

Genomic and proteomic repository of chitin degrading bacterium *Serratia proteamaculans* 568

P.V. Parvati Sai Arun^{1*}

CR Rao Advanced Institute of Mathematics Statistics and Computer Science, Hyderabad, India

^{*}Corresponding Author: arun.uoh@gmail.com, Tel.: +91-70751-11851

Available online at: www.ijcseonline.org

Received: 02/Aug/2017, Revised: 18/Aug/2017, Accepted: 10/Sep/2017, Published: 30/Sep/2017

Abstract— In this paper, we describe about a repository which is composed of the information related to genes sequences, proteins sequences, upstream sequences, codon usage in proteins, physico-chemical properties, secondary structures and biochemical pathway information of proteins of chitin degrading bacterium *Serratia proteamaculans* 568. The advantage of this repository is that it can be hosted in the user's computer and work without internet connection. The backend data for developing this repository was generated using different computational tools which were published earlier. The .faa, .fna, .ptt files of *S.proteamaculans* 568 were downloaded from NCBI was used as primary seed data for the generation backend data. Web technologies were used to retrieve and display the compiled data in the browser. The data retrieved out of this repository can be used as preliminary source for understanding various concepts related to genes and proteins of *Serratia proteamaculans* 568. This repository can be obtained from http://crraoaims.res.in/serratia_568/serratia_568.rar

Keywords— *Serratia proteamaculans* 568, gene sequences, protein sequences, physico-chemical properties, Secondary structures

I. INTRODUCTION

Serratia proteamaculans 568 is an endophyte which is isolated from *Populus trichocarpa* [1]. *Serratia proteamaculans* 568 was reported as plant growth promoting bacteria which induces the development of root and shoots of plants and help in plant growth. *S.proteamaculans* 568 became good model system for the researchers working on plant health, growth and also in the area of glyco-biology [2]. The chromosome of *S.proteamaculans* 568 codes for a total of 4891 proteins. Many researchers working on genomics and proteomics of *S.proteamaculans* 568 get data from different resources and analyze it. For biologists, it is very time consuming and laborious process to collect the data from various sources. To aid the biologists, who work on the *S.proteamaculans* 568, the data from different resources was collected and compiled in the form of a repository. Motivated from the databases developed on cyanobacteria such as Cyanobase [3] and CyanoPhyChe [4], this repository was developed on plant growth promoting bacteria *Serratia proteamaculans* 568. This resource provides the information such as gene sequences, upstream sequences, and directionality of the gene on the chromosome, protein sequence, its function along with the preliminary information about the physico-chemical properties and biochemical pathway information of proteins. Additional advantage of this repository is that it can be installed locally on the user's

computer using XAMPP which needs no internet. This repository is first of its kind developed for *S. proteamaculans* 568 serves as good source of genomic and proteomic resource.

II. RELATED WORK

Previously, similar kind of database was developed to aid the researchers who work in the area of cyanobacteria. The database such as Cyanobase serve mainly as the genomic repository whereas CyanoPhyChe serve as proteomic resource [3][4].

III. METHODOLOGY

The .faa, .fna and .ptt files of *S.proteamaculans* 568 chromosome was collected from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/). Initially the .faa file, which contains all the protein sequences, of both the chromosome was taken and split into individual protein files in FASTA format (Primary seed data). This primary seed data was taken and given as input for PEPSTATS from emboss package for the prediction of physico-chemical properties of the proteins [5]. The same primary seed data is also taken and given as input for PREDATOR for the prediction of secondary structure [6]. Biochemical pathway information of *S. Proteomaculans* 568

was collected from PATRIC database [7](<ftp.patricbrc.org>) and the primary seed data was mapped to this pathway information. Using the .fna and .ptt files, the gene sequences were retrieved and given as input to 'cusp' program of emboss package for calculation of codon usage. A Perl program was developed to retrieve the upstream sequences of all the genes present in the chromosome. The front end of the repository was developed using HTML, and PHP.

IV. RESULTS AND DISCUSSION

The home page of the repository contains “Home”, “Retrieve Data”, “Help”, “Contact” options. In the home page there is brief introduction about the repository (Figure ure .1).

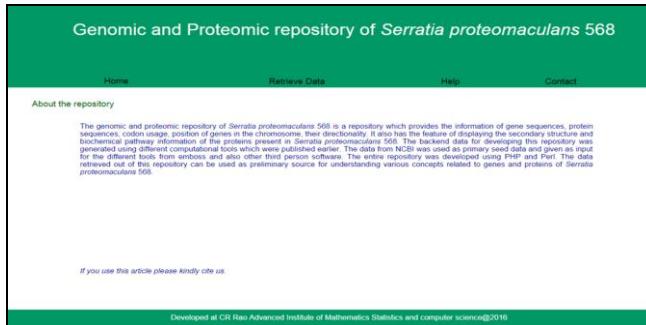


Figure 1: A snapshot showing the home page of the repository. Home page includes ‘Home’, ‘Retrieve Data’, ‘Help’ and ‘Contact’. Brief introduction about the repository is described.

A click on the retrieve data shows the list, which contain the ORF Ids, gene name, Protein ID, and its product (Figure 2). Upon single click on any of the ORF Id, the data is retrieved which shows the selected gene ORF Id, its directionality on the chromosome, start and end positions in the chromosome, Length of the protein, Gene name, and the function. For example, when *spro_0002* ORF Id is selected, then the page is navigated to a new page showing the details of the selected *spro_0002* ORF.

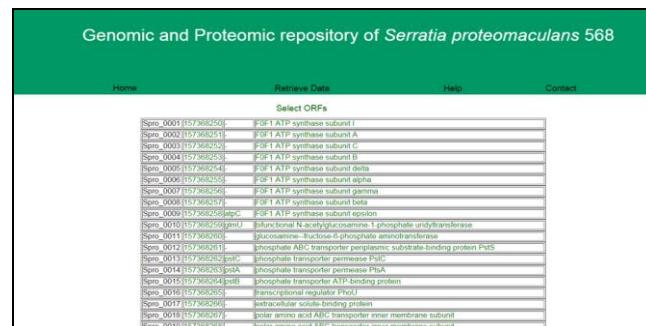


Figure 2: A snapshot showing the list of ORF Ids, gene names, Protein Ids, and their products.

The new page as shown in Figure 3a, shows that the directionality of the *spro_0002* gene was in 5' to 3' direct strand of the DNA. Moreover it also show that the ORF starts at 969th nucleotide and ends at 1170th nucleotide. It

also gives the information about the length of the protein it codes as 273 amino acids, with function as 'F0F1 ATP synthase subunit A' (Figure 3a).

Figure 3a: A snapshot showing the retrieved data for *spro_0002* gene. The data such as directionality start and end positions, protein length, and function along with the gene, protein and upstream sequences were displayed. There is also the information about the involvement of protein product of *spro_0002* various biochemical pathways such as methane metabolism.

Below this information, the page also displayed the gene, protein and upstream sequences in FASTA format. Figure 3a also shows the involvement of the gene product of the gene *spro_0002* in various pathways. From the retrieved data, it is found that the gene product of *spro_0002* is involved in Oxidative phosphorylation, Photosynthesis, and Methane metabolism (Figure 3a). Below the biochemical pathway information there is the codon usage information displayed in the form of a table. Next to the codon usage table, there is the information about the physico-chemical properties and secondary structure information for the protein encoded by *spro_0002* (Figure 3b).

Physico-chemical properties

Average Residue Weight = 771.17 Residues = 273
 Average Residue Weight = 111.250 Charge = 1.0
 A888 Molar Extinction Coefficient = 40540
 A888 Molar Refractivity = 101.000
 Improbability or expression in inclusion bodies = 0.641

Residue	Number	Mole%	DenehoffStat
Gly	05	0.000	0.000
Asp	06	0.000	0.000
Asx	07	0.000	0.000
Thr	08	0.000	0.000
Pro	09	0.000	0.000
Val	10	0.000	0.000
Glu	11	0.000	0.000
Leu	12	0.000	0.000
Ala	13	0.000	0.000
Leu	14	0.000	0.000
Leu	15	0.000	0.000
Asn	16	0.000	0.000
Asn	17	0.000	0.000
Asn	18	0.000	0.000
Gln	19	0.000	0.000
Gly	20	0.000	0.000
Leu	21	0.000	0.000
Ile	22	0.000	0.000
Ile	23	0.000	0.000
Ile	24	0.000	0.000
Leu	25	0.000	0.000
Leu	26	0.000	0.000
Leu	27	0.000	0.000
Leu	28	0.000	0.000
Leu	29	0.000	0.000
Leu	30	0.000	0.000
Leu	31	0.000	0.000
Leu	32	0.000	0.000
Leu	33	0.000	0.000
Leu	34	0.000	0.000
Leu	35	0.000	0.000
Leu	36	0.000	0.000
Leu	37	0.000	0.000
Leu	38	0.000	0.000
Leu	39	0.000	0.000
Leu	40	0.000	0.000
Leu	41	0.000	0.000
Leu	42	0.000	0.000
Leu	43	0.000	0.000
Leu	44	0.000	0.000
Leu	45	0.000	0.000
Leu	46	0.000	0.000
Leu	47	0.000	0.000
Leu	48	0.000	0.000
Leu	49	0.000	0.000
Leu	50	0.000	0.000
Leu	51	0.000	0.000
Leu	52	0.000	0.000
Leu	53	0.000	0.000
Leu	54	0.000	0.000
Leu	55	0.000	0.000
Leu	56	0.000	0.000
Leu	57	0.000	0.000
Leu	58	0.000	0.000
Leu	59	0.000	0.000
Leu	60	0.000	0.000
Leu	61	0.000	0.000
Leu	62	0.000	0.000
Leu	63	0.000	0.000
Leu	64	0.000	0.000
Leu	65	0.000	0.000
Leu	66	0.000	0.000
Leu	67	0.000	0.000
Leu	68	0.000	0.000
Leu	69	0.000	0.000
Leu	70	0.000	0.000
Leu	71	0.000	0.000
Leu	72	0.000	0.000
Leu	73	0.000	0.000
Leu	74	0.000	0.000
Leu	75	0.000	0.000
Leu	76	0.000	0.000
Leu	77	0.000	0.000
Leu	78	0.000	0.000
Leu	79	0.000	0.000
Leu	80	0.000	0.000
Leu	81	0.000	0.000
Leu	82	0.000	0.000
Leu	83	0.000	0.000
Leu	84	0.000	0.000
Leu	85	0.000	0.000
Leu	86	0.000	0.000
Leu	87	0.000	0.000
Leu	88	0.000	0.000
Leu	89	0.000	0.000
Leu	90	0.000	0.000
Leu	91	0.000	0.000
Leu	92	0.000	0.000
Leu	93	0.000	0.000
Leu	94	0.000	0.000
Leu	95	0.000	0.000
Leu	96	0.000	0.000
Leu	97	0.000	0.000
Leu	98	0.000	0.000
Leu	99	0.000	0.000
Leu	100	0.000	0.000
Leu	101	0.000	0.000
Leu	102	0.000	0.000
Leu	103	0.000	0.000
