PHYSICOCHEMICAL n-GRAMS TOOL (PnGT) TUTORIAL

Introduction

Physicochemical n-Grams Tool is an open source standalone software for calculating physicochemical descriptors of protein developed using the Python scripting language.

The algorithms used in the development of this tool are based on the n-Grams curation techniques which are mainly used for natural language processing. Apart from language processing, now a days n-Grams modelling is also making a room in the computational biology and genomics.

The software currently calculates 33 physicochemical descriptors and the sequence length for given protein primary sequence. To find more specific information about our tool, please contact us.

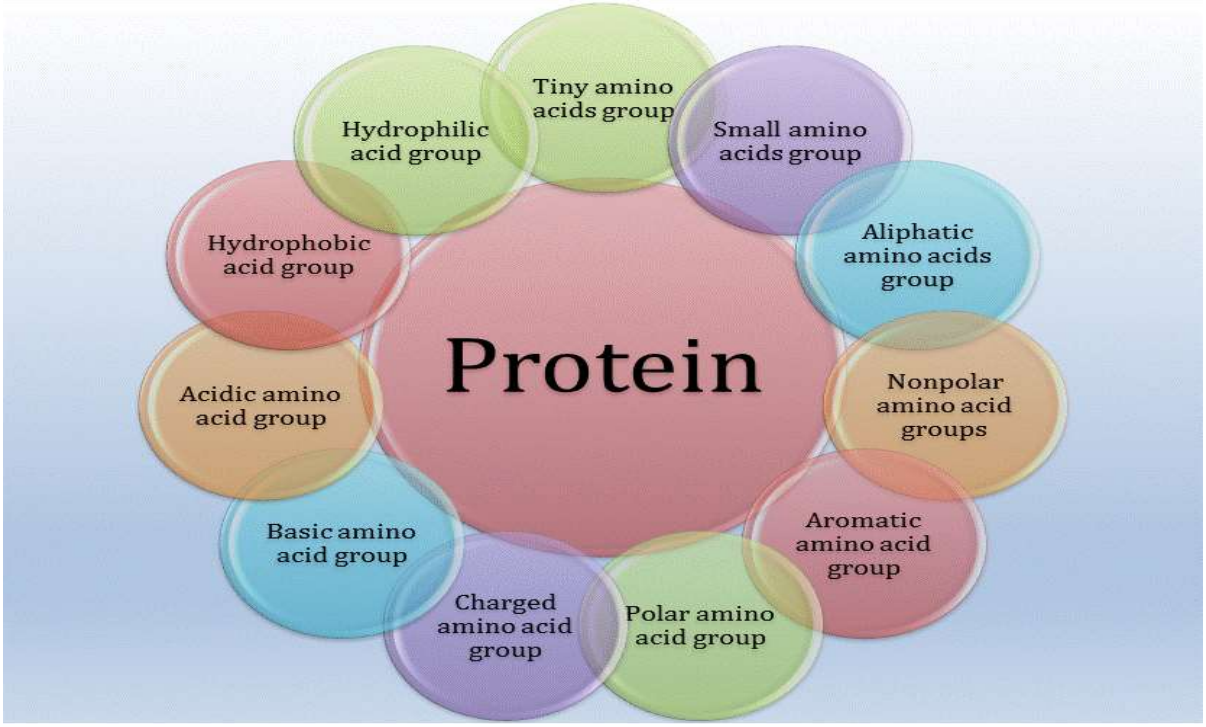

Figure.1 Physicochemical Amino Acids Property Groups

Utility: The tool is able to assess the **physicochemical properties** of the protein in an easier way with bi-gram, tri-Grams, tetra-Grams assessment of protein primary sequence.

Assessment of the physicochemical n-Grams properties of the protein: With the help of this PnGT tool, the analysis of proteomics data in the term of physicochemical n-Grams statistics is easy-to-do with the only primary sequence as feeding.

Implementation in machine learning model development as feeding: The outcomes in terms of physicochemical feature vectors from the Physicochemical n-Grams Tool can be directly implemented for the development of statistical (PCA, Regression Analysis) or machine learning model like SVM, *k*NN *etc.*

Keywords: Proteomics science, physicochemical parameter, n-Grams, machine learning, statistical modelling.

Physicochemical n-Grams Descriptors and Calculations: Conservation of multiple physicochemical groups along the sequence in a sliding window of length n residue. For example, physicochemical 2-grams: Aliphatic++; where n equals to 2

〖if aa〗_i ∈S^* AND 〖aa〗_(i+1)∈S^*

where, aai = ith residue of the protein

Table 1. Physicochemical property groups of amino acids

Sr. no.	Name of amino acid property group	Amino acids in the specific group (S*)
1.	Tiny amino acids group	Ala, Cys, Gly, Ser, Thr
2.	Small amino acids group	Ala, Cys, Asp, Gly, Asn, Pro, Ser, Thr, Val
3.	Aliphatic amino acids group	Ile, Leu and Val
4.	Nonpolar amino acid groups	Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp, Tyr
5.	Aromatic amino acid group	Phe, His, Trp, Tyr
6.	Polar amino acid group	Asp, Glu, His, Lys, Asn, Gln, Arg, Ser, Thr
7.	Charged amino acid group	Asp, Glu, His, Arg, Lys
8.	Basic amino acid group	His, Lys, Arg
9.	Acidic amino acid group	Asp, Glu
10.	Hydrophobic acid group	Ala, Cys, Phe, Ile, Leu, Met, Val, Trp, Tyr
11.	Hydrophilic acid group	Asp, Glu, Lys, Asn, Gln

Installation Steps for User

1. PnGT-Installer: To download “PnGT-Installer” go to Download tab and download PnGT-Installer to install Physicochemical n-Grams Tool.

2. Extract the zip file and open it, then locate “PnGT” as shown in Figure 2 below.

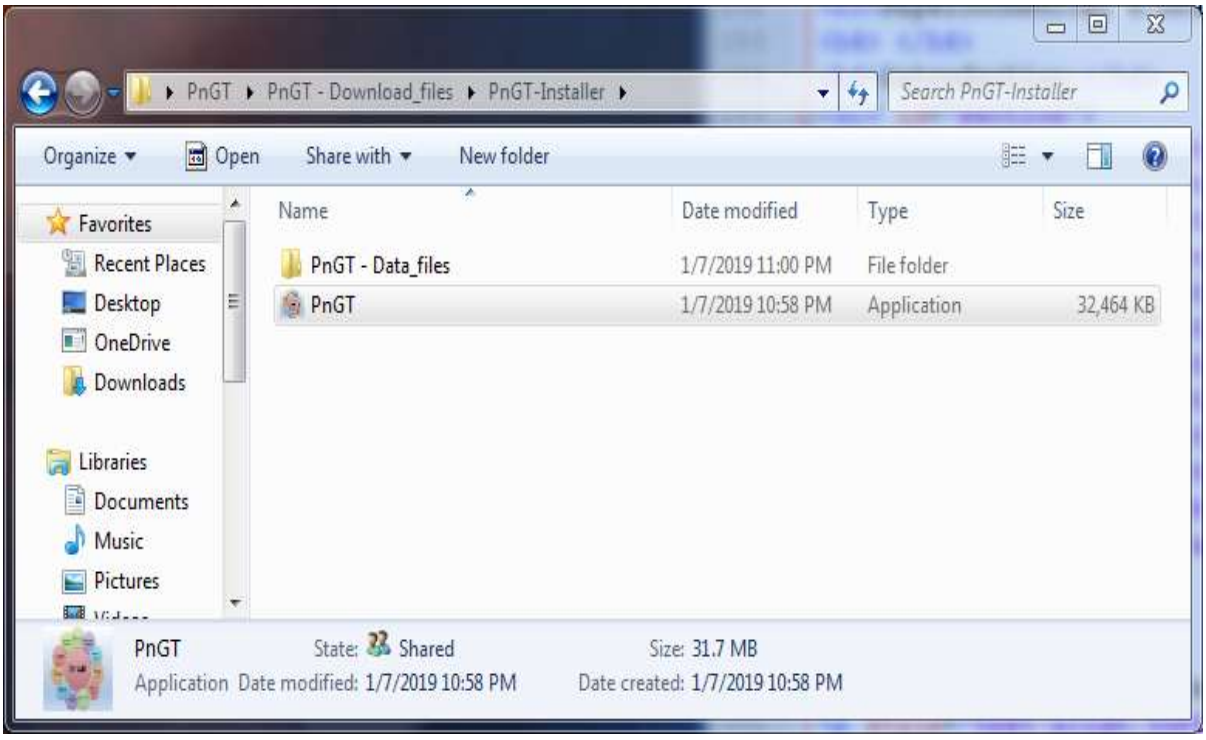

Figure.2 PnGT Installer

Home Page

1. The PnGT will open as shown in Figure 3 below.

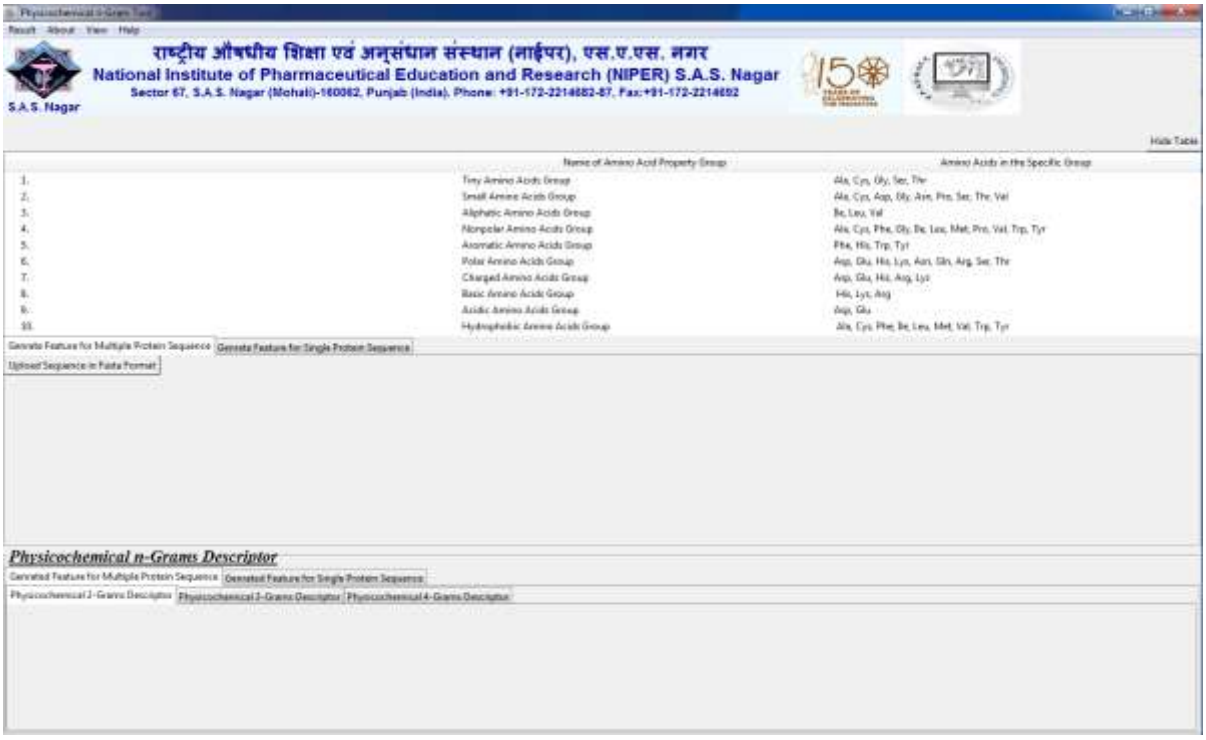

Figure.3 PnGT Home Page

Physicochemical n-Grams Descriptors Calculation

- The tool is leveraged with two input method: FASTA format and simple sequence pasting method, if the user doesn't have the primary sequence in FASTA format.

