# GPSRdocker: A Docker-based Resource for Genomics, Proteomics and Systems biology

**Piyush Agrawal[1,2#], Rajesh Kumar[1,2#], Salman Sadullah Usmani[1,2#], Anjali Dhall[1#], Sumeet Patiyal[1], Neelam Sharma[1], Harpreet Kaur[1,2], Vinod Kumar[1,2], Dilraj Kaur[1], Shipra Jain[1], Akshara Pande[1], Sherry Bhalla[1], Gajendra P.S. Raghava[1,2,*]**

[1] Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, India.

[2] Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India

[#] Equal contribution

* Corresponding authors

Gajendra P.S. Raghava, Head Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi 110020, India, Phone: +91-11-2690744

## Abstract

**Background:** In past number of web-based resources has been developed in the field of Bioinformatics. These resources are heavily used by scientific community to provide solution for challenges faced by experimental researchers particularly in the field of biomedical sciences. There are number of challenges in utilizing full potential of these services that includes internet speed, limits on computing power, and security of data. In order to enhance utilities of these web-based assets, we developed a docker-based container that integrates large number resources available in literature.

**Results:** This paper describes GPSRdocker a docker-based container developed for providing wide-range of computational tools in the field of bioinformatics particularly in genomics, proteomics and system biology. Majority of tools integrated in GPSRdocker are based on web services developed at Raghava's group in last two decades. Broadly, these tools can be categorized in three categories; i) general scripts, ii) supporting software and iii) major standalone software. In order to facilitate students or developers working in the field of bioinformatics, we developed general scripts in Perl and Python. These general-purpose scripts serve as building block for any bioinformatics tools like computing features/descriptors of a protein. Supporting software packages includes SCIKIT, WEKA, SVM$^{light}$, and PSI-BLAST; these software packages allow one to develop/implement bioinformatics software. Major Standalone software is core of this container which allows predicting function/class of biomolecules. These tools can be classified broadly in following categories; protein annotation, epitope-based vaccines, prediction of interaction and drug discovery.

**Conclusion:** A docker-based container has been developed which can be easily run on any operating system as well as it can be directly ported on cloud. Scripts can be run to build pipelines for addressing problems at system level like prediction of vaccine candidate for a

43 pathogen. GPSRdocker including manual is available free for academic use from

44 https://webs.iiitd.edu.in/gpsrdocker.

## 1. Introduction

Numerous software packages, libraries and web based services have been developed by bioinformaticians or computational biologist over the years. Only standalone software or libraries has been developed in pre-internet era, most of them were free for public use; mainly called domain software. Following archival sites have been created to maintain these software packages; EMBL File Serve [1], IUBIO archive [2], BioCatalog [3] and PDSB [4]. Our group developed, first scientific program ELISAeq [5] in 1990, for computing antigen/antibody concentration from ELISA data in GW-BASIC [6]. All these programs were standalone programs, developed for DOS/Windows using programming various languages like FORTRAN, PASCAL, C. These programs were distributed free for academic users via floppy, CD or via email-server. Though these programs were user-friendly, but one needs to have a hardware/software compatibility and knowledge of installation, in order to run these programs. In the era of Internet (1995 onwards), most of developers start to develop web based services. These webservers overcome limitations of standalone software packages where user only need to have a computer with browser and access to internet.

In last two decades, a wide range of web based services have been developed by scientific community particularly in the field of biology. Despite, numerous advantages of web based technology it has its own limitations. One of the major limitation of in the era of genomics is to predict function of all proteins or genes at genome level; transfer of huge data over internet is time consuming and costly. In addition, service provides cannot meet the computation requirement of user. Another concern is security of data, user do not wish to transfer confidential data over Internet. In order to provide service to community one need to use old technology of standalone software. There are number of challenges that include compatibilities of codes, dependencies, versions of libraries, compilers. This is nearly impossible for a user to install these software on their local machines. In order to overcome

above limitations, number of projects have been initiated to develop customize operating system for free software in bioinformatics. These projects include BioLinux, VigyaanCD, DNAlinux, NEBC Bio-Linux, Vlinux. These operating systems are mainly flavors of Unix/Linux, like Ubuntu, Red Hat, Debian. A wide range of bioinformatics software packages have been integrated in these packages. Despite numerous advantages of customize operating systems, it consumes lot of computational resources.

In order to provide alternate to VMS, docker based and singularity based containers which are light weight and requires minimum resources became popular. Using these available container numbers of Bioinformatics pipelines have been developed (Eg. DNAp, NGSeasy, LncPipe etc.). In this manuscript, we have developed GPSRdocker (docker based container) that integrates bioinformatics software to perform various tasks. Most of the software in this package are unique and not integrated in any container or VMS so far. These software's has been developed at Raghava's group over the years and their webservers are available on the webpage (https://webs.iiitd.edu.in/gpsrdocker/). This software would be useful for researchers for developing various pipelines on their local machines.

