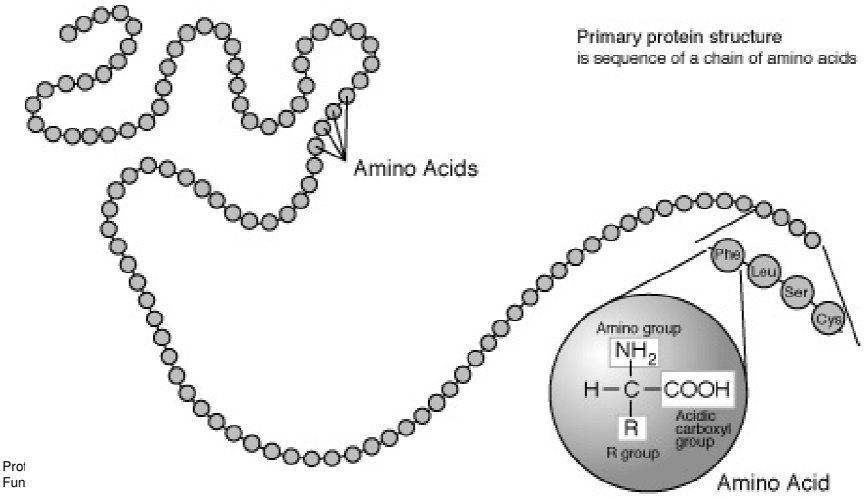
Sequence-Structure-Function Relationships Analyzed by Linguistic Models

Eva Sciacca (sciacca@dmi.unict.it)
University of Catania, Italy

The many functions of proteins

- Rhodopsin: allows vision
- Globins: transport oxygen
- Antibodies: immune system
- Enzymes: pepsin, renin, carboxypeptidase A
- Receptors: transmit messages through membranes
  - And hundreds of thousands more…

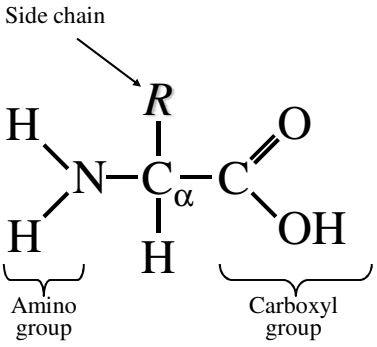
Proteins are chains of amino acids



Prot
Fun

Amino acid composition

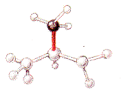
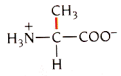
- Basic Amino Acid Structure:
  - The side chain, R, varies for each of the 20 amino acids

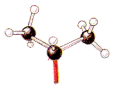


Side chain properties

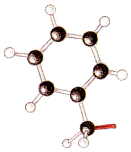
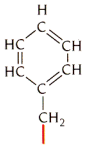
- Recall that the electronegativity of carbon is at about the middle of the scale for light elements
  - Carbon does not make hydrogen bonds with water easily – *hydrophobic*
  - O and N are generally more likely than C to h-bond to water – *hydrophilic*
- We group the amino acids into three general groups:
  - Hydrophobic
  - Charged (positive/basic & negative/acidic)
  - Polar

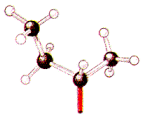
The Hydrophobic Amino Acids

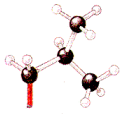
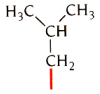


A Ala, Alanine

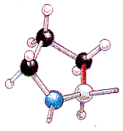
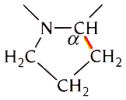


V Val, Valine

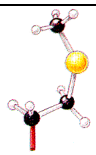
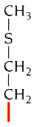


F Phe, Phenylalanine


I Ile, Isoleucine

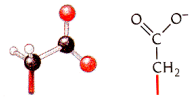


L Leu, Leucine



P Pro, Proline

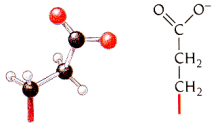
Proline severely limits allowable conformations!



