International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

Design and Implementation of Bioinformatic Tools to Analyze Protein Sequences

Sabbavarapu N V Siva Kumar & Kunjam Nageswara Rao
[1] M.Tech CST-Bioinformatics, CS&SE Department, Andhra University, Visakhapatnam [2]Associate Professor, CS&SE Department, Andhra University, Visakhapatnam.

***Abstract**:*

Proteins are one of the important fundamental units of all living cells. Proteins have a large number of functions within all the living organisms. Some of the important functions performed with the help of proteins such as DNA replication, catalysis of metabolic reactions, transportation of molecules from one place to another etc... There are 20 different types of amino acids that can be combined to make a protein sequence. Integrins are cell adhesion molecules that mediate cell-cell, cellextracellular matrix, and cell-pathogen interaction. They play crucial roles in leukocyte trafficking, migration, immunological synapse formation, costimulation, and phagocytosis. The studies of integrin molecule structure and function helps in better understanding the various pathological processes such as chronic inflammation, thrombosis and cancer metastasis. The study can help the researchers to discover a new drug to cure all these.To analyze the protein sequence with respect to primary structure, secondary structure and tertiary structure it needs the help of various bioinformatic tools. Most of these tools are web based tools and it requires internet connection to work with them. There are lot of challenges in maintaining proper internet connection all the time. So, developing an offline tool using high performance programming language like julia will be helpful for researchers to work with them at any time irrespective of internet connection.

Keywords

integrin $\alpha_v\beta_3$; protein analysis; phylogenetic analysis; biological pathways, offline bioinformatic tools, bioinformatic tools, protparam tool, protein structures

1.Introduction

Proteins plays a crucial role in cell functions. Proteins are large, complex molecules. Proteins are required for the structure, function and regulation of body tissues and organs. Proteins sequences are the combination of 20 amino acids that forms a long chain. Proteins are the responsible for various functions in our body. For example, Integrin are important members in the family of cell adhesion molecules mediating the signal transmission between extracellular matrix(ECM) and intracellular in many physiological reaction processes. They play an important part in regulating cell adhesion, proliferation, differentiation, metastasis and apoptosis.

Integrin $\alpha_v\beta_3$ is expressed in many cell types and combine with various ligands in many cellular activities, involved in many physiological and pathological processes. It is important on tumor angiogenesis [1]. It is capable of identifying the short Arg-Gly-Asp(RGD) peptide motif [2]. The ligands including FN, VN, TSP-1 and vWF, etc. Studies have found that monoclonal antibodies and antagonists of integrin $\alpha_v\beta_3$ can effectively inhibit the formation of tumor blood vessels, it may be a new mean of cancer treatment [3].

2.Materials & Methods

A.Tools for prediction of protein structure and function

ProtParam(http://kr.expasy.org/tools/protparam.html)[4], a network analyzer tool to predict protein primary structure, is used to predict the amino acid composition and characterization of the protein, including the molecular weight, extinction coefficient, theoretical PI, half life, aliphatic index, instability index and grand average of hydropathicity(GRAVY)[5]. To predict protein domain, signal peptide, transmembrane region and subcellular localization properties of primary structure Simple Modular Architecture Research Tool(SMART) (http://smart.embl-heidelberg.de/)[6] is a classification scheme used in the identification and analysis of protein domain. SignalP-4.1(http://www.cbs.dtu.dk/services/SignalP) is used to perform signal peptide prediction[7]. Transmembrane Helices Hidden Markov Model tool TMHMM(http://www.cbs.dtu.dk/services/TMHMM -2.0/) is a used for the prediction of the transmembrane domains. WoLF PSORT (https://wolfpsort.hgc.jp) is used to predict the subcellular localization[8]. Garnier-Osguthorpe-

Robson tool GOR4.0 (http://npsa-pbil.ibcp.fr/cgibin/npsa_automat.pl?page=npsa _gor4.html) is used for the prediction of the protein secondary structure.

ProtParam, a network analyzer about protein primary structure to predict the amino acid composition and characterization of the protein. ProtParam is a web based tool and it requires internet connection to work with it. An Offline tool has been developed using Julia programming language to speed up the process. Following parameters can be calculated using this offline tool.

Molecular Weight: In ProtParam, the molecular weight of protein is calculated by the addition of average isotopic masses of amino acids in the provided protein and the average isotopic mass of one water molecule.

Extinction Coefficient: The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this extinction coefficient of a protein from knowledge of its amino acid composition. From the molar extinction coefficient of tyrosine, tryptophan, and cystine (cysteine does not absorb appreciably at wavelengths >260 nm, while cystine does) at a given wavelength, the extinction coefficient of the native protein in water can be calculated using the following equation:

$$E(Prot) = Numb(Tyr) * Ext(Tyr) + Numb(Trp) * Ext(Trp) + Numb(Cystine)*Ext(Cystine)$$

Where (for proteins in water measured at 280 nm): Ext(Tyr) = 1490, Ext(Trp) = 5500, Ext(Cystine) = 125;

The absorbance (optical density) can be calculated using the following formula:

$$Absorb(Prot) = E(Prot) / Molecular\_weight$$

Theoretical pI: Theoretical pI of of protein is calculated using pKa values of amino acids. The pKa value of Amino acids depends on its side chain. It plays an important role in defining the pH dependent characteristics of a protein.

Half-life: The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell.

Grand average of hydropathicity (GRAVY): The GRAVY value for a protein or a peptide is calculated by adding the hydropathy of each amino acid residues and dividing by the number of residues in the sequence or length of the sequence. Increasing positive score indicates a greater hydrophobicity.

Aliphatic index: The aliphatic index of a protein is described as the relative volume occupied by the amino acids such as alanine, valine, isoleucine and leucine, which have an aliphatic side chain in their structure. The aliphatic index of a protein is computed using the following equation.

$$Aliphatic\ index = X(Ala) + a * X(Val) + b * (X(Ile) + X(Leu))$$

Where X(Ala), X(Val), X(Ile), and X(Leu) are mole percent (100 X mole fraction) of alanine, valine, isoleucine, and leucine.

The coefficients ‘a’ and ‘b’ are the relative volume of valine side chain (a = 2.9) and of Leu/Ile side chains (b = 3.9) to the side chain of alanine.

Instability Index: The instability index provides an estimation of the stability of your protein in a test tube. A protein with instability index smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable. Instability Index can be computed using the following equation.

$$II = (10/L) * \underset{i=1}{\overset{i=L-1}{Sum}} \quad DIWV(x[i]x[i+1])$$

where: L is the length of sequence

DIWV(x[i]x[i+1]) is the instability weight value for the dipeptide starting in position i.

B.Phylogenetic analysis

Using ClustalW, a multiple sequence alignment was performed between the amino acid sequences of $\alpha_v\beta_3$ in the different species, and after that to construct a phylogenetic tree for the molecular evolution based on the Neighbor-Joining using Molecular Evolutionary Genentic Analysis tool MEGA 6.0 tool [9].

C.Analysis of biological pathway

The intercomparison results between this species and others will be gotten from the database of Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathways (http://www.genome.jp/kegg/) [10].

3.Result

A.The Protein Structure of Integrin $\alpha_v\beta_3$

The accession number of α_v and β_3 subunit are P06756 and P05106 respectively in Uniprot database. By using ProtParam, the chemical property of integrin $\alpha_v\beta_3$ was gotten and arranged below.

