BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 10(2)

PHYSICO: An UNIX based Standalone Procedure for Computation of Individual and Group Properties of Protein Sequences

Parth Sarthi Sen Gupta[1], Shyamashree Banerjee[1], Rifat Nawaz Ul Islam[1], Sudipta Mondal[1], Buddhadev Mondal[2] & Amal K Bandyopadhyay[1]*

[1]Department of Biotechnology, The University of Burdwan, Golapbag, Burdwan, 713104, West Bengal, India; [2]Department of Zoology, Burdwan Raj College, The University of Burdwan, Golapbag, Burdwan, 713104, West Bengal, India; Amal K Bandyopadhyay - Email: akbanerjee@biotech.buruniv.ac.in; Phone - +91-342-2657231(O), 9474723882(M), Fax - +91-3422657231; *Corresponding author

Received January 20, 2014; Accepted January 26, 2014; Published February 19, 2014

Abstract:
In the genomic and proteomic era, efficient and automated analyses of sequence properties of protein have become an important task in bioinformatics. There are general public licensed (GPL) software tools to perform a part of the job. However, computations of mean properties of large number of orthologous sequences are not possible from the above mentioned GPL sets. Further, there is no GPL software or server which can calculate window dependent sequence properties for a large number of sequences in a single run. With a view to overcome above limitations, we have developed a standalone procedure i.e. PHYSICO, which performs various stages of computation in a single run based on the type of input provided either in RAW-FASTA or BLOCK-FASTA format and makes excel output for: a) Composition, Class composition, Mean molecular weight, Isoelectic point, Aliphatic index and GRAVY, b) column based compositions, variability and difference matrix, c) 25 kinds of window dependent sequence properties. The program is fast, efficient, error free and user friendly. Calculation of mean and standard deviation of homologous sequences sets, for comparison purpose when relevant, is another attribute of the program; a property seldom seen in existing GPL softwares.

Availability: PHYSICO is freely available for non-commercial/academic user in formal request to the corresponding author (akbanerjee@biotech.buruniv.ac.in).

Keywords: UNIX; physicochemical properties; FASTA; software; protein sequence.

Background:
In post genomic era, bioinformatics taking the lion share of research in the field of computational biology. Physico-chemical properties of amino acids (building block of protein) can be used to study protein sequence profiles, structures and functions **[1].** Various types of chromatographic procedures need apriori information about physicochemical properties such as molecular weight, net charge and binding affinity for the isolation of proteins. In general, Isoelectric point, Molecular weight, Aliphatic index and GRAVY are most popular among all physicochemical properties. However, these values do not reveal anything about the code of proteins i.e. the sequence property. Sequence properties such as Polarity, hydrophobicity, hydrophilicity, relative mutability, bulkiness, refractivity and others, which are computed using experimentally determined amino acids index values, are most important. Determination of physicochemical properties of unknown proteins and finding its correlation with known ones is very important as it can provide insight into their structure and hence function. Three dimensional structures of proteins, which are more conserve,

are determined by more diverse amino acids sequences **[2].** Several online GPL softwares are present which analyze physicochemical properties of proteins. Most authentic among them is from ExPASy **[3].** However, major bottle neck in this is that one has to input one sequence at a time for computation of above properties, which is time consuming and limited by the speed of internet. Further, managing outputs may create confusion. Propas **[4]** a standalone software, which can perform calculation on a number of sequences but only of three kinds of physicochemical properties. Thus, available software tools suffer from one or the other limitations. We, therefore, have developed PHYSICO which not only computes 25 kinds of sequence properties and physicochemical properties of large set of protein sequences but also computes mean properties as well as standard deviation of all input sequences for comparison. Our program thus would be helpful for researchers in the field of bioinformatics and computational biology.

Program input:

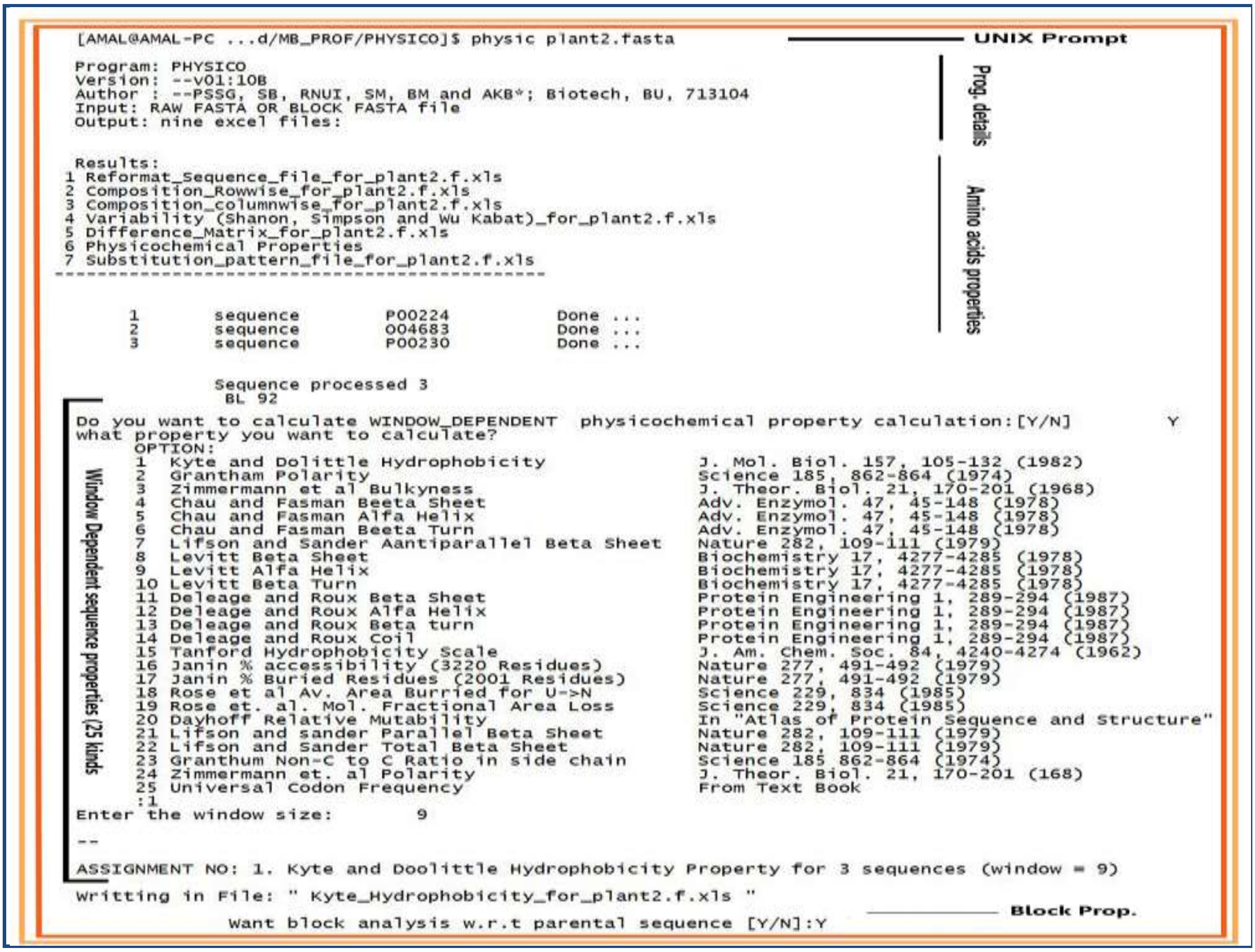
PHYSICO takes protein FASTA file (raw or block) as input. User needs to be on the same directory where the FASTA file is located. In the command prompt, the program is to be run with the name of FASTA file as argument. Upon thorough check of the input file, PHYSICO starts computation and output redirection.

Preparation of RAW-FASTA file

In this, full length sequence is to be kept in one line, following the title line (FASTA FORMAT). Any number of sequences of any length can be kept in the file. Sequences are to be retrieved from an authentic database and fed in a local text file along with header line.