M Met, Methionine

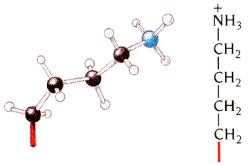
The Charged Amino Acids



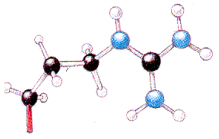
D Asp, Aspartic acid

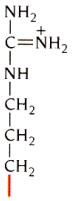


E Glu, Glutamic acid



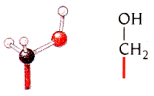
K Lys, Lysine



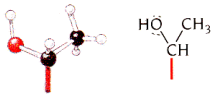


R Arg, Arginine

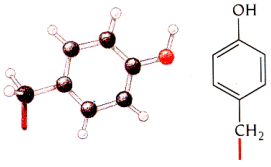
The Polar Amino Acids



S Ser, Serine

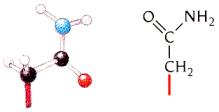


T Thr, Threonine

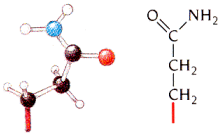


Y Tyr, Tyrosine

C Cys, Cysteine



N Asn, Asparagine

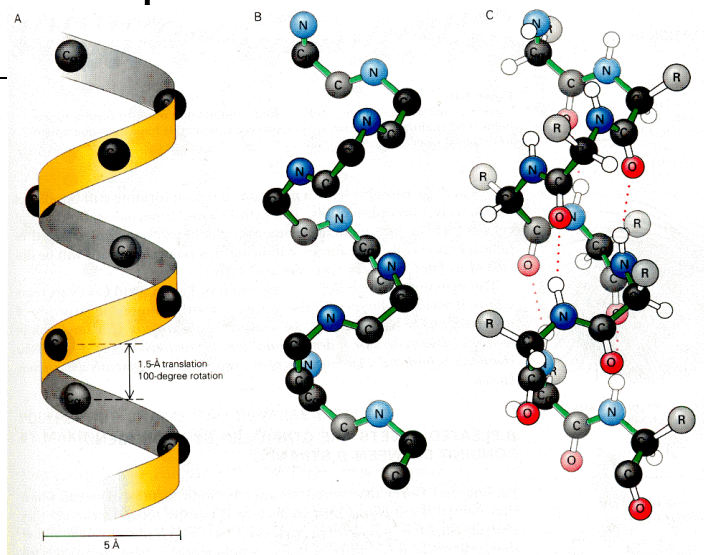


Q Gln, Glutamine

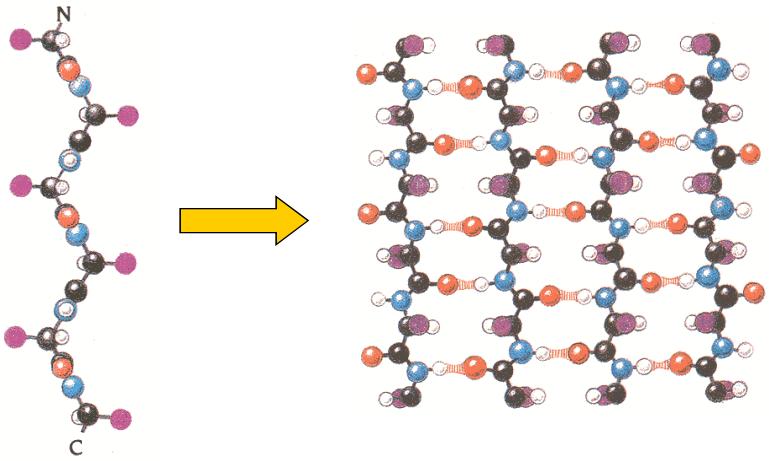
Primary & Secondary Structure

- Primary structure = the linear *sequence* of amino acids comprising a protein:
  AGVGTVPMTAYGNDIQYYGQVT…
- Secondary structure
  - Regular patterns of hydrogen bonding in proteins result in two patterns that emerge in nearly every protein structure known: the α-*helix* and the β-*sheet*
  - The location of direction of these periodic, repeating structures is known as the *secondary structure* of the protein