Leu	104	0.000	0.000
Leu	105	0.000	0.000
Leu	106	0.000	0.000
Leu	107	0.000	0.000
Leu	108	0.000	0.000
Leu	109	0.000	0.000
Leu	110	0.000	0.000
Leu	111	0.000	0.000
Leu	112	0.000	0.000
Leu	113	0.000	0.000
Leu	114	0.000	0.000
Leu	115	0.000	0.000
Leu	116	0.000	0.000
Leu	117	0.000	0.000
Leu	118	0.000	0.000
Leu	119	0.000	0.000
Leu	120	0.000	0.000
Leu	121	0.000	0.000
Leu	122	0.000	0.000
Leu	123	0.000	0.000
Leu	124	0.000	0.000
Leu	125	0.000	0.000
Leu	126	0.000	0.000
Leu	127	0.000	0.000
Leu	128	0.000	0.000
Leu	129	0.000	0.000
Leu	130	0.000	0.000
Leu	131	0.000	0.000
Leu	132	0.000	0.000
Leu	133	0.000	0.000
Leu	134	0.000	0.000
Leu	135	0.000	0.000
Leu	136	0.000	0.000
Leu	137	0.000	0.000
Leu	138	0.000	0.000
Leu	139	0.000	0.000
Leu	140	0.000	0.000
Leu	141	0.000	0.000
Leu	142	0.000	0.000
Leu	143	0.000	0.000
Leu	144	0.000	0.000
Leu	145	0.000	0.000
Leu	146	0.000	0.000
Leu	147	0.000	0.000
Leu	148	0.000	0.000
Leu	149	0.000	0.000
Leu	150	0.000	0.000
Leu	151	0.000	0.000
Leu	152	0.000	0.000
Leu	153	0.000	0.000
Leu	154	0.000	0.000
Leu	155	0.000	0.000
Leu	156	0.000	0.000
Leu	157	0.000	0.000
Leu	158	0.000	0.000
Leu	159	0.000	0.000
Leu	160	0.000	0.000
Leu	161	0.000	0.000
Leu	162	0.000	0.000
Leu	163	0.000	0.000
Leu	164	0.000	0.000
Leu	165	0.000	0.000
Leu	166	0.000	0.000
Leu	167	0.000	0.000
Leu	168	0.000	0.000
Leu	169	0.000	0.000
Leu	170	0.000	0.000
Leu	171	0.000	0.000
Leu	172	0.000	0.000
Leu	173	0.000	0.000
Leu	174	0.000	0.000
Leu	175	0.000	0.000
Leu	176	0.000	0.000
Leu	177	0.000	0.000
Leu	178	0.000	0.000
Leu	179	0.000	0.000
Leu	180	0.000	0.000
Leu	181	0.000	0.000
Leu	182	0.000	0.000
Leu	183	0.000	0.000
Leu	184	0.000	0.000
Leu	185	0.000	0.000
Leu	186	0.000	0.000
Leu	187	0.000	0.000
Leu	188	0.000	0.000
Leu	189	0.000	0.000
Leu	190	0.000	0.000
Leu	191	0.000	0.000
Leu	192	0.000	0.000
Leu	193	0.000	0.000
Leu	194	0.000	0.000
Leu	195	0.000	0.000
Leu	196	0.000	0.000
Leu	197	0.000	0.000
Leu	198	0.000	0.000
Leu	199	0.000	0.000
Leu	200	0.000	0.000
Leu	201	0.000	0.000
Leu	202	0.000	0.000
Leu	203	0.000	0.000
Leu	204	0.000	0.000
Leu	205	0.000	0.000
Leu	206	0.000	0.000
Leu	207	0.000	0.000
Leu	208	0.000	0.000
Leu	209	0.000	0.000
Leu	210	0.000	0.000
Leu	211	0.000	0.000
Leu	212	0.000	0.000
Leu	213	0.000	0.000
Leu	214	0.000	0.000
Leu	215	0.000	0.000
Leu	216	0.000	0.000
Leu	217	0.000	0.000
Leu	218	0.000	0.000
Leu	219	0.000	0.000
Leu	220	0.000	0.000
Leu	221	0.000	0.000
Leu	222	0.000	0.000
Leu	223	0.000	0.000
Leu	224	0.000	0.000
Leu	225	0.000	0.000
Leu	226	0.000	0.000
Leu	227	0.000	0.000
Leu	228	0.000	0.000
Leu	229	0.000	0.000
Leu	230	0.000	0.000
Leu	231	0.000	0.000
Leu	232	0.000	0.000
Leu	233	0.000	0.000
Leu	234	0.000	0.000
Leu	235	0.000	0.000
Leu	236	0.000	0.000
Leu	237	0.000	0.000
Leu	238	0.000	0.000
Leu	239	0.000	0.000
Leu	240	0.000	0.000
Leu	241	0.000	0.000
Leu	242	0.000	0.000
Leu	243	0.000	0.000
Leu	244	0.000	0.000
Leu	245	0.000	0.000
Leu	246	0.000	0.000
Leu	247	0.000	0.000
Leu	248	0.000	0.000
Leu	249	0.000	0.000
Leu	250	0.000	0.000
Leu	251	0.000	0.000
Leu	252	0.000	0.000
Leu	253	0.000	0.000
Leu	254	0.000	0.000
Leu	255	0.000	0.000
Leu	256	0.000	0.000
Leu	257	0.000	0.000
Leu	258	0.000	0.000
Leu	259	0.000	0.000
Leu	260	0.000	0.000
Leu	261	0.000	0.000
Leu	262	0.000	0.000
Leu	263	0.000	0.000
Leu	264	0.000	0.000
Leu	265	0.000	0.000
Leu	266	0.000	0.000
Leu	267	0.000	0.000
Leu	268	0.000	0.000
Leu	269	0.000	0.000
Leu	270	0.000	0.000
Leu	271	0.000	0.000
Leu	272	0.000	0.000
Leu	273	0.000	0.000
Leu	274	0.000	0.000
Leu	275	0.000	0.000
Leu	276	0.000	0.000
Leu	277	0.000	0.000
Leu	278	0.000	0.000
Leu	279	0.000	0.000
Leu	280	0.000	0.000
Leu	281	0.000	0.000
Leu	282	0.000	0.000
Leu	283	0.000	0.000
Leu	284	0.000	0.000
Leu	285	0.000	0.000
Leu	286	0.000	0.000
Leu	287	0.000	0.000
Leu	288	0.000	0.000
Leu	289	0.000	0.000
Leu	290	0.000	0.000
Leu	291	0.000	0.000
Leu	292	0.000	0.000
Leu	293	0.000	0.000
Leu	294	0.000	0.000
Leu	295	0.000	0.000
Leu	296	0.000	0.000
Leu	297	0.000	0.000
Leu	298	0.000	0.000
Leu	299	0.000	0.000
Leu	300	0.000	0.000
Leu	301	0.000	0.000
Leu	302	0.000	0.000
Leu	303	0.000	0.000
Leu	304	0.000	0.000
Leu	305	0.000	0.000
Leu	306	0.000	0.000
Leu	307	0.000	0.000
Leu	308	0.000	0.000
Leu	309	0.000	0.000
Leu	310	0.000	0.000
Leu	311	0.000	0.000
Leu	312	0.000	0.000
Leu	313	0.000	0.000
Leu	314	0.000	0.000
Leu	315	0.000	0.000
Leu	316	0.000	0.000
Leu	317	0.000	0.000
Leu	318	0.000	0.000
Leu	319	0.000	0.000
Leu	320</		

