Biog5: A bioinformatic system for the analysis of the Human Papillomavirus

Ernesto Rafael Pérez[1], Sofía Erdozain[2], Leonardo Gómez Chávez[1], Germán Conti[1], Emilio Angelina[1], Nélida Peruchena[1]

[1]Laboratorio de Estructura Molecular y Propiedades (LEMyP), Instituto de Química Básica y Aplicada del Nordeste Argentino, (IQUIBA-NEA).

[2]Facultad de Ciencias Exactas y Naturales-Universidad de Buenos Aires

Introduction

Currently, there are more than 200 Human Papillomavirus strains (HPV), out of which at least 14 can cause cancer. HPV16 y HPV18 are the main variants found in cervix cancer and in precancerous lesions and they are covered by available vaccines. However, there are several less common strains that can still cause cancer and be resistant to vaccines (Table 1).

Biog5 is a system that integrates multiple computational biology and bioinformatics tools to find relationships between HPV strains. It relies on available sequence and structural data of viral proteins L1, L2, E1, E2 and E7 (Table 2) for making predictions that could be useful to address the virulence of less prevalent strains.

To perform the predictions, Biog5 relies on the premise that VPH strains with higher potential for inducing cancer should be genetically related, i.e. they should cluster togheter in a philogenetic tree.

Table 1. HPV strains

Risk	Strains*
High	**16**, **18**, **31**, **33**, 35, 39, **45**, 51, **52**, 56, **58**, 59, 68, 73, 8
Low	**6**, **11**, 40, 42, 43, 44, 54, 61, 70, 72, 81
Undetermined	26, 53, 66

* vaccine available in Argentina **(green)**, vaccines available abroad **(blue)**

Table 2. HPV proteins

Proteins	Function
L1, L2	Viral capsid proteins
E6, E7	Promote uncontrolled cellular proliferation and inhibit apoptosis by blocking p53 and pRB.
E1, E2, E5	Viral DNA replication and transcription

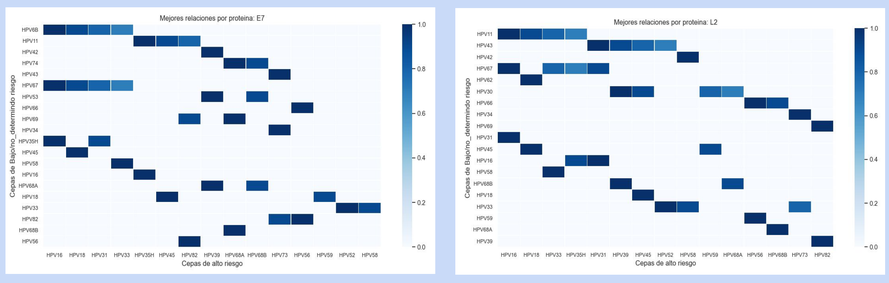
METHODS

Biog5 is implemented under Python 3, it consists in several bioinformatic tools that can be accessed *via* an interactive user interface, in the command-line. An iterative incremental model that divides the system's functionality into small increments called sprints was employed for software development. This software process model, known as Agile Scrum, provides great flexibility, adaptability and integration in various project modules.

FIGURE 1. BIOG5 structures

Paired Alignments

- All available strains are aligned, one against the other, using Blastp, subdivided into risk groups and proteins, then the ones that bought the best bitcore values are filtered and plotted.



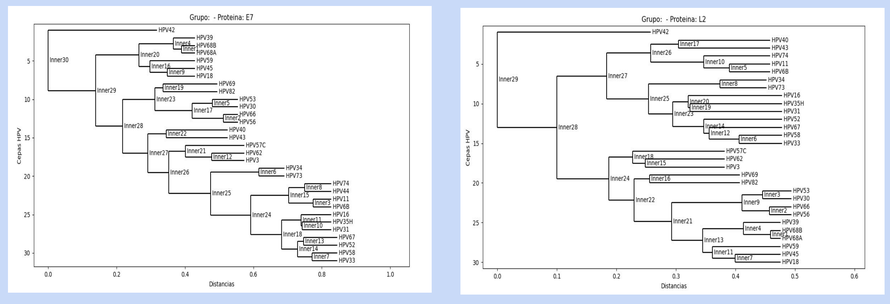
@biog5

```
####################################
###### Bienvenidos a BIOG5 ######
####################################

Que desea realizar:
1- Alineamientos de a pares usando Blast
2- Alineamientos multiples usando Clustal Omega
3- Modelos HMM
4- Analisis de variantes en proteinas
5- Analisis de variantes en genomas
6- Estructuras NCBI
7- Configuar Riesgos
8- Salir
Ingrese una opción:
```