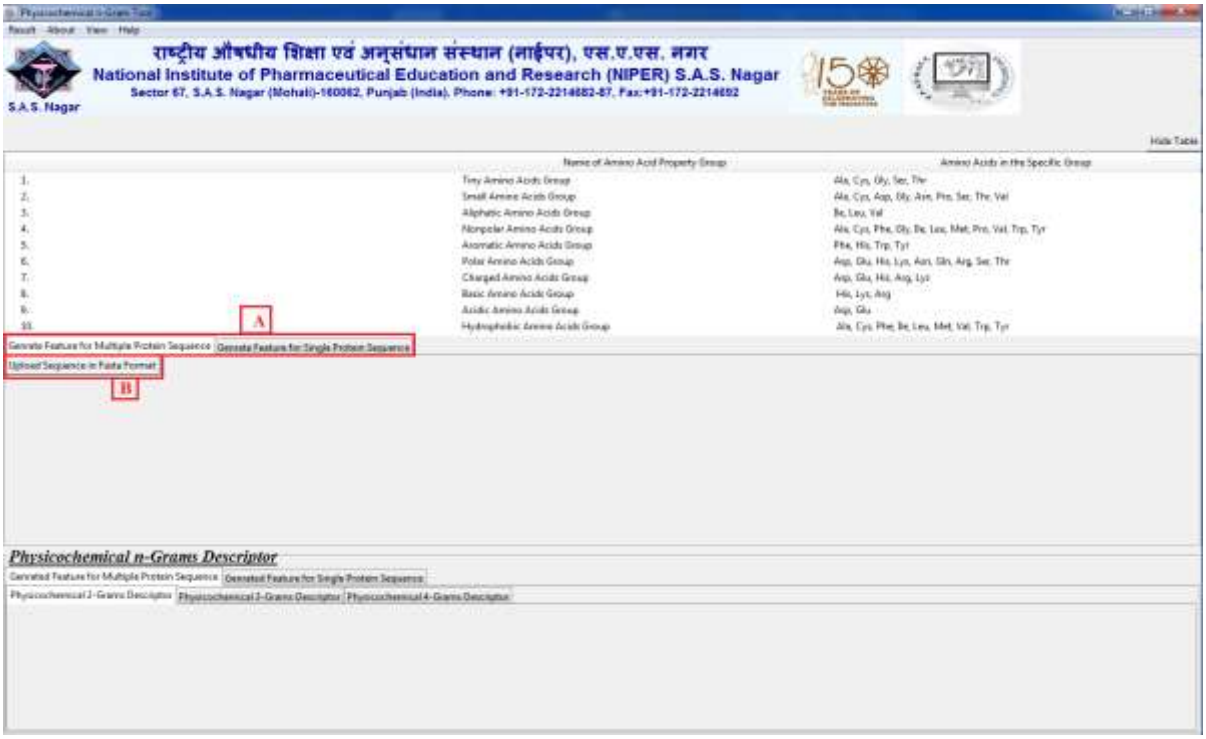

Figure.4A

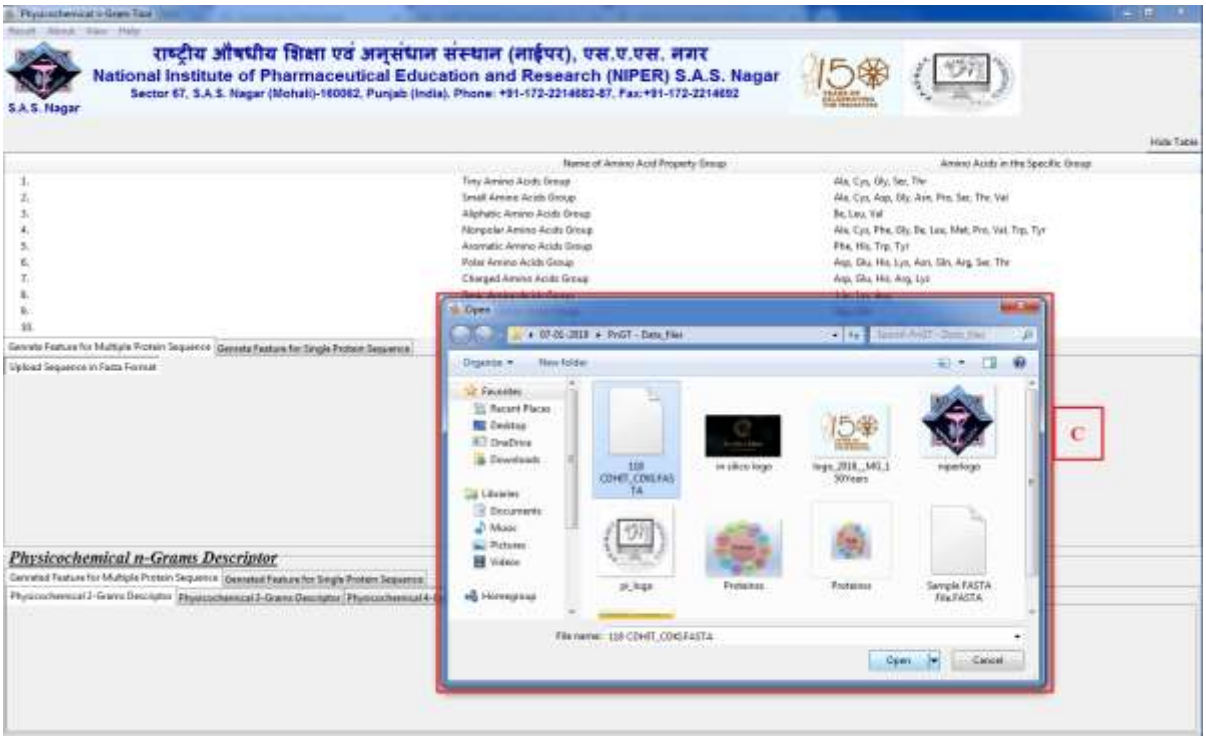

Figure.4B

Figure.4C

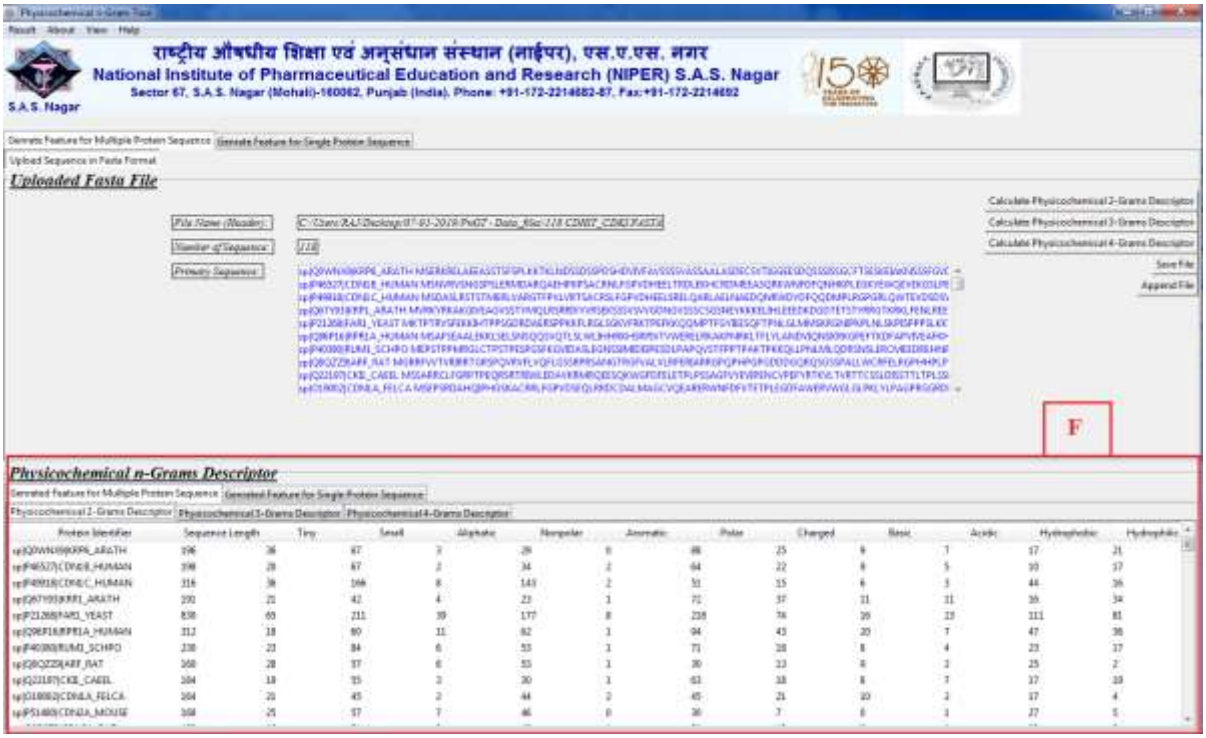

Figure.4D

- The user can download the feature vector generated through physicochemical n-Grams assessment in CSV file format (by default). Figure 5 showing output descriptor in CSV file format.