Preparation of BLOCK-FASTA file

In block FASTA file all sequences are of same length. Analysis of this file is useful for comparison among different orthologous sets adapted in different environment (such as mesophilic, thermophilic etc). Firstly a raw FASTA file is made using all available orthologous sequences as above and then it is subjected for multiple sequence alignment. Indels are to be removed and FASTA format is to be made. The block FASTA file thus obtained can be subdivided based on environment specific proteins (thermophilic or mesophilic etc) or can be separated in the same FASTA file by putting a specific marker (such as ##1, ##2 etc) just after > in the header line. Program can automatically identify the marker if it exists and process accordingly.



```
 [AMAL@AMAL-PC ...d/MB_PROF/PHYSICO]$ physic plant2.fasta

 Program: PHYSICO
 Version: --v01:10B
 Author : --PSSG, SB, RNUI, SM, BM and AKB*; Biotech, BU, 713104
 Input: RAW FASTA OR BLOCK FASTA file
 Output: nine excel files:


 Results:
1 Reformat_Sequence_file_for_plant2.f.xls
2 Composition_Rowwise_for_plant2.f.xls
3 Composition_columnwise_for_plant2.f.xls
4 Variability (Shanon, Simpson and Wu Kabat)_for_plant2.f.xls
5 Difference_Matrix_for_plant2.f.xls
6 Physicochemical Properties
7 Substitution_pattern_file_for_plant2.f.xls
-----------------------------------------


      1      sequence      P00224         Done ...
      2      sequence      004683         Done ...
      3      sequence      P00230         Done ...


             Sequence processed 3
              BL 92

 Do you want to calculate WINDOW_DEPENDENT  physicochemical property calculation:[Y/N]        Y
 what property you want to calculate?
      OPTION:
      1  Kyte and Dolittle Hydrophobicity                J. Mol. Biol. 157, 105-132 (1982)
      2  Grantham Polarity                               Science 185, 862-864 (1974)
      3  Zimmermann et al Bulkyness                      J. Theor. Biol. 21, 170-201 (1968)
      4  Chau and Fasman Beeta Sheet                     Adv. Enzymol. 47, 45-148 (1978)
      5  Chau and Fasman Alfa Helix                      Adv. Enzymol. 47, 45-148 (1978)
      6  Chau and Fasman Beeta Turn                      Adv. Enzymol. 47, 45-148 (1978)
      7  Lifson and Sander Aantiparallel Beta Sheet      Nature 282, 109-111 (1979)
      8  Levitt Beta Sheet                               Biochemistry 17, 4277-4285 (1978)
      9  Levitt Alfa Helix                               Biochemistry 17, 4277-4285 (1978)
      10 Levitt Beta Turn                                Biochemistry 17, 4277-4285 (1978)
      11 Deleage and Roux Beta Sheet                     Protein Engineering 1, 289-294 (1987)
      12 Deleage and Roux Alfa Helix                     Protein Engineering 1, 289-294 (1987)
      13 Deleage and Roux Beta turn                      Protein Engineering 1, 289-294 (1987)
      14 Deleage and Roux Coil                           Protein Engineering 1, 289-294 (1987)
      15 Tanford Hydrophobicity Scale                    J. Am. Chem. Soc. 84, 4240-4274 (1962)
      16 Janin % accessibility (3220 Residues)           Nature 277, 491-492 (1979)
      17 Janin % Buried Residues (2001 Residues)         Nature 277, 491-492 (1979)
      18 Rose et al Av. Area Burried for U->N            Science 229, 834 (1985)
      19 Rose et. al. Mol. Fractional Area Loss          Science 229, 834 (1985)
      20 Dayhoff Relative Mutability                     In "Atlas of Protein Sequence and Structure"
      21 Lifson and sander Parallel Beta Sheet           Nature 282, 109-111 (1979)
      22 Lifson and Sander Total Beta Sheet              Nature 282, 109-111 (1979)
      23 Granthum Non-C to C Ratio in side chain         Science 185 862-864 (1974)
      24 Zimmermann et. al Polarity                      J. Theor. Biol. 21, 170-201 (168)
      25 Universal Codon Frequency                       From Text Book
      :1
 Enter the window size:        9

 --

 ASSIGNMENT NO: 1. Kyte and Doolittle Hydrophobicity Property for 3 sequences (window = 9)

 Writting in File: " Kyte_Hydrophobicity_for_plant2.f.xls "

             Want block analysis w.r.t parental sequence [Y/N]:Y
```

Figure 1: Details and execution of the program in UNIX environment.

ISSN 0973-2063 (online) 0973-8894 (print)

Program output:

PHYSICO writes itemized excel outputs for block or raw FASTA files as input. In block FASTA file homologous position based computations are relevant and thus the program redirects extra relevant outputs for block FASTA files apart from common outputs. **Figure 1** show details of the program.

Caveats and future development:

PHYSICO is written in GAWK (interactive) language which can preferably run in any C shell UNIX prompt and also be made run on B shell Linux and window command prompts. We are developing a web based interface for PHYSICO and other tools developed in our lab.

Acknowledgement:

We thankfully acknowledge the computational facility Laboratory of the Department of Biotechnology, Govt. of India, The University of Burdwan.

References:

[1] Cai CZ *et al. Nucleic Acids Res*. 2003 **31**: 3692 [PMID: 12824396]
[2] Alain J Cozzone, *Nature* 2002 (www.els.net)
[3] Wilkins MR *et al. Methods Mol Biol*. 1999 **112**: 531 [PMID: 10027275]
[4] Songfeng Wu *et al. Bioinformation* 2012 **8**: 167 [PMID: 22368391]

Edited by P Kangueane
Citation: **Gupta *et al.*** Bioinformation 10(2): 105-107 (2014)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited