Statistical approach for lysosomal membrane proteins (LMPs) identification

Vijay Tripathi · Pooja Tripathi · Dwijendra Gupta

Received: 11 March 2014 / Revised: 11 June 2014 / Accepted: 26 July 2014 / Published online: 2 August 2014
© Springer Science+Business Media Dordrecht 2014

Abstract Discrimination of Lysosomal membrane proteins (LMP's) from folding types of globular (GPs) and other membrane proteins (OtMPs) is an important task both for identifying LMPs from genomic sequences and for the successful prediction of their secondary and tertiary structures. We have systematically analyzed the amino acid frequencies as well as dipeptide count of GPs, LMPs and OtMPs. Based on the above calculated single amino acid frequency combined with dipeptide count information, we statistically discriminated LMPs from GPs and OtMPs. This approach correctly classified the LMPs with an accuracy of 95 %. On the other hand, the amino acid frequency alone can discriminate LMPs with an accuracy of only 79 %. Similarly dipeptide count alone has an accuracy of 87 % for the discrimination of LMPs. Thus the combined information of both amino acid frequencies and dipeptide composition gives us significant high accurate results.

Keywords Lysosomal membrane proteins · Globular proteins · Other membrane proteins · Amino acid frequency · Dipeptide composition

Abbreviations

AA	Amino acid
LMPs	Lysosome membrane proteins
LSDs	Lysosomal Storage Disorders
GPs	Globular proteins
OtMPs	Other membrane proteins
PseAA	Pseudo amino acid
PDB	Protein Databank

Introduction

Lysosomes are cytoplasmic organelles present in almost all eukaryotic cells, which play a fundamental role in aspects of cellular homeostasis such as membrane repair, autophagy, endocytosis and protein metabolism. LMPs are synthesized as soluble or membrane integrated glycoproteins. These have demonstrated an important role in lysosomal acidification, transport of metabolites resulting from hydrolytic degradation, interaction and fusion with other cellular membrane systems. Mutations that cause lysosomal enzyme deficiencies result in different syndromes, known as Lysosomal storage disorders (LSDs). Most of the LSDs are associated with abnormal brain development and mental retardation. Lysosomal disorders represent a group of at least 40 genetic diseases, each of which results from a deficiency of one or more proteins involved in the degradation of macromolecules in lysosomes. This work emphasizes on the discrimination of LMPs from various other types of

Electronic supplementary material The online version of this article (doi:10.1007/s11693-014-9153-7) contains supplementary material, which is available to authorized users.

V. Tripathi · P. Tripathi · D. Gupta
Center of Bioinformatics, University of Allahabad, Allahabad, India

Present Address:

V. Tripathi (✉)
Genome Diversity Center, The Institute of Evolution, University of Haifa, Haifa, Israel
e-mail: vtripath@campus.haifa.ac.il

D. Gupta
Department of Biochemistry, University of Allahabad, Allahabad, India

membrane proteins as well as globular proteins on the basis of specific characteristics of membrane proteins. Previously most of the machine learning techniques used for the prediction of membrane protein types like Tripathi et al. (2014) used artificial neural network to discriminate LMPs from GPs to OtMPs, Yang et al. predicted membrane protein types on the basis of dipeptide as well as amino acid (AA) composition, Cai et al. (2004) used support vector machine and amino acid composition, Sonnhammer et al. (1998) have used the hidden markov model for predicting topology of membrane protein types. Similarly, Liu et al. (2005) have employed the Fourier spectrum and SVM, while Wang et al. (2004) have used weighted Support vector machine and pseudo amino acid composition. Wang et al. (2006) have used PseAA and stacked generalization. Chou and Shen (Chou and Shen 2007, 2009) developed a web server for the prediction membrane protein types. In this work the large dataset was used and computed the amino acid composition for 20 amino acid residues in GPs, OtMPs and LMPs. We have systematically analyzed the differences and similarities in these groups of proteins and devised a statistical method based on the deviation of amino acid and dipeptide composition for discriminating LMPs. We have tested our approach with several sets of globular proteins, other membrane proteins (OtMPs) belonging to different structural classes and LMPs obtained from well annotated sequences. The predicted results showed an accuracy of 95 % for correctly picking up the LMPs from known annotated sequences. This method is able to exclude up to 87 % of GPs and membrane proteins. These accuracy levels are comparable to or better than other methods previously applied.