The alpha helix

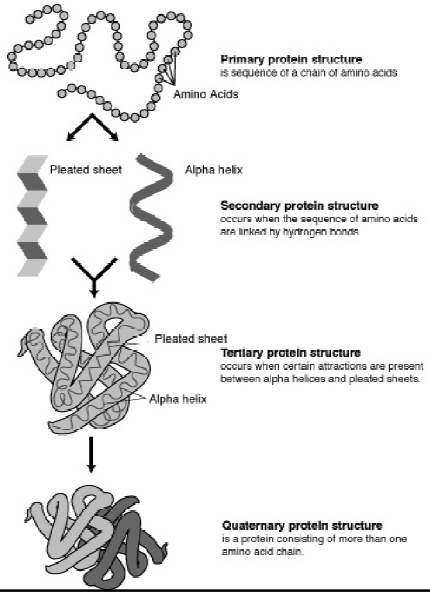


The beta strand (& sheet)

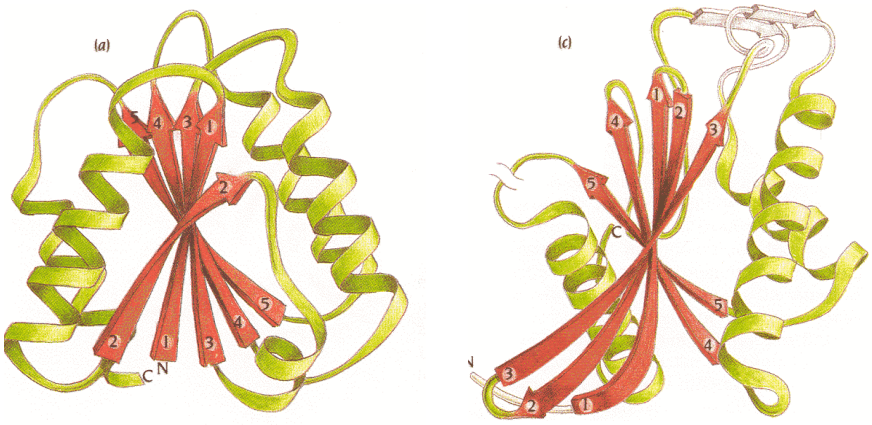


Levels of Protein Structure

- Secondary structure elements combine to form tertiary structure
- Quaternary structure occurs in multienzyme complexes
  - Many proteins are active only as homodimers, homotetramers, etc.

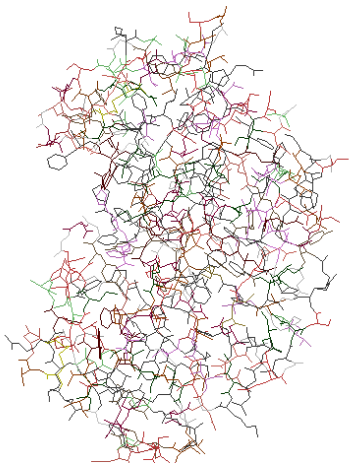


Protein Structure Examples

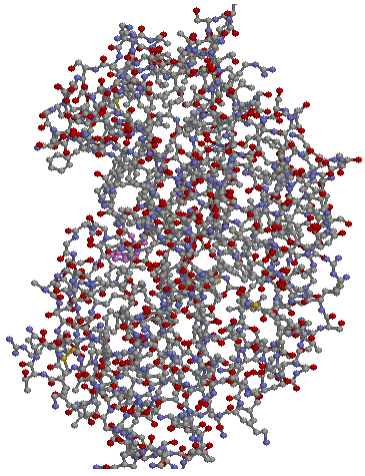


Views of a protein

Wireframe

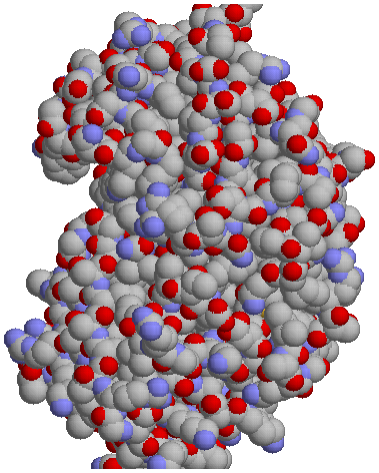


Ball and stick



Views of a protein

Spacefill



Cartoon

CPK colors

Carbon = green, black, or grey

Nitrogen = blue

Oxygen = red

Sulfur = yellow

Hydrogen = white

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

IEEE International Conference on BioInformatics and BioMedicine Workshops, 2007, pp. 53-59.

