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

2

3 **Piyush Agrawal^{1,2#}, Rajesh Kumar^{1,2#}, Salman Sadullah Usmani^{1,2#}, Anjali Dhall^{1#},**

4 **Sumeet Patiyal¹, Neelam Sharma¹, Harpreet Kaur^{1,2}, Vinod Kumar^{1,2}, Dilraj Kaur¹,**

5 **Shipra Jain¹, Akshara Pande¹, Sherry Bhalla¹, Gajendra P.S. Raghava^{1,2,*}**

6

7 ¹ Department of Computational Biology, Indraprastha Institute of Information Technology,
8 Delhi, India.

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

10

11

12 [#]Equal contribution

13

14 * Corresponding authors

15

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

18 **Abstract**

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

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

40 **Conclusion:** A docker-based container has been developed which can be easily run on any
 41 operating system as well as it can be directly ported on cloud. Scripts can be run to build
 42 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 webserver 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 webserver 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.

85

86

87

88 **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 webserver 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.iitd.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

knowledge of programming skills. Following are important programs included in this package used to generate features for developing prediction models from protein and DNA/RNA sequences. These programs are developed in PERL and Python following the set standards. In order to run these codes, user needs to have Python 3.0 or above version installed in their system.

Table 1: List of the scripts along with their purpose incorporated in the GPSRdocker.

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

	profile based on Eqn: $1/(1+e^{-x})$		on Eqn: $(\text{numb} - \text{min})/(\text{max} - \text{min})$
<i>pssm_n3</i>	Normalize pssm profile based on Eqn: $(\text{numb} - \text{min}) * 100 / (\text{max} - \text{min})$	<i>pssm_n4</i>	Normalize pssm profile based on Eqn: $1/(1+e^{-(x/100)})$
<i>pssm_comp</i>	Compute PSSM composition (400)	<i>pssm_smooth</i>	Design smooth pssm profile for plot
<i>pssm2pat</i>	To generate patterns of given size from PSSM matrix	<i>seq2motif</i>	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)

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

142
143 **Fig 1: This figure shows various types of standalone packages in GPSRDocker.**

144
145 **2.3.1. Protein Structure Prediction:**

146
147 Protein structure is traditionally determined using X-Ray crystallography, NMR spectroscopy
148 and Cryo-electron microscopy. These methods are accurate but have certain limitations such
149 as labor intensive, higher cost, etc. In order to bridge this gap, computational biologist have
150 come up with *in silico* methods. *In silico* protein structure prediction is usually done by three
151 methods (i) Homology Modelling; (ii) Threading based and (iii) ab initio method. In this
152 section, we have described various methods developed by our group to predict 2D and 3D
153 protein structure based on above principles. Described below are servers which can predict
154 alpha turn, beta turns, gamma turns, phi-psi angle in protein, tertiary structure of proteins and
155 peptides, etc.

156
157 **Table 2: List of Protein Structure Prediction software incorporated in the GPSRdocker.**

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

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

160

2.3.2. Functional annotation of proteins:

162

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.

170

171 **Table 3: List of Protein Functional annotation software incorporated in the**
 172 **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 immunogenetics and immunogenomics with systems biology and immune profiling, to aid in understanding personalized or precision medicine. Vaccinomics is

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

184

185 **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 proteasomal 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

186 # 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4:
187 Genomics; 5: BioDrugs; 6: Interactome.

188

189 2.3.4. Genomics: Genome annotation and application

190

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

197

198 **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 neuropeptide.	5
11	AntiTbPred [69]	To predict peptides with bactericidal activity against Mycobacterium species.	5

217 # 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4:
218 Genomics; 5: BioDrugs; 6: Interactome.

219

220 2.3.6. Interactome: Biomolecular based therapeutics

221

222 Molecular interactions among biomolecules inside a cell are known as “Interactome”.

223 Interactomics is the study of molecular interactions among proteins or small molecules and

224 their consequences in the cell. These Bio-molecules could be proteins, nucleic acids,

225 carbohydrates and lipid molecules. Interactomics could aid in identifying disease

226 development and alteration in molecular mechanism of disease state. Below are few tools

227 developed in our group to study interaction among biomolecules.

228

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

230 # 1: Protein Structure Prediction; 2: Functional annotation of proteins; 3: Vaccinomics; 4:

231 Genomics; 5: BioDrugs; 6: Interactome.

232

233 3. Applications of GPSRDocker

234

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

240

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

can be used as potential vaccine candidates. ZikaVR [79], EbolaVCR [80] and VacTarBac [81] are few examples of such type of pipelines.

(ii) Cancer Risk stage prediction: The suite also comprises of packages like CancerCSP [55] which provides user to identify the potential biomarkers using gene expression data and predict the possible risk stage of a cancer patient. This will help user to start treatment faster.

(iii) Annotating large amount of protein sequences: The image provides number of methods developed for predicting small molecule binding site in the protein sequence. The ligand includes ATP, GTP, NAD, FAD and SAM. These are important ligands used previously for designing drugs. User can annotate the protein function for its protein sequences by predicting the interacting site of these small ligands. Also, user can predict whether the protein sequence of their interest in nucleic acid binding (DNA & RNA), whether they are allergen proteins, etc. Packages like ESLpred2 [24], PSLpred [34] allows user to predict the subcellular localization of their proteins sequence.

(iv) Designing novel therapeutic peptides: Number of packages have been incorporated into this docker image for designing various types of peptide therapeutics. User can design novel antimicrobial peptides such as antibacterial peptides, antifungal peptides, cell penetrating peptides, hemolytic peptides, antiangiogenic peptides, anticancer peptides, toxicity predicting peptides, chemically modified cell penetrating peptides, chemically modified antimicrobial peptide, etc.

(v) Easy to use on large dataset: One of the main problems with the web-based services is 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 web servers 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 web servers 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. Stoeckl 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/svmcatbook/>
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

- 366 and secondary structure information. *Proteins*. 2004;55: 83–90.
367 doi:10.1002/prot.10569
- 368 16. Kaur H, Raghava GPS. Role of evolutionary information in prediction of aromatic-
369 backbone NH interactions in proteins. *FEBS Lett*. 2004;564: 47–57.
370 doi:10.1016/S0014-5793(04)00305-9
- 371 17. Kaur H, Raghava GPS. Prediction of beta-turns in proteins from multiple alignment
372 using neural network. *Protein Sci*. 2003;12: 627–34. doi:10.1110/ps.0228903
- 373 18. Kumar M, Bhasin M, Natt NK, Raghava GPS. BhairPred: prediction of beta-hairpins
374 in a protein from multiple alignment information using ANN and SVM techniques.
375 *Nucleic Acids Res*. 2005;33: W154-9. doi:10.1093/nar/gki588
- 376 19. Garg A, Kaur H, Raghava GPS. Real value prediction of solvent accessibility in
377 proteins using multiple sequence alignment and secondary structure. *Proteins*.
378 2005;61: 318–24. doi:10.1002/prot.20630
- 379 20. Natt NK, Kaur H, Raghava GPS. Prediction of transmembrane regions of beta-barrel
380 proteins using ANN- and SVM-based methods. *Proteins*. 2004;56: 11–8.
381 doi:10.1002/prot.20092
- 382 21. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE
383 epitopes. *Nucleic Acids Res*. 2006;34: W202-9. doi:10.1093/nar/gkl343
- 384 22. Lata S, Raghava GPS. Prediction and classification of chemokines and their receptors.
385 *Protein Eng Des Sel*. 2009;22: 441–4. doi:10.1093/protein/gzp016
- 386 23. Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of
387 eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*.
388 2004;32: W414-9. doi:10.1093/nar/gkh350
- 389 24. Garg A, Raghava GPS. ESLpred2: improved method for predicting subcellular
390 localization of eukaryotic proteins. *BMC Bioinformatics*. 2008;9: 503.
391 doi:10.1186/1471-2105-9-503
- 392 25. Bhasin M, Raghava GPS. GPCRpred: an SVM-based method for prediction of families
393 and subfamilies of G-protein coupled receptors. *Nucleic Acids Res*. 2004;32: W383-9.
394 doi:10.1093/nar/gkh416
- 395 26. Bhasin M, Raghava GPS. GPCRsclass: a web tool for the classification of amine type
396 of G-protein-coupled receptors. *Nucleic Acids Res*. 2005;33: W143-7.
397 doi:10.1093/nar/gki351
- 398 27. Mishra NK, Kumar M, Raghava GPS. Support vector machine based prediction of
399 glutathione S-transferase proteins. *Protein Pept Lett*. 2007;14: 575–80. Available:
400 <http://www.ncbi.nlm.nih.gov/pubmed/17627599>
- 401 28. Garg A, Bhasin M, Raghava GPS. Support vector machine-based method for
402 subcellular localization of human proteins using amino acid compositions, their order,
403 and similarity search. *J Biol Chem*. 2005;280: 14427–32.
404 doi:10.1074/jbc.M411789200
- 405 29. Kumar M, Verma R, Raghava GPS. Prediction of mitochondrial proteins using support
406 vector machine and hidden Markov model. *J Biol Chem*. 2006;281: 5357–63.
407 doi:10.1074/jbc.M511061200
- 408 30. Kumar M, Raghava GPS. Prediction of nuclear proteins using SVM and HMM
409 models. *BMC Bioinformatics*. 2009;10: 22. doi:10.1186/1471-2105-10-22
- 410 31. Bhasin M, Raghava GPS. Classification of nuclear receptors based on amino acid
411 composition and dipeptide composition. *J Biol Chem*. 2004;279: 23262–6.
412 doi:10.1074/jbc.M401932200
- 413 32. Verma R, Varshney GC, Raghava GPS. Prediction of mitochondrial proteins of
414 malaria parasite using split amino acid composition and PSSM profile. *Amino Acids*.
415 2010;39: 101–10. doi:10.1007/s00726-009-0381-1

- 416 33. Verma R, Tiwari A, Kaur S, Varshney GC, Raghava GP. Identification of proteins
417 secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. BMC
418 Bioinformatics. 2008;9: 201. doi:10.1186/1471-2105-9-201
- 419 34. Bhasin M, Garg A, Raghava GPS. PSLpred: prediction of subcellular localization of
420 bacterial proteins. Bioinformatics. 2005;21: 2522–4.
421 doi:10.1093/bioinformatics/bti309
- 422 35. Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding
423 proteins using binding residues and evolutionary information. J Mol Recognit.
424 2011;24: 303–13. doi:10.1002/jmr.1061
- 425 36. Kaundal R, Raghava GPS. RSLpred: an integrative system for predicting subcellular
426 localization of rice proteins combining compositional and evolutionary information.
427 Proteomics. 2009;9: 2324–42. doi:10.1002/pmic.200700597
- 428 37. Garg A, Raghava GPS. A machine learning based method for the prediction of
429 secretory proteins using amino acid composition, their order and similarity-search. In
430 Silico Biol. 2008;8: 129–40. Available:
431 <http://www.ncbi.nlm.nih.gov/pubmed/18928201>
- 432 38. Rashid M, Saha S, Raghava GP. Support Vector Machine-based method for predicting
433 subcellular localization of mycobacterial proteins using evolutionary information and
434 motifs. BMC Bioinformatics. 2007;8: 337. doi:10.1186/1471-2105-8-337
- 435 39. Panwar B, Raghava GPS. Prediction of uridine modifications in tRNA sequences.
436 BMC Bioinformatics. 2014;15: 326. doi:10.1186/1471-2105-15-326
- 437 40. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using
438 recurrent neural network. Proteins. 2006;65: 40–8. doi:10.1002/prot.21078
- 439 41. Saha S, Raghava GPS. BTXpred: prediction of bacterial toxins. In Silico Biol. 2007;7:
440 405–12. Available: <http://www.ncbi.nlm.nih.gov/pubmed/18391233>
- 441 42. Ansari HR, Raghava GP. Identification of conformational B-cell Epitopes in an
442 antigen from its primary sequence. Immunome Res. 2010;6: 6. doi:10.1186/1745-
443 7580-6-6
- 444 43. Dhanda SK, Vir P, Raghava GPS. Designing of interferon-gamma inducing MHC
445 class-II binders. Biol Direct. 2013;8: 30. doi:10.1186/1745-6150-8-30
- 446 44. Gupta S, Ansari HR, Gautam A, Open Source Drug Discovery Consortium GP,
447 Raghava GPS. Identification of B-cell epitopes in an antigen for inducing specific
448 class of antibodies. Biol Direct. 2013;8: 27. doi:10.1186/1745-6150-8-27
- 449 45. Nagpal G, Usmani SS, Dhanda SK, Kaur H, Singh S, Sharma M, et al. Computer-
450 aided designing of immunosuppressive peptides based on IL-10 inducing potential. Sci
451 Rep. 2017;7: 42851. doi:10.1038/srep42851
- 452 46. Dhanda SK, Gupta S, Vir P, Raghava GPS. Prediction of IL4 inducing peptides. Clin
453 Dev Immunol. 2013;2013: 263952. doi:10.1155/2013/263952
- 454 47. Singh H, Ansari HR, Raghava GPS. Improved method for linear B-cell epitope
455 prediction using antigen's primary sequence. Schönbach C, editor. PLoS One. 2013;8:
456 e62216. doi:10.1371/journal.pone.0062216
- 457 48. Bhasin M, Raghava GPS. Pcleavage: an SVM based method for prediction of
458 constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences.
459 Nucleic Acids Res. 2005;33: W202–7. doi:10.1093/nar/gki587
- 460 49. Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. Bioinformatics.
461 2001;17: 1236–7. Available: <http://www.ncbi.nlm.nih.gov/pubmed/11751237>
- 462 50. Singh H, Raghava GPS. ProPred1: prediction of promiscuous MHC Class-I binding
463 sites. Bioinformatics. 2003;19: 1009–14. Available:
464 <http://www.ncbi.nlm.nih.gov/pubmed/12761064>
- 465 51. Bhasin M, Raghava GPS. Analysis and prediction of affinity of TAP binding peptides

- 466 using cascade SVM. *Protein Sci.* 2004;13: 596–607. doi:10.1110/ps.03373104
- 467 52. Nagpal G, Chaudhary K, Agrawal P, Raghava GPS. Computer-aided prediction of
468 antigen presenting cell modulators for designing peptide-based vaccine adjuvants. *J*
469 *Transl Med.* 2018;16: 181. doi:10.1186/s12967-018-1560-1
- 470 53. Kumar R, Panwar B, Chauhan JS, Raghava GP. Analysis and prediction of
471 cancerlectins using evolutionary and domain information. *BMC Res Notes.* 2011;4:
472 237. doi:10.1186/1756-0500-4-237
- 473 54. Panwar B, Raghava GPS. Identification of protein-interacting nucleotides in a RNA
474 sequence using composition profile of tri-nucleotides. *Genomics.* 2015;105: 197–203.
475 doi:10.1016/j.ygeno.2015.01.005
- 476 55. Bhalla S, Chaudhary K, Kumar R, Sehgal M, Kaur H, Sharma S, et al. Gene
477 expression-based biomarkers for discriminating early and late stage of clear cell renal
478 cancer. *Sci Rep.* 2017;7: 44997. doi:10.1038/srep44997
- 479 56. Bhalla S, Kaur H, Dhali A, Raghava GPS. Prediction and analysis of skin cancer
480 progression using genomics profiles of patients. *bioRxiv.* Cold Spring Harbor
481 Laboratory; 2018; 393454. doi:10.1101/393454
- 482 57. Bhalla S, Kaur H, Kaur R, Sharma S, Raghava GPS. Expression based biomarkers and
483 models to classify early and late stage samples of Papillary Thyroid Carcinoma.
484 *bioRxiv.* Cold Spring Harbor Laboratory; 2018; 393975. doi:10.1101/393975
- 485 58. Ahmed F, Raghava GPS. Designing of highly effective complementary and mismatch
486 siRNAs for silencing a gene. Preiss T, editor. *PLoS One.* 2011;6: e23443.
487 doi:10.1371/journal.pone.0023443
- 488 59. Kumar S, Li H. In Silico Design of Anticancer Peptides. *Methods Mol Biol.*
489 2017;1647: 245–254. doi:10.1007/978-1-4939-7201-2_17
- 490 60. Kumar R, Chaudhary K, Singh Chauhan J, Nagpal G, Kumar R, Sharma M, et al. An
491 in silico platform for predicting, screening and designing of antihypertensive peptides.
492 *Sci Rep.* 2015;5: 12512. doi:10.1038/srep12512
- 493 61. Lata S, Sharma BK, Raghava GPS. Analysis and prediction of antibacterial peptides.
494 *BMC Bioinformatics.* 2007;8: 263. doi:10.1186/1471-2105-8-263
- 495 62. Lata S, Mishra NK, Raghava GPS. AntiBP2: improved version of antibacterial peptide
496 prediction. *BMC Bioinformatics.* 2010;11 Suppl 1: S19. doi:10.1186/1471-2105-11-
497 S1-S19
- 498 63. Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, et al. In silico
499 approaches for designing highly effective cell penetrating peptides. *J Transl Med.*
500 2013;11: 74. doi:10.1186/1479-5876-11-74
- 501 64. Sharma A, Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, et al.
502 Computational approach for designing tumor homing peptides. *Sci Rep.* 2013;3: 1607.
503 doi:10.1038/srep01607
- 504 65. Chaudhary K, Kumar R, Singh S, Tuknait A, Gautam A, Mathur D, et al. A Web
505 Server and Mobile App for Computing Hemolytic Potency of Peptides. *Sci Rep.*
506 2016;6: 22843. doi:10.1038/srep22843
- 507 66. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Open Source Drug Discovery
508 Consortium GPS, et al. In silico approach for predicting toxicity of peptides and
509 proteins. Patterson RL, editor. *PLoS One.* 2013;8: e73957.
510 doi:10.1371/journal.pone.0073957
- 511 67. Saha S, Raghava GPS. VICMpred: an SVM-based method for the prediction of
512 functional proteins of Gram-negative bacteria using amino acid patterns and
513 composition. *Genomics Proteomics Bioinformatics.* 2006;4: 42–7. doi:10.1016/S1672-
514 0229(06)60015-6
- 515 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