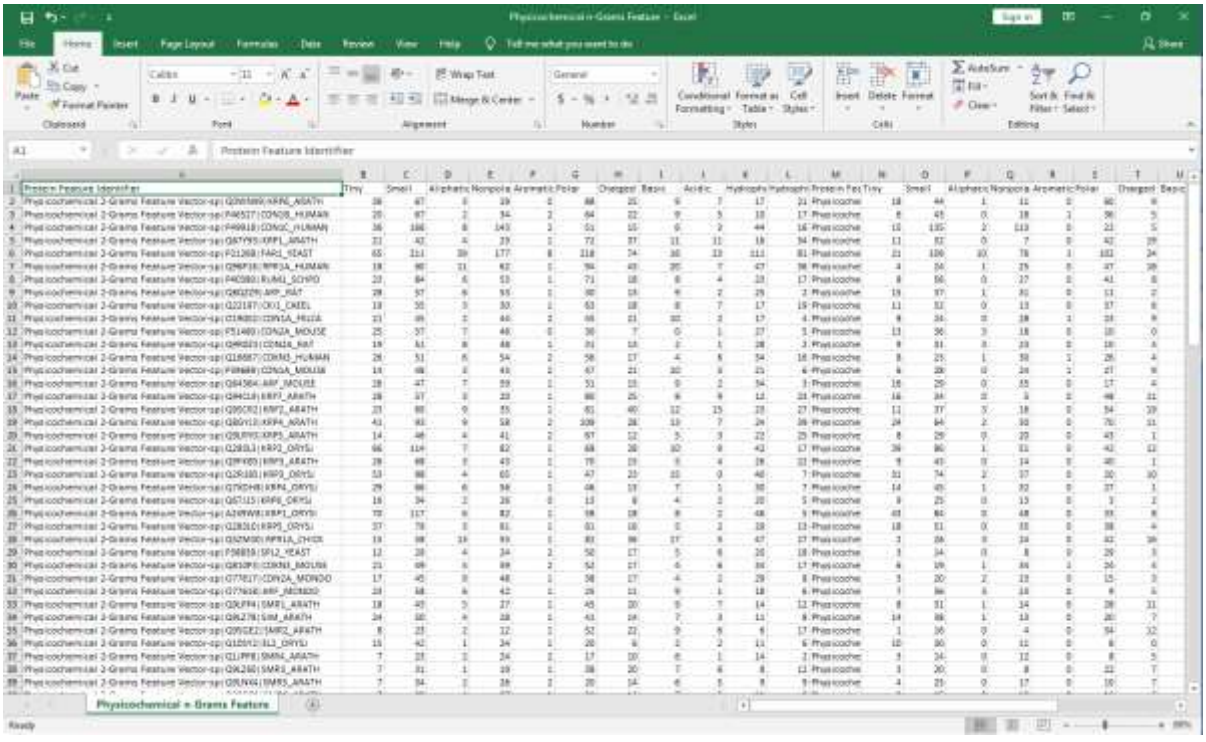

Figure.5 Figure showing output descriptor in CSV file format; Only one click for CSV file generation (Save File).

- The PnGT tool consists of a menu bar, a graphical control element that contains few selectable drop-downs for user convenient such as Physicochemical Properties, References *etc.*

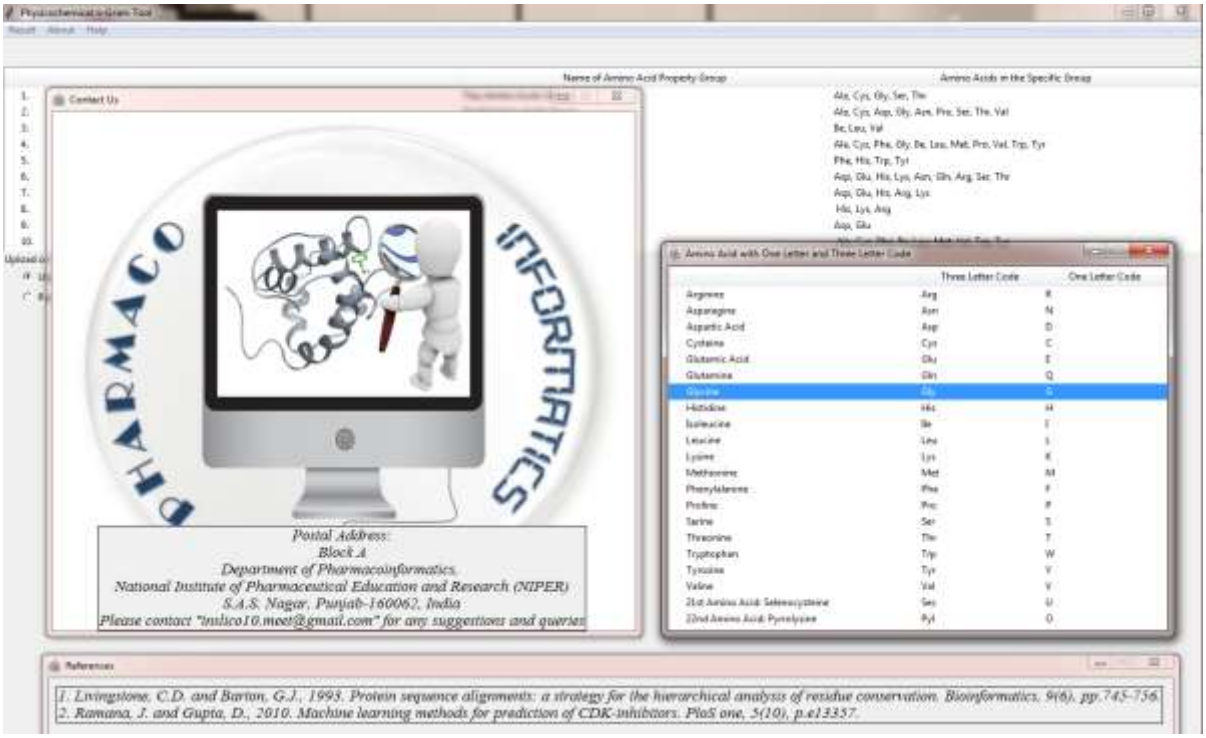

Figure.6 The Snapshot showing widget of About Menu: Amino Acid and widgets of Help Menu: Contact Us & References.