## 2. Implementation

Docker provides a platform to perform operating system level virtualization or containerization. In brief, it provides a platform to develop, employ and run applications within a flexible and lightweight container. Containers are basically a software package, which are isolated from each other and pack their own configuration files, tools and libraries. All the containers can be interconnected for ease communication through well-defined

channels and are actually run by a single operating system kernel. Containers are launched by running an image, which specify their precise contents. An image is basically the executable package constituting essentials needed to run a software i.e. code, libraries, configuration files and environment variables. GPSRdocker is a docker-based container that provides a resource on Genomics, Proteomics and System Biology. Concisely, GPSRdocker, is based on docker suite where customized container of all our webservers are available. User can run GPSRdocker on their machine using following steps

1. Install the docker into your system and create account at docker hub.
2. Make sure the docker is running before installing GPSRdocker.
3. Type following command install: **docker pull raghavagps/gpsrdocker**
4. Run docker on your machine using command: **docker run -i -t raghavagps/gpsrdocker**
5. Above step allow to work in docker image, now user can install software packages using command **"/gpsr/gpsr_install.pl"**. This will allow to users to install packages and provides instruction to run these software packages.

Complete manual on GPSRdocker is available from its web site https://webs.iiitd.edu.in/gpsrdocker/ , user may read manual to install and run software packages.

### 2.1. General Scripts

In this section, we have described small programs developed at our group to generate features which can be used as building block to develop complex prediction modules. These programs are different from existing software libraries or modules like BioPERL, BioPython, as user should have programming skills in order to use these modules/subroutines. In GPSR 2.0 package we have developed small programs, which can be run by any person with minimal

118 knowledge of programming skills. Following are important programs included in this
119 package used to generate features for developing prediction models from protein and
120 DNA/RNA sequences. These programs are developed in PERL and Python following the set
121 standards. In order to run these codes, user needs to have Python 3.0 or above version
122 installed in their system.
123
124 **Table 1: List of the scripts along with their purpose incorporated in the GPSRdocker.**

| Program | Purpose | Program | Purpose |
|---|---|---|---|
| *pro2aac* | Amino Acid composition | *pro2dpc* | Dipeptide composition |
| *pro2tpc* | Tripeptide composition | *fasta2sfasta* | FASTA to single fasta format |
| *blast_similarity* | To perform BLAST | *motif2bin* | Binary input from the multifasta motif file |
| *add_cols* | Add columns of two files | *col2svm* | Generate SVM$^{light}$ input format file |
| *col_mult* | Multiply each column of input file with a number | *col_mult_sel* | Multiply selective columns with a number |
| *col_avg* | Average column of two files | *col_ext* | Extract selective columns from a file |
| *col_corr* | Correlation co-efficient between two column | *perl col_rem* | Remove selective columns from a file |
| *col_sig* | Significance of columns in two column files | *seq2pssm_imp* | Compute PSSM matrix in column format without any normalization |
| *pssm_n1* | To normalize pssm | *pssm_n2* | Normalize pssm profile based |

| | | | |
|---|---|---|---|
| | profile based on Eqn: 1/(1+e-x) | | on Eqn: (numb -min)/(max -min) |
| *pssm_n3* | Normalize pssm profile based on Eqn: (numb -min)*100/(max -min) | *pssm_n4* | Normalize pssm profile based on Eqn: 1/(1+e-(x/100) |
| *pssm_comp* | Compute PSSM composition (400) | *pssm_smooth* | Design smooth pssm profile for plot |
| *pssm2pat* | To generate patterns of given size from PSSM matrix | *seq2motif* | Create motifs by sliding window of user defined length with option of adding terminal X |

## 2.2. Supporting Software

We utilized service of various software for developing and implementing our software. These supporting software include PSI-BLAST [7], CD-HIT [8], LPC [9], PSIPRED [10], machine learning packages like scikit-learn [11], SVM$^{light}$ [12], SNNS [13], WEKA [14]. These software were used for data processing, developing dataset, performing alignment, removing redundancy among sequences, developing machine learning models and implementing them.

## 2.3. Standalone Packages

In this section, we are describing various standalone prediction packages developed in our group. For the ease of user we have classified this section into broad six categories such as 1)

Protein Structure Prediction; 2) Functional annotation of proteins; 3) Vaccinomics; 4) Genomics: Genome annotation and application; 5) BioDrugs: Biomolecules based therapeutics; 6) Interactome: Biomolecular based therapeutics. We have categorized our prediction packages in these classes.

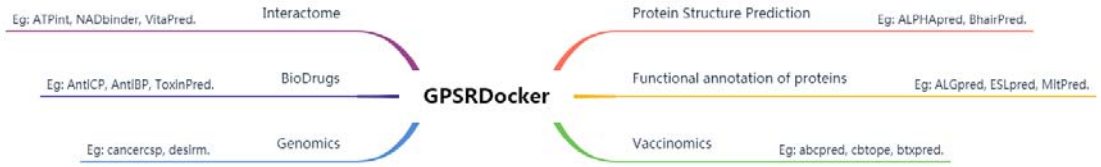


**Fig 1: This figure shows various types of standalone packages in GPSRDocker.**

### 2.3.1. Protein Structure Prediction:

Protein structure is traditionally determined using X-Ray crystallography, NMR spectroscopy and Cryo-electron microscopy. These methods are accurate but have certain limitations such as labor intensive, higher cost, etc. In order to bridge this gap, computational biologist have come up with *in silico* methods. *In silico* protein structure prediction is usually done by three methods (i) Homology Modelling; (ii) Threading based and (iii) ab initio method. In this section, we have described various methods developed by our group to predict 2D and 3D protein structure based on above principles. Described below are servers which can predict alpha turn, beta turns, gamma turns, phi-psi angle in protein, tertiary structure of proteins and peptides, etc.

**Table 2: List of Protein Structure Prediction software incorporated in the GPSRdocker.**

| Sr. | Server Name | Description | Category |
|---|---|---|---|

| No. | | | # |
|---|---|---|---|
| 1 | ALPHApred [15] | Neural network based method for predicting alpha-turn in a protein. | 1 |
| 2 | Ar_NHpred [16] | Identification of aromatic-backbone NH interaction in protein residues. | 1 |
| 3 | BetaTPred2 [17] | Statistical-based method for predicting Beta Turns in a protein. | 1 |
| 4 | BhairPred [18] | Prediction of beta hairpins in proteins using ANN and SVM techniques. | 1 |
| 5 | SARpred [19] | A neural network based method predicts the real value of surface accessibility. | 1 |
| 6 | TBBpred [20] | A webserver for the prediction of transmembrane Beta barrel regions in a given protein sequence. | 1 |

*# 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4: Genomics; 5: BioDrugs; 6: Interactome.*

### 2.3.2. Functional annotation of proteins:

Proteins are key components in various biological processes. Protein interactions with other molecules in a biological system are responsible for signaling pathways, up regulation/down regulation of processes etc. Alteration or mutation in a protein sequence can lead to altered protein function, growth of various diseases, etc. Due to advancement in next generation sequencing techniques, large number of genome projects has been sequenced providing pool of protein sequences. However, functional annotation of these proteins is yet to be unfolded. Our group has developed number of tools to predict the function of proteins.

**Table 3: List of Protein Functional annotation software incorporated in the GPSRdocker.**

| Sr. No. | Server Name | Description | Category[#] |
|---|---|---|---|
| 1 | ALGpred [21] | Prediction of allergenic proteins and mapping of IgE epitopes in antigens. | 2; 3 |
| 2 | ChemoPred [22] | A server to predict chemokines and their receptor | 2 |
| 3 | ESLpred [23] | Subcellular localization of the eukaryotic proteins using | 2 |
| 4 | ESLpred2 [24] | Advanced method for subcellular localization of eukaryotic proteins. | 2 |
| 5 | GPCRpred [25] | Prediction of families and superfamilies' of G-protein coupled receptors (GPCR). | 2 |
| 6 | GPCRsclass [26] | This webserver predicts amine type of G-protein coupled receptors | 2 |
| 7 | GSTpred [27] | SVM-based method for predicting Glutathione S-transferase protein. | 2 |
| 8 | HSLpred [28] | Prediction of subcellular localization of human proteins with high accuracy | 2 |
| 9 | MitPred [29] | Prediction of mitochondrial proteins using SVM and hidden Markov model. | 2 |
| 10 | Nppred [30] | A webserver for the prediction of nuclear proteins. | 2 |

| | | | |
|---|---|---|---|
| 11 | Nrpred [31] | A SVM based method for the classification of nuclear receptors | 2 |
| 12 | PFMpred [32] | Predicting mitochondrial proteins of malaria parasite Plasmodium falciparum. | 2 |
| 13 | PSEApred [33] | Prediction of Plasmodium Secretory and Infected Erythrocyte Associated Proteins. | 2 |
| 14 | PSLpred [34] | Predict subcellular localization of prokaryotic proteins. | 2 |
| 15 | RNApred [35] | A webserver for the prediction of RNA binding proteins. | 2 |
| 16 | RSLpred [36] | A method for the subcellular localization prediction of rice proteins. | 2 |
| 17 | SRTpred [37] | A method for the classification of protein sequence as secretory or non-secretory protein. | 2 |
| 18 | Tbpred [38] | A webserver that predicts four subcellular localization of mycobacterial proteins. | 2 |
| 19 | tRNAmod [39] | Prediction of post transcriptional modifications in transfer-RNA (tRNA) sequence. | 2 |

*# 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4: Genomics; 5: BioDrugs; 6: Interactome.*

### 2.3.3. Vaccinomics

Vaccinomics combines immunogentics and immunogenomics with systems biology and immune profiling, to aid in understanding personalized or precision medicine. Vaccinomics is

to comprehend biological immune system response towards vaccine-induced immunity of an individual. This paves the way for scientific community to design effective vaccines against hypervariable or resistant pathogens. Below are the tools developed by our group, which helps in development, administration and monitorization of potential vaccines.

**Table 4: List of Vaccinomics software incorporated in the GPSRdocker.**

| Sr. No. | Server Name | Description | Category[#] |
|---|---|---|---|
| 1 | abcpred [40] | Mapping of B-cell epitope(s) in an antigen sequence, using artificial neural network. | 3 |
| 2 | bcepred [40] | Prediction of linear B-cell epitopes, using Physico-chemical properties. | 3 |
| 3 | btxpred [41] | Prediction of bacterial toxins. | 3 |
| 4 | cbtope [42] | Conformational B-cell Epitope prediction. | 3 |
| 5 | ifnepitope [43] | Prediction and designing interferon-gamma inducing epitopes. | 3 |
| 6 | igpred [44] | Prediction of antibody specific B-cell epitope. | 3 |
| 7 | il10pred [45] | Prediction of Interleukin-10 inducing peptides. | 3 |
| 8 | il4pred [46] | In silico platform for designing and discovering of Interleukin-4 inducing peptides. | 3 |
| 9 | lbtope [47] | Prediction of linear B-cell epitopes. | 3 |
| 10 | pcleavage [48] | Identification of protesosomal cleavage sites in a protein sequence. | 3 |
| 11 | propred [49] | Prediction of MHC Class-II binding regions in an antigen sequence. | 3 |

| | | | |
|---|---|---|---|
| 12 | propred1 [50] | Prediction of promiscuous MHC Class-I binders. | 3 |
| 13 | tappred [51] | Prediction of binding affinity of peptides toward the TAP transporter. | 3 |
| 14 | vaxinpad [52] | Designing of peptide based vaccine adjuvant. | 3 |
| 15 | cancer_pred [53] | Prediction of the cancer lectins. | 3 |
| 16 | rnapin [54] | Prediction of Protein Interacting Nucleotides (PINs) in RNA sequences. | 3; 6 |

*# 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4: Genomics; 5: BioDrugs; 6: Interactome.*

### 2.3.4. Genomics: Genome annotation and application

With the advent of genomics era and next generation sequencing technologies, bioinformaticians have developed various tools for sequencing, assembling, structural and functional annotation of genomes. The sequencing data is increasing exponentially, and is available in public domain for analysis. Therefore, there is a need to develop tools which can effectively search, analyze and infer genomic information. In this direction, our group has developed many tools listed below:

**Table 5: List of Genomics software incorporated in the GPSRdocker.**

| Sr. No. | Server Name | Description | Category# |
|---|---|---|---|
| 1 | cancercsp [55] | Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. | 4, 3 |

| | | | |
|---|---|---|---|
| 2 | cancerspp [56] | Prediction and analysis of primary and metastatic tumor of SKCM using signature genes expression data. | 4, 3 |
| 3 | cancertsp [57] | Gene expression-based biomarkers for discriminating early and late stage of Papillary Thyroid Carcinoma (PTC). | 3 |
| 4 | desirm [58] | Designing of highly efficient siRNA with minimum mutation approach. | 4, 3 |

*# 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4: Genomics; 5: BioDrugs; 6: Interactome.*

### 2.3.5. BioDrugs: Biomolecules based therapeutics

BioDrugs also known as “bioactive drugs” are released in gastrointestinal tract by living orally administered recombinant microorganisms. These microorganisms are responsible for bioconversion or biosynthesis in the digestive environment. Bioactive drugs in model organisms like bacteria, yeasts etc. is a tedious and cost intensive experimental design. Chemoinformatics is the study of chemicals using various chemical databases, quantitative-structure activity relationship (QSAR), prediction of chemical properties or spectral. It plays a significant role in efficient drug discovery and development process. It has become an integral part of research in various fields like biochemistry, molecular biology, chemical genomics, bioinformatics etc. To facilitate this various bioinformatics tools have been developed which integrate experimentally validated data and used them to design new drugs. Our group has developed number of software’s to screen potential biodrugs in silico.

216 **Table 6: List of Biodrugs software incorporated in the GPSRdocker.**

| Sr. No. | Server Name | Description | Category[#] |
|---|---|---|---|
| 1 | AntiCP [59] | Prediction and design of anticancer peptides. | 5 |
| 2 | AHTpin [60] | Designing and virtual screening of antihypertensive peptides. | 5 |
| 3 | AntiBP [61] | Mapping of antibacterial peptides in a protein sequence. | 5 |
| 4 | AntiBP2 [62] | Advanced server for predicting antibacterial peptides with high precision. | 5 |
| 5 | CellPPD [63] | Computer-aided Designing of efficient cell penetrating peptides. | 5 |
| 6 | TumorHPD [64] | Server dedicated for designing tumor homing peptides. | 5 |
| 7 | HemoPI [65] | Prediction and virtual screening of hemolytic peptides. | 5 |
| 8 | ToxinPred [66] | An in silico method, which is developed to predict and design toxic/non-toxic peptides. | 5, 2, 3 |
| 9 | VICMPpred [67] | Prediction of Virulence factors, Information molecule, Cellular process and Metabolism molecule in the Bacterial proteins. | 5, 2 |
| 10 | NeuroPIpred [68] | Predict peptide as a neuropeptide or non neuro-peptide. | 5 |
| 11 | AntiTbPred [69] | To predict peptides with bactericidal activity against Mycobacterium species. | 5 |

*# 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4: Genomics; 5: BioDrugs; 6: Interactome.*

### 2.3.6. Interactome: Biomolecular based therapeutics

Molecular interactions among biomolecules inside a cell are known as "Interactome". Interactomics is the study of molecular interactions among proteins or small molecules and their consequences in the cell. These Bio-molecules could be proteins, nucleic acids, carbohydrates and lipid molecules. Interactomics could aid in identifying disease development and alteration in molecular mechanism of disease state. Below are few tools developed in our group to study interaction among biomolecules.

**Table 7: List of Interactome software incorporated in the GPSRdocker.**

| Sr. No. | Server Name | Description | Category[#] |
|---|---|---|---|
| 1 | ATPint [70] | Identification of ATP binding sites in ATP-binding proteins. | 6 |
| 2 | GlycoEP [71] | Prediction of C-, N- and O-glycosylation site in eukaryotic proteins. | 6,2 |
| 3 | GlycoPP [72] | Prediction of potential N-and O-glycosites in prokaryotic proteins. | 6,2 |
| 4 | GTPbinder [73] | Identification of GTP binding residue in protein sequences. | 6,2 |
| 5 | NADbinder [74] | Prediction of NAD binding proteins and their interacting residues. | 6,2 |

| | | | |
|---|---|---|---|
| 6 | Pprint [75] | ANN based method for identification of RNA-interacting residues in a protein. | 6 |
| 7 | PreMieR [76] | Identification of mannose interacting residues (MIRs) in protein sequences. | 6 |
| 8 | VitaPred [77] | Identification of different class of vitamin interacting residues in a protein. | 6 |
| 9 | SAMbinder [78] | A webserver for predict SAM interacting residue in a given protein sequence. | 6 |

*# 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4: Genomics; 5: BioDrugs; 6: Interactome.*

## 3. Applications of GPSRDocker

In this paper, we have launched GPSRdocker, which brings together various standalone versions of webservers developed by our group in various fields of bioinformatics. In this package we have tried to bring various open source software's to serve scientific community in a user friendly manner. GPSRdocker provides number of applications in our scientific filed and these applications are discussed below in detail.

**(i) Development of Novel Therapeutic Pipelines:** This is one of the biggest advantage of this docker container which provides an option to the user for developing new therapeutic pipelines. In the past decades, number of viruses and bacterial strains have been evolved which required immediate treatment to prevent their outbreak. User can implement different software present in the package for designing novel vaccine and drugs. User can utilize the genome of the new strains to identify the various epitopes (B-cell, T-cell and A-cell) which

247 can be used as potential vaccine candidates. ZikaVR [79], EbolaVCR [80] and VacTarBac
248 [81] are few examples of such type of pipelines.
249
250 **(ii) Cancer Risk stage prediction:** The suite also comprises of packages like CancerCSP
251 [55] which provides user to identify the potential biomarkers using gene expression data and
252 predict the possible risk stage of a cancer patient. This will help user to start treatment faster.
253
254 **(iii) Annotating large amount of protein sequences:** The image provides number of
255 methods developed for predicting small molecule binding site in the protein sequence. The
256 ligand includes ATP, GTP, NAD, FAD and SAM. These are important ligands used
257 previously for designing drugs. User can annotate the protein function for its protein
258 sequences by predicting the interacting site of these small ligands. Also, user can predict
259 whether the protein sequence of their interest in nucleic acid binding (DNA & RNA),
260 whether they are allergen proteins, etc. Packages like ESLpred2 [24], PSLpred [34] allows
261 user to predict the subcellular localization of their proteins sequence.
262
263 **(iv) Designing novel therapeutic peptides:** Number of packages have been incorporated
264 into this docker image for designing various types of peptide therapeutics. User can design
265 novel antimicrobial peptides such as antibacterial peptides, antifungal peptides, cell
266 penetrating peptides, hemolytic peptides, antiangiogenic peptides, anticancer peptides,
267 toxicity predicting peptides, chemically modified cell penetrating peptides, chemically
268 modified antimicrobial peptide, etc.
269
270 **(v) Easy to use on large dataset:** One of the main problems with the web-based services is
271 that they are not able to process after a certain limit and size of the dataset. GPSRdocker

allows user to implement and use the same service on the large dataset without any issue of file size and space. User needs to check the space availability of its local machine and thereafter can use the package the way he wants too.

**(vi) Data security:** Data security is one of the majors concerned nowadays. User can use the standalone service for securing its data and without providing any personal details. Also user can store any amount of data in the image without any loss of information by following the standard protocols while working in the docker container. User can also keep the data safe by saving the data in another image.

**(vii) No internet requirement:** internet availability is prerequisite for accessing the web based service. However, in case of docker standalone package, there is no requirement of the internet once the image is pulled on the local machine. User can work anywhere in the container without the internet presence.

**(viii) Comprehensive resource of software:** GPSRdocker provides a comprehensive resource of software related to different field of science such as immunoinformatic, protein structure and function annotation, cheminformatics, bio drugs, vaccine designing, genomics, etc. To the best of author's knowledge there is no such platform developed previously which comprises of more than 60 standalone version of the software. Therefore, this package will be very useful to the researchers both working in the wet lab as well as in dry lab. The package could be useful to various pharmaceutical companies as well as to the students who are starting their career in the area of bioinformatics.

## 4. Conclusion

GPSRdocker is a user friendly docker based container developed by our group which can be used to run standalone versions of various web servers. At present, GPSRdocker contains around 65 standalone software of the webservers developed by our group which are highly cited in the literature. Each server included in this container is used to address various questions in the field of computational biology. Aim of developing GPSRdocker is to integrate various freely available resources on a platform which is compatible with all type of operating systems. With the rapid advancement in the field of bioinformatics, there is a need to implement cloud based technologies such as Docker to make resources easily accessible to the users. The only limitation of this work is that it includes software developed specifically in our group only. However, there are various other useful bioinformatics containers available in market. We are working to include all the possible general bioinformatics modules as well as other new bioinformatics webservers in GPSRdocker version2.

**Acknowledgements**

Authors are thankful to funding agencies Department of Biotechnology (DBT) and Department of Science and Technology (DST-INSPIRE) and Council of Scientific and Industrial Research (CSIR), Govt. of India and Indraprastha Institute for Information Technology for financial support and fellowships.

**Author contribution**

PA, SP, AD, NS, AP, RK, VK and DK developed the python codes. PA, SSU, RK, AD, SP, NS, HK, VK, DK, SJ, and AP developed the standalone versions of the software. PA, SSU, AD, SP, NS, DK, SJ and VK developed the tables and figure. RK and VK developed the website and manual was written by PA, SSU and GPSR. PA, SJ, HK, and GPSR wrote the manuscript. GPSR conceived the idea and coordinated the project. All authors read and

approved the final paper.

**Funding**

This work was supported by J. C. Bose Fellowship, Department of Science and Technology, India.

**References**:

1. Stoehr PJ, Omond RA. The EMBL Network File Server. Nucleic Acids Res. 1989;17: 6763–4. Available: http://www.ncbi.nlm.nih.gov/pubmed/2780308
2. IUBio-Archive | re3data.org [Internet]. [cited 2 May 2019]. Available: https://www.re3data.org/repository/r3d100010090
3. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, et al. BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Res. 2010;38: W689-94. doi:10.1093/nar/gkq394
4. Raghava GPS. PDSB: Public Domain Software in Biology. Biotech Softw Internet Rep. 2001;2: 154–156. doi:10.1089/152791601753204313
5. Raghava GP, Joshi AK, Agrewala JN. Calculation of antibody and antigen concentrations from ELISA data using a graphical method. J Immunol Methods. 1992;153: 263–4. Available: http://www.ncbi.nlm.nih.gov/pubmed/1517598
6. Inman D, Albrecht B. The GW-BASIC reference [Internet]. Osborne McGraw-Hill; 1990. Available: https://dl.acm.org/citation.cfm?id=103255
7. PSIBLAST [Internet]. [cited 2 May 2019]. Available: http://www.biology.wustl.edu/gcg/psiblast.html
8. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22: 1658–9. doi:10.1093/bioinformatics/btl158
9. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics. 1999;15: 327–32. Available: http://www.ncbi.nlm.nih.gov/pubmed/10320401
10. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000;16: 404–5. Available: http://www.ncbi.nlm.nih.gov/pubmed/10869041
11. About us — scikit-learn 0.20.3 documentation [Internet]. [cited 2 May 2019]. Available: https://scikit-learn.org/stable/about.html
12. Joachims T. Learning to classify text using support vector machines [Internet]. Kluwer Academic Publishers; 2002. Available: http://www.cs.cornell.edu/people/tj/svmtcatbook/
13. Zell A, Mache N, Hübner R, Mamier G, Vogt M, Schmalzl M, et al. SNNS (Stuttgart Neural Network Simulator). Springer, Boston, MA; 1994. pp. 165–186. doi:10.1007/978-1-4615-2736-7_9
14. Smith TC, Frank E. Introducing Machine Learning Concepts with WEKA. Methods Mol Biol. 2016;1418: 353–78. doi:10.1007/978-1-4939-3578-9_17
15. Kaur H, Raghava GPS. Prediction of alpha-turns in proteins using PSI-BLAST profiles

and secondary structure information. Proteins. 2004;55: 83–90. doi:10.1002/prot.10569

16. Kaur H, Raghava GPS. Role of evolutionary information in prediction of aromatic-backbone NH interactions in proteins. FEBS Lett. 2004;564: 47–57. doi:10.1016/S0014-5793(04)00305-9

17. Kaur H, Raghava GPS. Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci. 2003;12: 627–34. doi:10.1110/ps.0228903

18. Kumar M, Bhasin M, Natt NK, Raghava GPS. BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res. 2005;33: W154-9. doi:10.1093/nar/gki588

19. Garg A, Kaur H, Raghava GPS. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins. 2005;61: 318–24. doi:10.1002/prot.20630

20. Natt NK, Kaur H, Raghava GPS. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. Proteins. 2004;56: 11–8. doi:10.1002/prot.20092

21. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res. 2006;34: W202-9. doi:10.1093/nar/gkl343

22. Lata S, Raghava GPS. Prediction and classification of chemokines and their receptors. Protein Eng Des Sel. 2009;22: 441–4. doi:10.1093/protein/gzp016

23. Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. 2004;32: W414-9. doi:10.1093/nar/gkh350

24. Garg A, Raghava GPS. ESLpred2: improved method for predicting subcellular localization of eukaryotic proteins. BMC Bioinformatics. 2008;9: 503. doi:10.1186/1471-2105-9-503

25. Bhasin M, Raghava GPS. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res. 2004;32: W383-9. doi:10.1093/nar/gkh416

26. Bhasin M, Raghava GPS. GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. Nucleic Acids Res. 2005;33: W143-7. doi:10.1093/nar/gki351