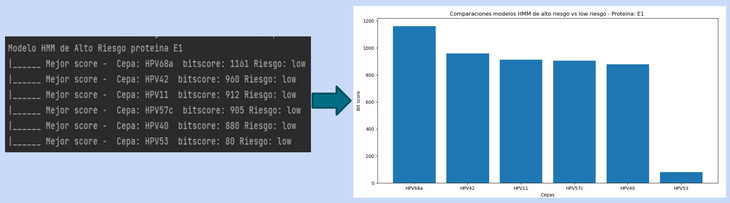
Multiple Sequence Alignments (MSA)

- Protein groups are aligned by risk and protein type, using Clustal Omega, then associated phylogenetic trees are generated.



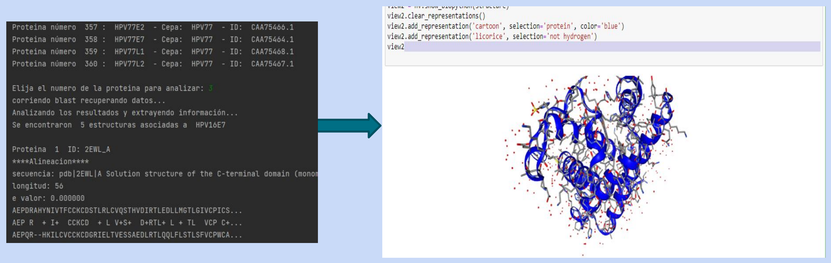
Hidden Markov Models (HMM)

- It uses results from the biog5 pipeline developed for Linux that generates 5 high-risk HMM models (E1, E2, E7, L1, L2), against which the other strains are compared, and then those that obtained the best bitcore values are filtered and graphed.



NCBI structures

- Searches the NCBI server for structures (pdb) aligning fasta sequences of selected proteins of interest, then allows them to be downloaded and viewed.



Variants in genomes and proteins

- A multiple alignment is performed by proteins and risk groups, or whole genomes. From each group of aligned sequences, allows you to filter amino acids by an entered percentage of conservation.

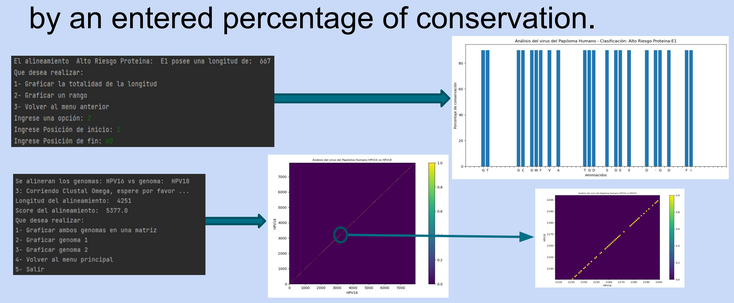


FIGURE 2. BIOG5 functionalities

Results

Alignment of test HPV sequences against entire genomes and individual proteins of HPV strains of well-known phenotype yields the following results:

* Sequence comparison of capsid proteins (i.e. L1 and L2) suggest a possible relationship with strain prevalence rather than with risk level. For instance, low risk HPV11 strain (it causes genital warts) shows high similarity with high risk HPV16 strain, both have in common a high population prevalence.

* On the other hand, similarity in oncoproteins E6 and E7 can be linked to risk level for developing cancer. Alignment modules demonstrate similarities between E6 and E7 proteins from the same risk group. For example, HPV11 strain is clearly classified as low risk, based on E7 sequence.

Conclusions

The trend and accessibility of sequencing technologies hint at their potential use within healthcare systems. With this in mind, Biog5 was conceived as a tool that could be useful for epidemiological surveillance or for prognosis of disease outcome.