TABLE I.The Result of Primary Structure Analysis of $\alpha v\beta_3$

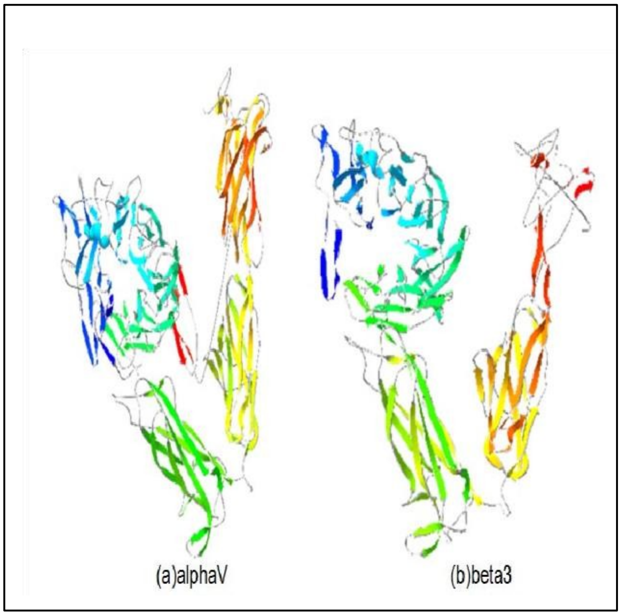
Items	Online Values		Offline Values	
	α_v	β_3	α_v	β_3
The number of amino acids	1048	788	1048	788
The total number of atoms	16245	11981	16245	11981
Theoretical pI	5.45	5.09	5.452536	5.090929
Aliphatic Index	86.15	72.72	86.14902	72.72067
GRAVY	-0.220 (hydrophile)	-0.332 (hydrophile)	-0.220115 (hydrophile)	-0.33215 (hydrophile)
Molecular weight	116051.8	87057.8	116051.8	87057.8
Extinction coefficient	111255/110130	86240/82740	111255 /110130	86240/82740
Instability Index	37.69	45.10	37.69012	45.10495
Half-life period (Mammals reticulocyte, out of body)	30 hours	30 hours	30 hours	30 hours

In Table-I the results were compared between existing protparam tool which is in expasy server and offline tool which was developed using high performance programming language “Julia”. The results have been shown that offline tool values and online tool values are equal and even the offline tool generating results faster than online tool because results generated by online tool depends on the speed of internet connection.

The most common types of secondary structures are the α helix and the β pleated sheet. Both structures are held in shape by hydrogen bonds, which form between the carbonyl O of one amino acid and the amino H of another.The secondary structures of α_v and β_3 were analyzed with SMART. The results showed that the subunit of $\alpha_v$is consisted with the lamellas occupy 25.38%, the spirals occupy 4.48%, and the rings occupy 70.13%. There is no curly spiral. For the subunit of β_3, the lamellas occupy 6.22%, the spirals occupy 13.83%, and the rings occupy 79.95%. There are four potential curly spirals.

The tertiary structure of related proteins of $\alpha_v\beta_3$ are analyzed with Swiss-Model. The tertiary structure model corresponded with α_v and β_3 is the crystal structure 1L5G(Figure-1). In Figure-1, each secondary structure is signed by different color. Green color indicates lamellas, yellow color indicates spirals, blue and navy-blue combination indicates rings.

Figure 1. The tertiary template of α_v(a) and β_3(b) subunits.



B.The Evolution Analysis of Integrin $\alpha_v\beta_3$

Using MEGA tool, the protein sequences of different species were aligned with human α_v and β_3 sequences and constructed the evolutionary trees, and the relationship was shown in Figure-2. For α_v subunit, only the part within the purple frame is different from this before excluding. For β_3 subunit, only the position of Sus scrofa is different in the evolutionary tree.

In different species, the structural domains are conserved in the process of different species evolution. For α_v subunit, the conservatism of transmembrane is highest. For β_3 subunit, the conservatism of integrin tail is highest among all the structural domains. The relationship shown in the following Figure-2 is what after getting the evolutionary trees by comparing protein sequences of different species from uniport.

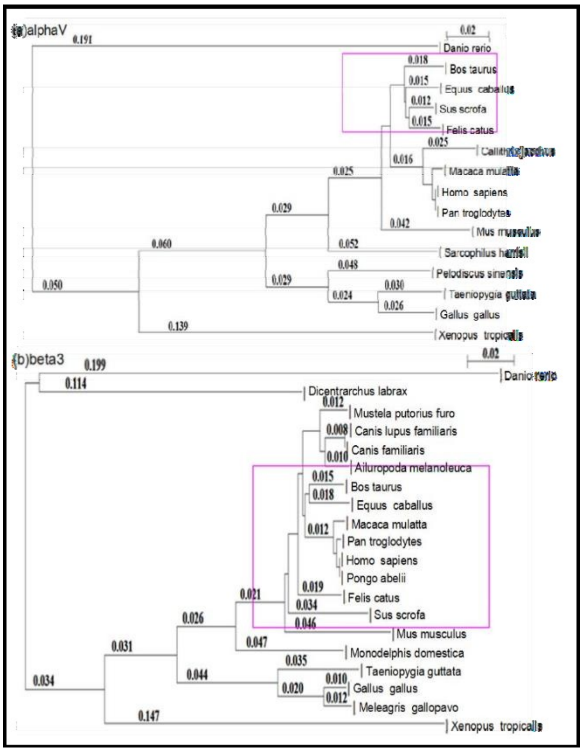


Figure 2. The Commonness Analysis of α_v(a) and β_3(b) in Different Species.

TABLE II. The Conservatism Analysis Of αV Subunit and β_3 Subunit

The name of structural domain	he position of Structural Domain	Conservative average	Conservatism
α_v subunit			
FG-GAP 1	32-98	57.57	middle
FG-GAP 2	109-170		
FG-GAP 3	173-225		
FG-GAP 4	237-295		
FG-GAP 5	296-357		
FG-GAP 6	358-415		
FG-GAP 7	419-482		
Transmembrane	993-1016	100.00	high
Cytoplasmic	1017-1048	67.00	middle
β_3 subunit			
PSI	30-76	48.49	low
INB	38-461	79	middle
VWFA	135-377	85	high
IEGF-1	463-511	89.125	high
IEGF-2	512-553		
IEGF-3	554-592		
IEGF-4	593-629		
β tail	634-718	100	high
Transmembrane	719-741	96	high
Cytoplasmic	742-788	74	high

In following Table-II definition, the average of the mutation is taken from 0 to 50 is low, from 50 to 80 is middle and from 80 to 100 is high. The table contains the locations of protein sequence domains, transmembrane region and subcellular localization of αv and β3 protein sequences and displayed with respective conservatism average and conservatism level for obtaining the target protein sequence.

C.Analysis of Biological pathways of Integrin $\alpha_v\beta_3$

Biological pathways of homologous gene in α_v subunit and $\beta_3$subunit were analyzed with KEGG. In this study, human and bonobo are selected to compare.

Table-III contains the information that the biological pathways related to α_v subunits of humans is same as the pathways related to $\alpha_v$subunits of pan paniscus, only some biological factors are different. And β_3 subunit has the same situation as α_v subunit in the evolution of similar species. Thus, can conclude that the α_v subunit and β_3 subunit in the evolution of similar species are almost exactly the same, which reflected in its homology between different species. Both α_v and β_3 subunits biological pathways are compared between different species and bonobo is the nearest matching species among all other species. The pathway comparison is shown in the following table which clearly illustrates the matching biological pathways of α_v and β_3 subunits of homosapiens and pan paniscus.

After analyzing each pathway associated with integrin $\alpha_v\beta_3$, the results can be concluded that integrin $\alpha_v\beta_3$ influence all aspects of biological reactions and physiological processes must be through the mutual reaction and extracellular matrix (ECM) to really play its due role, so the research on interaction pathway in ECM-receptor is necessary.

TABLE III. The Pathway Comparison Between Human and Bonobo

hsa3685' Homo sapiens (human)		pps100975589 Pan paniscus (bonobo)	
Number of Pathway	Name ofPathway	umber of Pathway	Name ofPathway
α_v subunit			
hsa04145	Phagosome	pps04145	Phagosome
hsa04151	PI3K-Akt signaling pathway	pps04151	PI3K-Akt signaling pathway

hsa04510	Focal adhesion	pps04510	Focal adhesion
hsa04512	CM-receptor interaction	pps04512	CM-receptor interaction
hsa04514	ell adhesion molecules (CAMs)	pps04514	ell adhesion molecules(CAMs)
β_3 **subunit**			
hsa04015	p1 signaling pathway	ggo04015	p1 signaling pathway
hsa04145	Phagosome	ggo04145	Phagosome
hsa04151	PI3K-Akt signaling pathway	ggo04151	PI3K-Akt signaling pathway
hsa04380	Osteoclast differentiation	ggo04380	Osteoclast differentiation
hsa04510	Focal adhesion	ggo04510	Focal adhesion