27. Mishra NK, Kumar M, Raghava GPS. Support vector machine based prediction of glutathione S-transferase proteins. Protein Pept Lett. 2007;14: 575–80. Available: http://www.ncbi.nlm.nih.gov/pubmed/17627599

28. Garg A, Bhasin M, Raghava GPS. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem. 2005;280: 14427–32. doi:10.1074/jbc.M411789200

29. Kumar M, Verma R, Raghava GPS. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J Biol Chem. 2006;281: 5357–63. doi:10.1074/jbc.M511061200

30. Kumar M, Raghava GPS. Prediction of nuclear proteins using SVM and HMM models. BMC Bioinformatics. 2009;10: 22. doi:10.1186/1471-2105-10-22

31. Bhasin M, Raghava GPS. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem. 2004;279: 23262–6. doi:10.1074/jbc.M401932200

32. Verma R, Varshney GC, Raghava GPS. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids. 2010;39: 101–10. doi:10.1007/s00726-009-0381-1

33. Verma R, Tiwari A, Kaur S, Varshney GC, Raghava GP. Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. BMC Bioinformatics. 2008;9: 201. doi:10.1186/1471-2105-9-201
34. Bhasin M, Garg A, Raghava GPS. PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics. 2005;21: 2522–4. doi:10.1093/bioinformatics/bti309
35. Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recognit. 2011;24: 303–13. doi:10.1002/jmr.1061
36. Kaundal R, Raghava GPS. RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. Proteomics. 2009;9: 2324–42. doi:10.1002/pmic.200700597
37. Garg A, Raghava GPS. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. In Silico Biol. 2008;8: 129–40. Available: http://www.ncbi.nlm.nih.gov/pubmed/18928201
38. Rashid M, Saha S, Raghava GP. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics. 2007;8: 337. doi:10.1186/1471-2105-8-337
39. Panwar B, Raghava GPS. Prediction of uridine modifications in tRNA sequences. BMC Bioinformatics. 2014;15: 326. doi:10.1186/1471-2105-15-326
40. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins. 2006;65: 40–8. doi:10.1002/prot.21078
41. Saha S, Raghava GPS. BTXpred: prediction of bacterial toxins. In Silico Biol. 2007;7: 405–12. Available: http://www.ncbi.nlm.nih.gov/pubmed/18391233
42. Ansari HR, Raghava GP. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. Immunome Res. 2010;6: 6. doi:10.1186/1745-7580-6-6
43. Dhanda SK, Vir P, Raghava GPS. Designing of interferon-gamma inducing MHC class-II binders. Biol Direct. 2013;8: 30. doi:10.1186/1745-6150-8-30
44. Gupta S, Ansari HR, Gautam A, Open Source Drug Discovery Consortium GP, Raghava GPS. Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. Biol Direct. 2013;8: 27. doi:10.1186/1745-6150-8-27
45. Nagpal G, Usmani SS, Dhanda SK, Kaur H, Singh S, Sharma M, et al. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. Sci Rep. 2017;7: 42851. doi:10.1038/srep42851
46. Dhanda SK, Gupta S, Vir P, Raghava GPS. Prediction of IL4 inducing peptides. Clin Dev Immunol. 2013;2013: 263952. doi:10.1155/2013/263952
47. Singh H, Ansari HR, Raghava GPS. Improved method for linear B-cell epitope prediction using antigen's primary sequence. Schönbach C, editor. PLoS One. 2013;8: e62216. doi:10.1371/journal.pone.0062216
48. Bhasin M, Raghava GPS. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. Nucleic Acids Res. 2005;33: W202-7. doi:10.1093/nar/gki587
49. Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. Bioinformatics. 2001;17: 1236–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/11751237
50. Singh H, Raghava GPS. ProPred1: prediction of promiscuous MHC Class-I binding sites. Bioinformatics. 2003;19: 1009–14. Available: http://www.ncbi.nlm.nih.gov/pubmed/12761064
51. Bhasin M, Raghava GPS. Analysis and prediction of affinity of TAP binding peptides

using cascade SVM. Protein Sci. 2004;13: 596–607. doi:10.1110/ps.03373104

52. Nagpal G, Chaudhary K, Agrawal P, Raghava GPS. Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants. J Transl Med. 2018;16: 181. doi:10.1186/s12967-018-1560-1

53. Kumar R, Panwar B, Chauhan JS, Raghava GP. Analysis and prediction of cancerlectins using evolutionary and domain information. BMC Res Notes. 2011;4: 237. doi:10.1186/1756-0500-4-237

54. Panwar B, Raghava GPS. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. Genomics. 2015;105: 197–203. doi:10.1016/j.ygeno.2015.01.005

55. Bhalla S, Chaudhary K, Kumar R, Sehgal M, Kaur H, Sharma S, et al. Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. Sci Rep. 2017;7: 44997. doi:10.1038/srep44997

56. Bhalla S, Kaur H, Dhall A, Raghava GPS. Prediction and analysis of skin cancer progression using genomics profiles of patients. bioRxiv. Cold Spring Harbor Laboratory; 2018; 393454. doi:10.1101/393454

57. Bhalla S, Kaur H, Kaur R, Sharma S, Raghava GPS. Expression based biomarkers and models to classify early and late stage samples of Papillary Thyroid Carcinoma. bioRxiv. Cold Spring Harbor Laboratory; 2018; 393975. doi:10.1101/393975

