analysis, and then visualized with ClustalW [16]. The presence of gaps inside of the alignment is related to the primary sequence variability, among some of these proteins. Firstly, the possibility to single out HP information relative to inositol binding, suggests us a new approach in order to evidence structural information independently from the primary sequences of these proteins. Secondly, this analysis, brings to light the relationship structure/function of the PH protein family, due to the presence of some fundamental amino acids in the inositol interaction.

In an evolutionary sense, HP Pattern–Based grammars supply us information about amino acids substitutions. The fact that probabilities of equal substitutions among various residues exists, could give us the ability to synthesize different sequences, which could share high probability to work in the same way.

5. Conclusions

Our work points out some fundamental aspects regarding the relationship between sequence, structure and function of proteins. Using HP Pattern–Based Grammars, we easily understand that a certain biological function can independently be described from the primary sequence of a protein.

HP Pattern–Based Grammars put in evidence chemical conservations among the sequences of a binding domain as

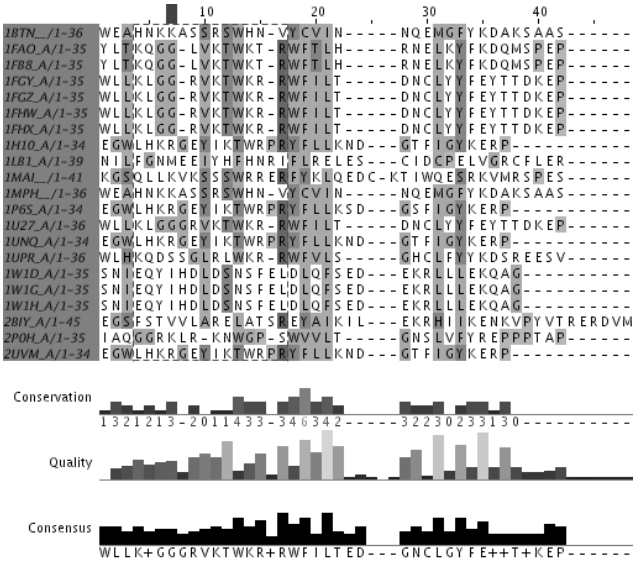

Figure 5. Alignment of the 21 PH proteins, which are known as inositol binding proteins, obtained by means of software SSM and visualized with ClustalW.

the PH pocket which binds the inositol phosphates. Such chemical conservation, is fundamentally important when we study the 3D structure of a domain.

Considering our preliminary results, in future we would apply our model in predictions of other PH domain binding sites. Our approach could represent a new way to single out hidden PH domain functions inside unexpected PH proteins, or proteins in which the PH domain was not highlighted.

In a study of Haslam et all (1993) [6], the authors stated that there is a PH domain comparable in many respects with SH2 and SH3 domains, which are binding domains for phosho-Tyrosine. These observations rose from a series of alignment of protein sequences similar to the N- or C-terminal domains of pleckstrin. Interestingly, some proteins with SH2 and SH3 domains, such as ras-GA, also contain a PH domain. It means that, in some cases, a convergent evolution could be happened during separation of this two different domains [6].

Our approach could allow us to analyze some of these evolutionary aspects, centralizing our attention on the binding of the PH domain proteins with inositol. If we can observe similarities comparing different protein domains, it means that a common information among protein structures exists; and it seems that, sometimes, such structures interact in a more wide molecular vision than how we can imagine.

If Nature is so redundant, we should obtain some advantages from this redundance. In fact, the aim of this work goes beyond the single identification of new, not canonical, PH domains in proteins.

On the other hand our study supplies a new methodology of survey, which considers structure biology as a fresh and powerful instrument of acquaintance, and at the same time, our approach, throws new light on protein evolution. The power to design new peptides oriented to specific targets, using our acquaintances on the relationship structure-function of catalytic domains, as PH domain, derives from how much information on protein chemistry are not stored inside the DNA, but at protein level.

The main idea, is to apply these knowledge to peptide design and therapeutic targets, and create antagonists of normal proteins containing PH domains in the binding to inositol. We explained in section 4 how these factors are implied in surviving, so, for example, we could modulate the molecular response of Akt family members using synthesized peptides.

In any case, even if a successful antagonism between synthetic peptides and protein containing PH domains was not directly possible, we are sure that the production of aimed inositol–binding peptides, could improve the research of a specific antagonist for therapeutic aims.

References

[1] V. Bennet and D. M. Gilligan. The spectrin-based membrane skeleton and micron-scale organization of the plasma membrane. *Annu. Rev. Cell. Biol.*, 9:27–66, 1993.

[2] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank". *Nucleic Acids Research*, (28):235–242, 2000.

[3] A. Brunet, A. Bonni, M. Zigmond, M. Lin, P. Juo, L. Hu, M. Anderson, K. Arden, J. Blenis, and M. Greenberg. Akt promotes cell survival by phosphorylating and inhibiting a forkhead transcription factor. *Cell*, 96:857–68, 1999.

[4] K. M. Ferguson, J. M. Kavran, and et al. Structural basis for discrimination of 3-phosphoinositides by pleckstrin homology domains. *Mol Cell*, 6(2):373–384, 2000.

[5] R. E. Hancock and A. Patrzykat. Clinical development of cationic antimicrobial peptides: from natural to novel antibiotics. *Curr. Drug Targets Infect. Disord.*, (2):79–83, 2002.

[6] R. J. Haslam, H. B. Koide, and B. A. Hemmings. Pleckstrin domain homology. *Nature*, 363(27), 1993.

[7] T. P. Hopp and K. R. Woods. Prediction of protein antigenic determinants from amino acid sequences. *PNAS*, (78):3824–3828, 1981.

[8] J. H. Hurley and S. Misra. Signaling and subcellular targeting by membrane-binding domains. *Annu. Rev. Biophys. Biomol. Struct*, 29:4979, 2000.

[9] E. Krissinel and K. Henrick. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*, (60):2256–2268, 2004.

[10] J. Kyte and R. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, (157):105–132, 1982.

[11] C. Loose, K. Jensen, I. Rigoutsos, and G. Stephanopoulos. A linguistic model for the rational design of antimicrobial peptides. *Nature*, (443):867–869, 2006.

[12] J. M. Otaki, S. Ienaka, T. Gotoh, and H. Yamamoto. Availability of short amino acid sequences in proteins. *Protein Science*, (14):617–625, 2005.

[13] I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequences: The teiresias algorithm. *Bioinformatics*, (14):55–67, 1998.

[14] G. Rose, A. Gesolowittz, G. Lesser, R. Lee, and M. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, (229):834–838, 1985.

[15] J. Schultz, F. Milpetz, P. Bork, and C. Ponting. Smart, a simple modular architecture research tool: Identification of signaling domains. *PNAS*, (95):5857–5864, 1998.

[16] J. Thompson, D. Higgins, and T. Gibson. Clustal w: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, (22):4673–4680, 1994.

[17] Z. Wang and G. Wang. Apd: the antimicrobial peptide database. *Nucleic Acids Research*, (32):590–592, 2004.

[18] H. S. Yoon, P. J. Hajduk, A. M. Petros, E. T. Olejniczak, R. P. Meadows, and S. W. Fesik. Solution structure of a pleckstrin-homology domain. *Nature*, (369):672 – 675, 1994.

[19] M. Zasloff. Antimicrobial peptides of multicellular organisms. *Nature*, (415):389–395, 2002.