It showed that the research of integrin and tumor cancer have a high prospect for the future, because from the table get the results that pathways related with integrin $\alpha_v\beta_3$ can lead to cell proliferation, avoidance of apoptosis, and sustained angiogenesis. While other physiological outcomes such as the metabolism, transport and translation of glucose, and the stabilization of actin are also likely to be provided for the above three major results. The occurrences of cancer are inextricably linked with those results, because when cells are able to proliferate, evading apoptosis and generate endothelial cell consistently, can think the cells have the ability of infinite proliferation, which is meaned cancerous.

The related molecules such as some protein kinase that SRC, FAK, PI3K, PIP3, PKB/Akt play key roles in the process of biological pathways. Thus, can verify the molecular by experiment, analyzing the position which the drug target on can make an effect of inhibiting the tumor cell proliferation, therefore obtain the effective treatment for cancer.

Activated Akt can not only limit Raf-1, GSA3, P21, P27, FOXO, BAD, Casp, RXR and NUR77, but also promote eNOS(the NO synthetase), GREB, IKK and MDM2, keeping them from entering the nucleus to play normal function, thus affect the process of cell cycle. The PI3K Akt signaling pathway that associated with this pathway and integrin lead to cell survival, the normal process of cell cycle, metabolism, cell proliferation, angiogenesis and DNA repair. As is known to all, the characteristics of cancer cells is unlimited, endless proliferation. Therefore, the study of these signaling pathways has a great significance.

4.Summary

Work shown in this paper is mainly about studying the biological information related to integrin $\alpha_v\beta_3$, including the protein structure and function prediction, the system evolution analysis, and the analysis of biological pathways of integrin $\alpha_v\beta_3$. By using the related biological informatic tools to understand the physiological processes that integrin $\alpha_v\beta_3$ involved cancer and other diseases. It is helpful to provide new insights for the treatment of these diseases. In the process of predicting the chemical properties of protein sequence an offline tool has been developed to save the time by producing the accurate results of the chemical properties of protein sequence without using protparam tool from expasy server that requires internet connection.

5.References

[1] M. Kim, C. V. Carman, and T. A. Springer, "Bidirectional transmembrane signaling by cytoplasmic domain separation in integrins," *Science,* vol. 301, pp. 1720-1725, 2003.

[2] E. Ruoslahti, "RGD and other recognition sequences for integrins," *Annual review of cell and developmental biology,* vol. 12, pp. 697-715, 1996.

[3] F. Danhier, A. L. Breton, and V. r. Préat, "RGD-based strategies to target alpha (v) beta (3) integrin in cancer therapy and diagnosis," *Molecular pharmaceutics,* vol. 9, pp. 2961-2973, 2012.

[4] *Protein Identification and Analysis Tools on the ExPASy Server;* Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005). pp. 571-607.

[5] A. Sudhakar, H. Sugimoto, C. Yang, J. Lively, M. Zeisberg, and R. Kalluri, "Human tumstatin and human endostatin exhibit distinct antiangiogenic activities mediated by v3 and 51 integrins," *Proceedings of the National Academy of Sciences,* vol. 100, pp. 47664771, 2003.

[6] I. Letunic, R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, *et al.*, "SMART 4.0: towards genomic data integration," *Nucleic acids research,* vol. 32, pp. D142-D144, 2004.

[7] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0," *Journal of molecular biology,* vol. 340, pp. 783795, 2004.

[8] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. Adams-Collier, *et al.*, "WoLF PSORT: protein localization predictor," *Nucleic acids research,* vol. 35, pp. W585W587, 2007.

[9] S. Kumar, M. Nei, J. Dudley, and K. Tamura, "MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences," *Briefings in bioinformatics,* vol. 9, pp. 299306, 2008.

[10] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research,* vol. 28, pp. 27-30, 2000.