58. Ahmed F, Raghava GPS. Designing of highly effective complementary and mismatch siRNAs for silencing a gene. Preiss T, editor. PLoS One. 2011;6: e23443. doi:10.1371/journal.pone.0023443

59. Kumar S, Li H. In Silico Design of Anticancer Peptides. Methods Mol Biol. 2017;1647: 245–254. doi:10.1007/978-1-4939-7201-2_17

60. Kumar R, Chaudhary K, Singh Chauhan J, Nagpal G, Kumar R, Sharma M, et al. An in silico platform for predicting, screening and designing of antihypertensive peptides. Sci Rep. 2015;5: 12512. doi:10.1038/srep12512

61. Lata S, Sharma BK, Raghava GPS. Analysis and prediction of antibacterial peptides. BMC Bioinformatics. 2007;8: 263. doi:10.1186/1471-2105-8-263

62. Lata S, Mishra NK, Raghava GPS. AntiBP2: improved version of antibacterial peptide prediction. BMC Bioinformatics. 2010;11 Suppl 1: S19. doi:10.1186/1471-2105-11-S1-S19

63. Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, et al. In silico approaches for designing highly effective cell penetrating peptides. J Transl Med. 2013;11: 74. doi:10.1186/1479-5876-11-74

64. Sharma A, Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, et al. Computational approach for designing tumor homing peptides. Sci Rep. 2013;3: 1607. doi:10.1038/srep01607

65. Chaudhary K, Kumar R, Singh S, Tuknait A, Gautam A, Mathur D, et al. A Web Server and Mobile App for Computing Hemolytic Potency of Peptides. Sci Rep. 2016;6: 22843. doi:10.1038/srep22843

66. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Open Source Drug Discovery Consortium GPS, et al. In silico approach for predicting toxicity of peptides and proteins. Patterson RL, editor. PLoS One. 2013;8: e73957. doi:10.1371/journal.pone.0073957

67. Saha S, Raghava GPS. VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. Genomics Proteomics Bioinformatics. 2006;4: 42–7. doi:10.1016/S1672-0229(06)60015-6

68. Agrawal P, Kumar S, Singh A, Raghava GPS, Singh IK. NeuroPIpred: a tool to

predict, design and scan insect neuropeptides. Sci Rep. 2019;9: 5129. doi:10.1038/s41598-019-41538-x

69. Usmani SS, Bhalla S, Raghava GPS. Prediction of Antitubercular Peptides From Sequence Information Using Ensemble Classifier and Hybrid Features. Front Pharmacol. 2018;9: 954. doi:10.3389/fphar.2018.00954

70. Chauhan JS, Mishra NK, Raghava GPS. Identification of ATP binding residues of a protein from its primary sequence. BMC Bioinformatics. 2009;10: 434. doi:10.1186/1471-2105-10-434

71. Chauhan JS, Rao A, Raghava GPS. In silico platform for prediction of N-, O- and C-glycosites in eukaryotic protein sequences. E Tosatto SC, editor. PLoS One. 2013;8: e67008. doi:10.1371/journal.pone.0067008

72. Chauhan JS, Bhat AH, Raghava GPS, Rao A. GlycoPP: a webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. Burchell JM, editor. PLoS One. 2012;7: e40155. doi:10.1371/journal.pone.0040155

73. Chauhan JS, Mishra NK, Raghava GPS. Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. BMC Bioinformatics. 2010;11: 301. doi:10.1186/1471-2105-11-301

74. Ansari HR, Raghava GPS. Identification of NAD interacting residues in proteins. BMC Bioinformatics. 2010;11: 160. doi:10.1186/1471-2105-11-160

75. Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins. 2008;71: 189–94. doi:10.1002/prot.21677

76. Agarwal S, Mishra NK, Singh H, Raghava GPS. Identification of mannose interacting residues using local composition. Tramontano A, editor. PLoS One. 2011;6: e24039. doi:10.1371/journal.pone.0024039

77. Panwar B, Gupta S, Raghava GPS. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. BMC Bioinformatics. 2013;14: 44. doi:10.1186/1471-2105-14-44

78. Agrawal P, Mishra G, Raghava GPS. SAMbinder: A web server for predicting SAM binding residues of a protein from its amino acid sequence. bioRxiv. Cold Spring Harbor Laboratory; 2019; 625806. doi:10.1101/625806

79. Gupta AK, Kaur K, Rajput A, Dhanda SK, Sehgal M, Khan MS, et al. ZikaVR: An Integrated Zika Virus Resource for Genomics, Proteomics, Phylogenetic and Therapeutic Analysis. Sci Rep. 2016;6: 32713. doi:10.1038/srep32713

80. Dhanda SK, Chaudhary K, Gupta S, Brahmachari SK, Raghava GPS. A web-based resource for designing therapeutics against Ebola Virus. Sci Rep. 2016;6: 24782. doi:10.1038/srep24782

81. Nagpal G, Usmani SS, Raghava GPS. A Web Resource for Designing Subunit Vaccine Against Major Pathogenic Species of Bacteria. Front Immunol. 2018;9: 2280. doi:10.3389/fimmu.2018.02280