Giuseppe Nicosia

Eva Sciacca (sciacca@dmi.unict.it)

University of Catania, Italy

Luca Zammataro

IRCC

University of Turin, Italy

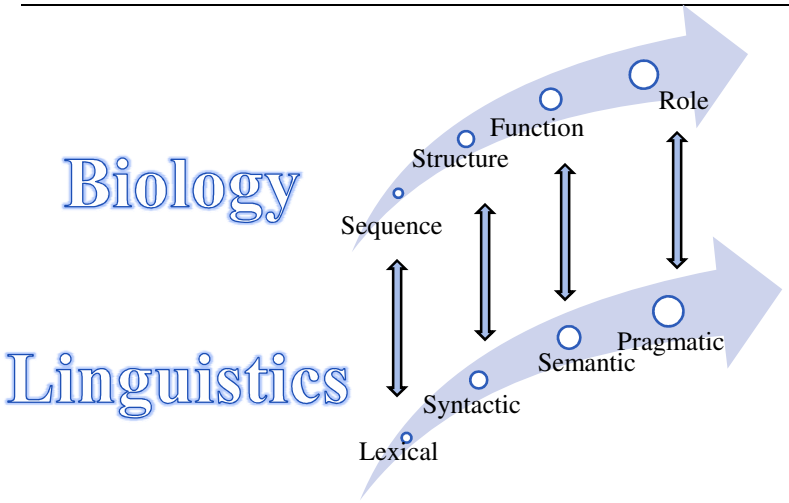
Outline

- Introduction
- Linguistic models for the design of proteins
- The methodology:
  - HP Pattern-Based Grammar
- AmPs Test Banch
- PH domains: inositol phosphates binding
- Conclusions

Introduction

- The discovery of protein sequences similarity in the primary structure usually corresponds to residues ***conserved*** during evolution due to an important ***structural*** or ***functional*** role.
- The analysis of biological sequences is a crucial task to synthesize ***new*** artificial protein sequences with specific 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.

Linguistic models for the design of proteins



D.B.Searls, Reading the Book of Life, 2001, Bioinformatics 17(7) 579-580

Linguistic models for the design of proteins

- ***Linguistic metaphors*** have been woven into the basics of molecular biology since its inception. In fact, many techniques used in bioinformatics, may be seen to be grounded in linguistics.
- Protein sequences as a ***formal language***: a set of sentences using words from a fixed ***vocabulary*** (i.e. the set of naturally occurring amino acids).
- The language of a class of proteins could be described by a set of ***regular grammars***: simple ***rules*** that describes the allowed arrangements of words.

The Methodology

- Every protein sequence belonging to a specific class of proteins has been represented in a ***formal language***.
- In order to face with the structure of the proteins, we used the ***HP model*** and every sequence was "translated" into the binary sequence of H and P
- We deduced the ***HP patterns*** (rigid motifs) of the translated sequences using ***TEIRESIAS***[*] algorithm.

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

TEIRESIAS Algorithm

• TEIRESIAS searches for patterns consisting of characters of the alphabet Σ and wild-card characters '.'.

• **Ambiguous Character** is a character corresponding to a subset of Σ.

Ex. A-[LF]-G

• **Wild-card or Don't care** is a special kind of ambiguous character that matches any character in Σ.

• **Flexible Gap** is a gap of variable length. Ex. X(4,6) matches any gap with length 4,5 or 6. X(I) denotes a fixed gap of length I.

(L,W) Patterns

- Pattern P is a (L,W) pattern iff
  - P is a string of characters from **Σ** and wild cards '**.**'.
  - P starts and ends with a character from Σ
  - Any sub pattern of P(i.e subsequence starting and ending with a character from Σ) containing exactly **L** non-wildcard characters has length of at most **W**.

  Ex. For L=3 and W=5

  AF..C

Algorithm

- Idea: If a pattern P is a (L,W) pattern occurring in at least K sequences, then its sub patterns are also (L,W) patterns occurring in at least K sequences.

- Necessary Condition: K >= 2

Two Phases

The algorithm works in two phases.

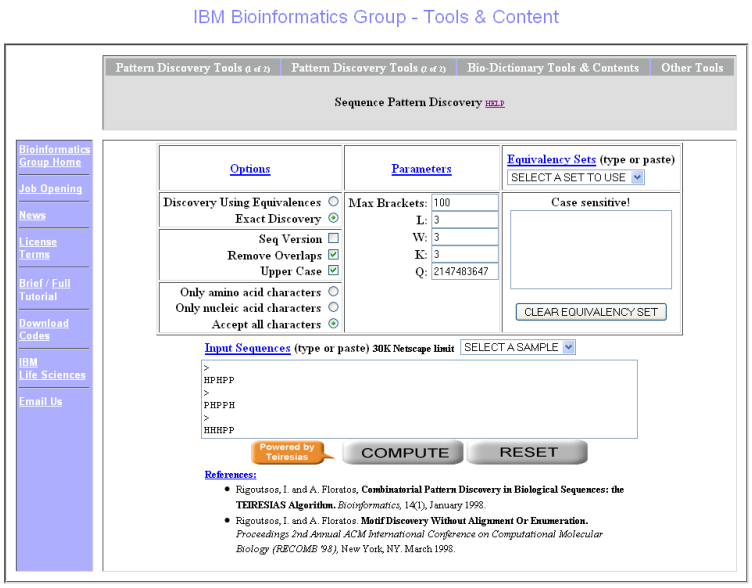
Scanning phase: it finds all (L,W) patterns occurring in at least K sequences that contain exactly L non-wildcards.

Pruned Exhaustive Search:

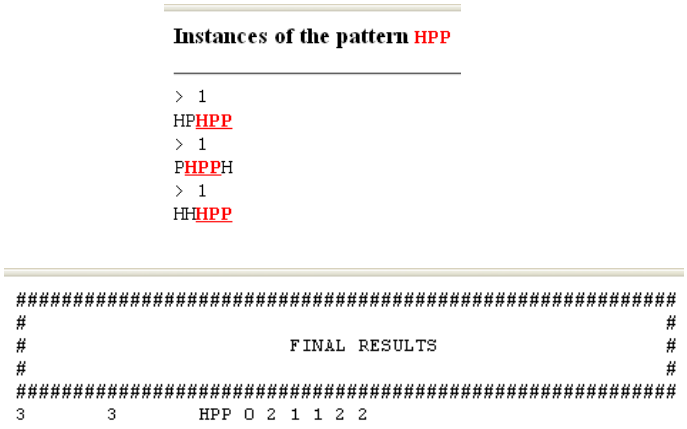
- find a short pattern that appears in K input sequences
- extend them until the support doesn't go below K
- once we find pattern that cannot be extended further, we can say that the patters in maximal and can be written to output.

Convolution phase: For each elementary pattern P, try to extend the pattern with other elementary patterns

Snapshots: http://cbcsrv.watson.ibm.com/Tspd.html



Snapshots(Contd….)

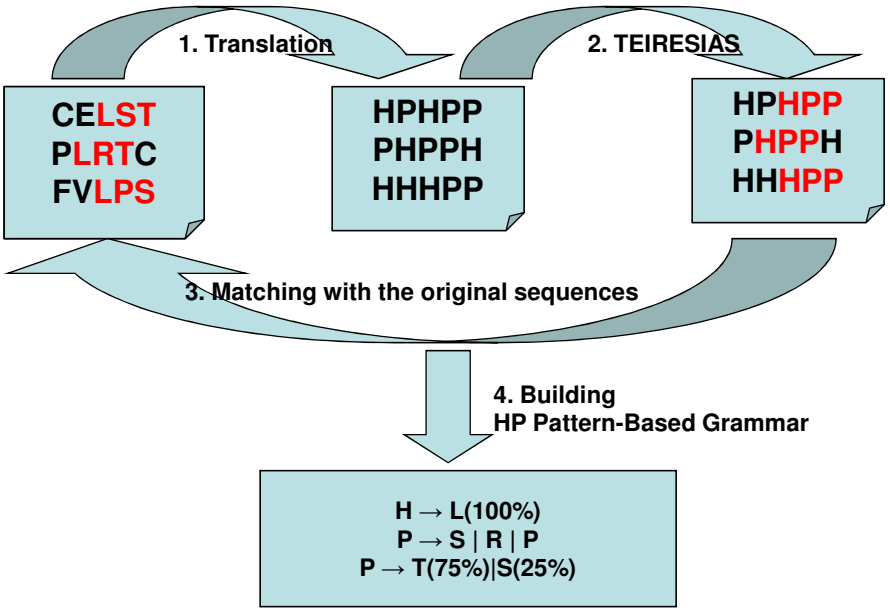


HP Pattern-Based Grammar

- For every ***HP pattern*** of length ***L*** of the translated sequences we identified the derivation rules which bind every H and P of the HP pattern with the amino acids ***aa*** in the correspondent patterns of the original sequences.
- These amino acids ***aa*** are coupled with their corresponding frequencies ***f*** in which they appear within the patterns of the original sequences.
- Every ***derivation rule*** is in the form:

$$H|P \rightarrow (aa_1; f_1) \ldots (aa_n; f_n)$$

- To discard the less frequent amino acids and to consider only the most frequent, the sum of the frequencies of the considered amino acids was chosen at least equal to a ***threshold*** cut–off value.



Validation Stage: AmPs

- ***Antimicrobial Peptides*** (AmPs) are small proteins that are used by the innate immune system to combat bacterial infection in multicellular eukaryotes.
- Data set: 526 well characterized eukaryotic AmPs from APD.
- We validated and compared our resulting grammar set with the set of regular grammars designed by [*] which was used to create new, unnatural AmPs sequences.

E.g.
P[KAYS] [ILN] [FGI]C [KPSA] [IV] [TS] [RKC][KR]

[*] C. Loose, K. Jensen, I. Rigoutsos, and G. Stephanopoulos. A linguistic model for the rational design of antimicrobial peptides. *Nature*, 2006.

Validation Stage: Method

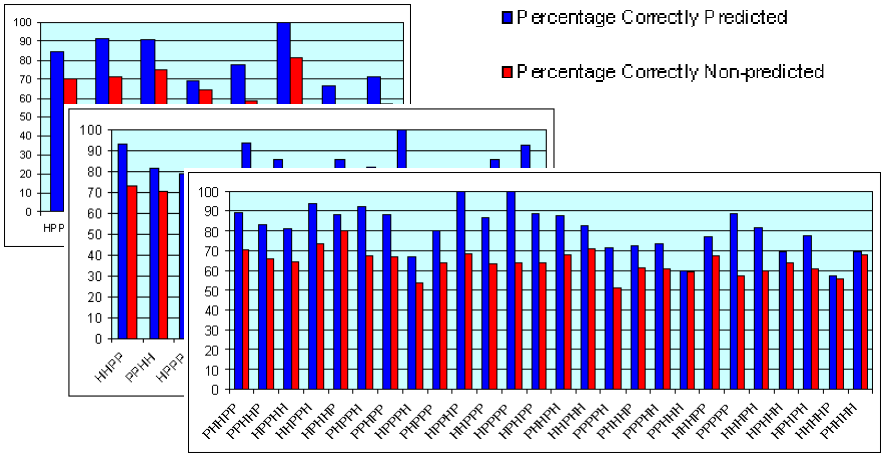
- We examined all possible combinatorial sets of three, four, and five amino acids in the form of H and P (triplets, quartets and pentats), collectively called "***constituent sequences***"[*]. Setting the maximum number of literals in the patterns and the number of non wild-card of Teiresias algorithm to three, four and five.
- Information on 3D structures and functions exists in the context of connections of short constituent sequences and proteins are composed of evolutionary selected constituent sequences[*].

[*] J.M. Otaki, S. Ienaka, T. Gotoh, and H. Yamamoto. Availability of short amino acid sequences in proteins. *Protein Science*, 2005

Validation Stage: Comparison

- In order to compare the two grammars we deduced the HP Pattern-Based derivation rules by means of the grammars built by [Loose et al.].
- Let us call **M** the set of the HP Pattern-Based Grammars built from scratch 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 [Loose et al.] (i.e. the reference data).
- We performed the comparison showing the PCP (the ***sensitivity***) and the PCN (the ***specificity***).

Validation Stage: Results



PH domains: inositol phosphates binding

- ***Pleckstrin Homology(PH) domains***, is a domain of about 100 residues that occurs in a wide range of proteins involved in ***intracellular signaling*** but the details of the function and binding partners of these domains have not been well characterized.
- PH domains are remarkably dissimilar with respect to their primary structure. However they have a ***characteristic secondary and tertiary fold*** consisting of an up and down beta barrel of 7 antiparallel β strands and a C-terminal α helix.

