BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 11(7)

# PHYSICO2: an UNIX based standalone procedure for computation of physicochemical, window-dependent and substitution based evolutionary properties of protein sequences along with automated block preparation tool, version 2

**Shyamashree Banerjee†, Parth Sarthi Sen Gupta†, Arnab Nayek†, Sunit Das, Vishma Pratap Sur, Pratyay Seth, Rifat Nawaz Ul Islam & Amal K Bandyopadhyay***

Department of Biotechnology, 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
†- Authors equally contributed

Received May 08, 2015; Revised June 15, 2015; Accepted June 16, 2015; Published July 31, 2015

**Abstract:**
Automated genome sequencing procedure is enriching the sequence database very fast. To achieve a balance between the entry of sequences in the database and their analyses, efficient software is required. In this end PHYSICO2, compare to earlier PHYSICO and other public domain tools, is most efficient in that it i] extracts physicochemical, window-dependent and homologous-position-based-substitution (PWS) properties including positional and BLOCK-specific diversity and conservation, ii] provides users with optional-flexibility in setting relevant input-parameters, iii] helps users to prepare BLOCK-FASTA-file by the use of *Automated Block Preparation Tool* of the program, iv] performs fast, accurate and user-friendly analyses and v] redirects itemized outputs in excel format along with detailed methodology. The program package contains documentation describing application of methods. Overall the program acts as efficient PWS-analyzer and finds application in sequence-bioinformatics.

**Availability:** PHYSICO2: is freely available at http://sourceforge.net/projects/physico2/ along with its documentation at https://sourceforge.net/projects/physico2/files/Documentation.pdf/download for all users.

**Keywords:** CYGWIN; ABPT; FASTA; BLOCK; Program; Protein sequence.

**Background:**
Candidate sequences of almost every protein in the database belong to a named family (e.g cytochrome c) with many taxonomic groups (e.g. metazoan, cyanobacteriaum etc). Comprehensive statistical analyses of a) physicochemical [**1, 2, 3**], b) window-dependent [**2**], c) homologous position-based substitution properties (PWS) [**4**] and their comparison among different taxonomic groups have been the recent trend in sequence bioinformatics [**1, 2**]. Considering the evergrowing sequence databases of about 6000 genomes, rapid yet accurate PWS-analyses are sought to bring a balance between sequences entry via genome project and their studies.

 

There are different web-tools that perform either physicochemical [**5, 6**] or window-dependent [**6**] analysis on per-sequence basis for one [**5, 6**] or few properties **[6].** Web-tools are also there that use amino acid index values [**7**] for prediction of interaction profile of sequence [**8, 9**] or sequences [**9**] per run. However, web-tools are rare that allow mass-scale, user-friendly analyses of PWS (3-in-1) properties in a single run using any form of FASTA-file. Gaining insight into PWS differential among different taxonomic groups, is of great significance in sequence bioinformatics [**1, 2, 4**], would be computationally costly by analyzing one sequence at a time and then computing the average. Moreover management of analytical and graphical web-data by later procedure is very cumbersome and error-prone. Further, sharing of same web-software by worldwide users might cause lower processivity. While PHYSICO **[10],** in contrast performs batch analyses for PWS properties, their ranges are still less exhaustive. Now the later to serve better, an upgraded version seems urgent such that users i] could relish the flexibility in setting relevant input-parameters, ii] could procure additional outputs on window-dependent profiles for RAW-FASTA, pI-profiles and items in similar kind of output as PHYSICO that contains novel reports on substitution-based positional as well as BLOCK specific diversity and conservation, iii] can access detailed documentation on principle and methodologies used in the program. Although capable in analyzing, PHYSICO is unable to perform painstaking BLOCK-FASTA-file preparation that would not only be necessary for extraction of extra information but also for comparison of novel evolutionary properties among different taxonomic groups of a given family. PHYSICO2 incorporates all the above attributes along with comprehensive up-gradation of PWS properties in reference to earlier version and thus has been a unique tool in sequence bioinformatics.

**Methodology:**

The program works on input FASTA file of any form **Figure 1: (F1 & F2)**. Upon execution it optionally allows to change default input-parameters (DPAR) such as residue classes, pI-method and Shannon-threshold by users one (UPAR). Program then enters into first phase (P1) of analyses. In contrast to BLOCK-FASTA (F2), RAW-FASTA (F1) input produces only one output **(Figure 1:R1)** as in this case homologous positions are non-comparable and thus column specific analyses (that produce additional three outputs in the former one: B2, B3 and B4) are skipped. RAW-FASTA-file (F1) harboring sequences from one or more taxonomic groups that are readily converted into BLOCK-FASTA-file or files of identical width respectively using ABPT (F3) of the program. In second phase of computation (P2), the program performs window-dependent property analysis. In this case if the input is BLOCK-FASTA (where homologous positions are aligned), all sequence specific profiles are redirected into one excel table (R5) to facilitate easy computation of mean along with standard deviation for taxonomically related sequences (Documentation) otherwise each sequence specific profile is saved separately ($R2_1$, $R2_2$ etc) in named directory (Documentation).

**Program input:**

PHYSICO2 is extensively tested to function in CYGWIN (32-bit) environment. It takes either RAW **(Figure 1: F1)** or BLOCK - FASTA (F2) as input. While the former is directly usable upon downloading from the database, the later is to be prepared (either manually or programmatically) prior to its use as input. Unlike PHYSICO where one needs to prepare BLOCK-FASTA-file manually, PHYSICO2 includes ABPT for its preparation (Documentation). Users are also prompted for input-parameters such as residue-classes, pI-method and Shannon-threshold.

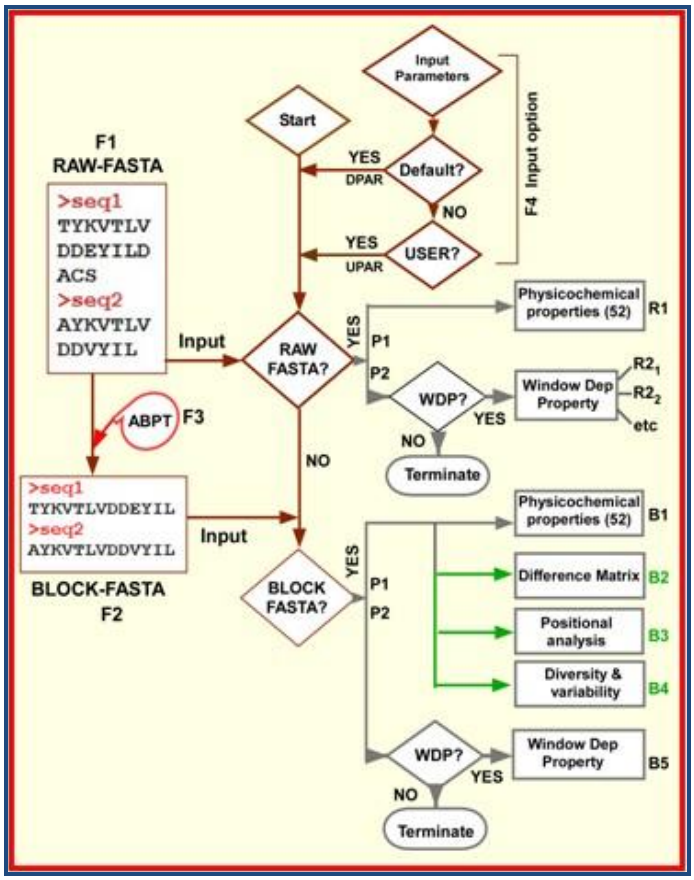


**Figure 1:** Flow chart for functioning of PHYSICO2.

**Method of computation, performance of the program and experimental validation:**

Detailed method precedes analytical results of each item in each output file. We performed PHYSICO2 based analyses on representative candidate sequences from two taxonomic groups (metazoa: 31 and cyanobacteria: 32 sequences) of "cytochrome c family" in Intel(R) core™ i3 CPU M330 @2.13 GHz PC-CYGWIN (32 bit) environment. We also performed same analysis using "PROTPARAM" for physicochemical (8 properties [**6**] and their averages) and "PROTSCALE" (individual and average profile) for window-dependent [**6**] properties using "Alliance Broadband: PRIME (54Mbps) package" internet connection. Efficient use of these tools took ≥ 10 hours for obtaining these results in excel. On the other hand, only 6 minutes was sufficient to obtain the above and other additional properties in PWS-format using PHYSICO2.

To compare itemized-results, same set of sequences were subjected for analysis using available web-tools [**5, 6**] and PHYSICO2. Although not all items of results could be compared due to lack of public domain program (e.g. substitution hetero-pair diversity) above items showed exactly similar results (data not shown).

**Program output:**
PHYSICO2 redirects itemized one output on physicochemical properties for each of BLOCK (**Figure 1**: **B1**) and RAW-FASTA-files (R1). Unlike RAW-FASTA-file, BLOCK-FASTA-file produces homologous position based three additional outputs (B2, B3 and B4). Window-dependent property in case of BLOCK-FASTA-file is in one compact output for all sequences **(Figure 1: B5)** and that are in separate outputs in case of RAW-FASTA-file (**Figure 1: $R1_1$, $R2_2$, etc**). Detailed method also precedes analytical results of each item of each output.

**Caveats and future development:**
PHYSICO2 is written in AWK language that runs from C or B shell of CYGWIN (32-bit) operating system. We are developing GUI-based application for the program.

**Conclusion:**
PHYSICO2, unlike other public domain programs acts as PWS analyzer as 3-in-1 form. Unlike earlier PHYSICO, it not only doubles PWS properties in outputs but also redirects new output; new items in outputs along with detailed methodology from identical input file as earlier. It provides users flexibility for input-parameters and helps auto-preparation of BLOCK-FASTA file.

**Reference:**

**[1]** Jaspard E *et al. PLlos One* 2012 **7**: e36968 [PMID: 22615859]
**[2]** Polyansky AA *et al. Nat Commun.* 2013 **4**: 2784 [PMID: 24253588]
**[3]** Brendel *et al. Proc Natl Acad Sci USA*. 1992 **89** 2002
**[4]** Ladunga I & Smith RF, Protein Eng. 1997 **10**: 187 [PMID: 9153083]
**[5]** Wilkins MR *et al*. *Methods Mol Biol*. 1999 **112**: 531 [PMID: 10027275]
**[6]** Gasteiger *et al. Protein Identification and Analysis Tools on the ExPASy Server;*(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press. 2005.pp. 571-607
**[7]** Kawashima *et al. Nucleic Acids Res* 1999 **27**: 368 [PMID: 9847231]
**[8]** Geourjon & Deleage, *Compute Appl Biosci* 1995 **11**: 681 [PMID: 8808585]
**[9]** Rao HB *et al. Nucleic Acids Res*. 2011 **39:** W385 [PMID: 21609959]
**[10]** Gupta *et al. Bioinformation* 2014 **10**: 105 [PMID: 24616564]


**Edited by P Kangueane**
**Citation**: **Banerjee *et al.*** Bioinformation 11(7): 366-368 (2015)