Figure 3b: A snapshot showing the predicted physico-chemical properties and secondary structure (partial) for the protein encoded by *sarc_0002*.

The physico-chemical properties include the information about the protein's molecular weight, iso-electric point, net charge and many other protein properties along with data of combination of amino acids and their distribution in the protein. The predicted secondary structure information includes the helices, coils and beta strands in the protein sequence.

V. SIGNIFICANCE OF RETRIEVED DATA

The retrieved gene sequences can be used as templates for performing the regular molecular biology works such as primer designing, gene cloning, protein expression etc. The retrieved protein sequences can be used for 3D modelling, docking and molecular dynamics simulations for better understanding the structure of the proteins. Upstream sequences of the genes can be used for prediction *cis* regulatory elements, non-coding RNA etc. Pathway information serves important information about the selected gene products which is very useful while doing experiments related to metabolomics. Understanding the codon usage and bias in codon usage is important tool for discovering new genes, gene expression, origin of species etc [8-10]. The predicted physico-chemical properties would provide preliminary information about the protein characteristics such as iso-electric point, charge etc which play crucial information in expression of protein in laboratory conditions. The secondary structure provides information about the composition of proteins such as helices, coils and beta turns which aid researcher in understanding the stability of the protein.

VI. CONCLUSION AND FUTURE SCOPE

As repository has gene and protein sequences in FASTA format, where the users can directly use them for research work. Apart from these sequences, the information about the location of the ORF along with its directionality enables the users in design of primers for their molecular biology work such as cloning and expression. Researchers who are working in the area of analysis of proteins and their substitutions of amino acids can make use of this codon usage table. The predicted physico-chemical properties provide the user for understanding the nature of the protein such that suitable medium can be used for its successful into soluble fraction, avoiding the formation of inclusion bodies. Secondary structure information can be useful for the users to understand about the structural stability. The retrieved upstream sequences can be used to perform the experiments such as protein DNA interactions. In coming future, more number of *Serratia* species will be added to this repository.

REFERENCES

[1]. S. Taghavi, C. Garafola, S. Monchy, L. Newman, A. Hoffman, N. Weyens, T. Barac, J. Vangronsveld, D. van der Lelie, "Genome survey and characterization of endophytic bacteria exhibiting a beneficial effect on growth and development of poplar trees", Applied and environmental microbiology, Vol.75, Issue.3, pp.748-757,2009.

[2]. P. Purushotham, P.V. Arun, J.S. Prakash, A.R. Podile, "Chitin binding proteins act synergistically with chitinases in *Serratia proteamaculans* 568", PloS one, Vol.7, Issue.5, pp.e36714, 2012.

[3]. M. Nakao, S. Okamoto, M. Kohara, T. Fujishiro, T. Fujisawa, S. Sato, S. Tabata, T. Kaneko, Y. Nakamura, "CyanoBase: the cyanobacteria genome database update 2010", Nucleic acids research, Vol.38, Issue.-, pp.D379-381, 2010.

[4]. P.V. Arun, R.K. Bakku, M. Subashini, P. Singh, N.P. Prabhu, I. Suzuki, J.S. Prakash, "CyanoPhyChe: a database for physico-chemical properties, structure and biochemical pathway information of cyanobacterial proteins", PloS one, Vol.7, Issue.11, pp.e49425, 2012.

[5]. P. Rice, I. Longden, A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite, Trends in genetics : TIG", Vol.16, Issue.6, pp.276-277, 2000.

[6]. D. Frishman, P. Argos, "Seventy-five percent accuracy in protein secondary structure prediction", Proteins, Vol.27, Issue.3, pp.329-335, 1997.

[7]. A.R. Wattam, D. Abraham, O. Dalay, T.L. Disz, T. Driscoll, J.L. Gabbard, J.J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E.K. Nordberg, R. Olson, R. Overbeek, G.D. Pusch, M. Shukla, J. Schulman, R.L. Stevens, D.E. Sullivan, V. Vonstein, A. Warren, R. Will, M.J. Wilson, H.S. Yoo, C. Zhang, Y. Zhang, B.W. Sobral, "PATRIC, the bacterial bioinformatics database and analysis resource", Nucleic acids research, Vol.42, Issue.-, pp.D581-591, 2014.

[8]. I. Ahn, B.J. Jeong, S.E. Bae, J. Jung, H.S. Son, "Genomic analysis of influenza A viruses, including avian flu (H5N1) strains", Eur J Epidemiol, Vol.21, Issue.7, pp.511-519, 2006.

[9]. J.F. Kane, "Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*", Curr Opin Biotechnol, Vol.6, Issue.5, pp.494-500, 1995.

[10]. X. Yang, X. Luo, X. Cai, "Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset", Parasit Vectors, Vol.7, Issue.1, pp.527, 2014.

Authors Profile

Dr. P.V Parvati Sai Arun pursued Bachelor of Technology from Jawaharlal Nehru Technological in 2007 and Master of Technology from University of Hyderabad in year 2009. He also received Ph.D. from University of Hyderabad and currently working as Post doctoral fellow since 2015. He has published more than 5 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences and it's also available online. His main research work focuses on Computational biology, Bioinformatics based education.