- A clear role of PH Domain in ***inositol phosphate interaction*** shed in light the importance of this domain in ***signal transduction***: there are several pathways in which inositol phosphate is implied.
- Data set: 81 sequences of PH Domain from PDB archive.

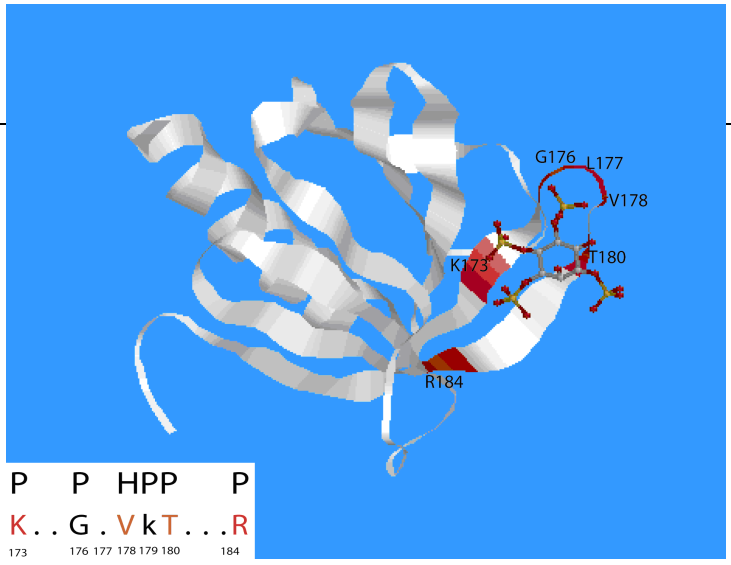
PH domains: inositol phosphates binding

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

KxxGxVKTxxxR

- The typical four phosphate groups of ***inositol phosphate*** are bound in the central part of the positively charged face of each PH domain.

[*] 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.



$\mathbf{P} \mapsto E(19\%)|S(13\%)|\mathbf{K}\ \mathbf{(11\%)}|G(9\%)|Q(9\%)|R(8\%)|D(7\%)|N(7\%)|P(4\%)|Y(4\%)$

x

x

$\mathbf{P} \mapsto E(17\%)|K(16\%)|\mathbf{G}\ \mathbf{(13\%)}|R(8\%)|Q(7\%)|H(7\%)|P(5\%)|T(5\%)|Y(5\%)|D(4\%)$

x

$\mathbf{H} \mapsto \mathbf{V}\ \mathbf{(27\%)}|L(23\%)|A(12\%)|F(11\%)|I(10\%)|C(8\%)$

$\mathbf{P} \mapsto \mathbf{K}\ \mathbf{(18\%)}|G(11\%)|R(10\%)|E(10\%)|Q(9\%)|Y(8\%)|D(6\%)|P(6\%)|T(5\%)|S(5\%)$

$\mathbf{P} \mapsto S(13\%)|K(13\%)|R(13\%)|E(11\%)|Y(8\%)|G(7\%)|\mathbf{T}\ \mathbf{(7\%)}|H(6\%)|D(4\%)|N(4\%)|P(4\%)$

x

x

x

$\mathbf{P} \mapsto S(13\%)|\mathbf{R}\ \mathbf{(13\%)}|E(12\%)|K(9\%)|Q(8\%)|H(7\%)|G(6\%)|P(6\%)|D(6\%)|N(5\%)|Y(4\%)$

PH domains: inositol phosphates binding

- Our HP Pattern–Based Grammar has been validated on a group of 21 proteins, which are known as ***inositol binding*** PH proteins.
- HP pattern describes the essential information for the description of the inositol binding pocket in 18 over 21 proteins.
- HP Pattern–Based Grammars put in evidence ***chemical conservations*** among the sequences of a binding domain as the PH pocket which binds the inositol phosphates.
- Such chemical conservation, is fundamentally important when we study the 3D structure of the domain.

Conclusions

- Using ***HP Pattern–Based Grammars***, we understand that a certain biological ***function*** can be described ***independently from the primary sequence*** of a protein.
- 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.
- The ***power*** to design new peptides oriented to specific targets, using our acquaintances on the relationship sequence-structure-function of catalytic domains, as PH domain, derives from how much information on ***protein chemistry*** is not stored inside the DNA, but at the protein level (secondary and tertiary structure).

Thank you for your kind attention!