Methods

Datasets

The protein sequences were retrieved from the UniProtKB database (<http://www.uniprot.org/>). This database has the category 'Subcellular location' in which protein sequences are categorized according there subcellular localization in the cell. In this study we have retrieved sets of data for calculating the amino acid composition, identifying LMPs and excluding GPs and OtMPs. The dataset contain 768 proteins of LMPs from lysosome membrane category, 595 OtMPs from membrane protein category and 557 GPs belonging to all structural classes from the cytoplasm category of the database, for computing the amino acid composition of 20 amino acid residues as well as dipeptide calculations of 400 dipeptides. Further, the method was tested with three other datasets: (1) a subset of 27 non-redundant OtMPs sequences with less than 35 % sequence identity: (2) a non-redundant dataset of 49 GPs available in

Protein Data Bank with a sequence identity of <25 % (Berman et al. 2000) and (3) a dataset of 17 well-annotated lysosome membrane proteins from the Orientations of membrane proteins in membrane database Lomize et al. (2006) (<http://opm.phar.umich.edu/>).

Discrimination of protein sequences

The statistical method earlier applied for discriminating membrane proteins by Gromiha et. al. (2005) using only amino acid frequency and Gromiha and Suwa (2005) using dipeptide composition only, with the accuracy 84 and 93 % in both the cases respectively. Here we have inherited their methods, but by combining them together, which shows an increase in efficiency.

Amino acid frequency

The amino acid frequencies for each set of proteins GPs, OtMPs and LMPs considered in this study was computed using the number of amino acids of each type and the total number of residues. It is defined as:

$$AA_freq(i) = \sum n_i / N \quad (1)$$

where i stands for the 20 amino acid residues. n_i is the number of residues of each type and N is the total number of residues. The summation is over all the residues in all the considered proteins. The total number of residues in the datasets of GPs, OtMPs and LMPs are 197804, 291543 and 333669 respectively.

Dipeptide composition

The composition of all the 400 (20×20) dipeptides based on the distribution of amino acid residues along the sequences of each set of proteins GPs, OtMPs and LMPs has been computed using the following expression:

$$Dipeptide(i,j) = \frac{\sum N_{ij}}{\sum N_i + \sum N_j} \quad (2)$$

where i, j stands for the distribution of 20 amino acid residues at positions i and $i + 1$. N_{ij} is the number of residues of type i followed by the residue j . $\sum N_i$ and $\sum N_j$ are the total number of residues of type i and j , respectively. The total number of dipeptide count in the datasets of GPs, OtMPs and LMPs, are 197227, 290721 and 332882, respectively.

Discrimination of Lysosome membrane proteins

Amino acid frequencies alone are not efficient for discrimination of any proteins, if dipeptide composition

information is added for protein sequences than the results significantly improves. We combined both the information of amino acid frequency and dipeptide composition results by just concatenating the two matrices.

We have followed the below mentioned steps to discriminate LMPs: (1) calculated the amino acid frequency and dipeptide composition for all three dataset *i.e.*, GPs, OtMPs and LMPs and combine them together in one matrix by simply concatenating them together, where the first column of the matrix is amino acid frequency and rest of the matrix is dipeptide composition matrix, the size of the matrix is now 20×21 . The difference between the matrices are also calculated (LMP–OtMP) and (LMP–GP); (2) for a new protein, P, the amino acid frequency and dipeptide composition calculated by using Eqs. (1) and (2), concatenate them as told previously and given weights to the matrix of protein M_P with (LMP–OtMP) and (LMP–GP), *i.e.*, $[M_P * (LMP\text{--}OtMP)]$ (3) calculated the sum of weighted $[M_P * (LMP\text{--}OtMP)]$ matrix and (4) the protein P is predicted to be an LMP if the total weighted amino acid frequency and dipeptide composition is positive and GP or OtMP otherwise.

Self-consistency and jack-knife test methods

The same proteins were used to predict whether each protein is of lysosome membrane proteins or globular or outer membrane type. This method is called self-consistency test (or back-check prediction). For the validity check prediction, followed the procedures that are widely used in the literature for protein secondary structure and solvent accessibility predictions: a set of N proteins is split into equally balanced subsets; parameters are developed on M proteins and then tested on the remaining N–M proteins (Cuff and Barton 1999; Ahmad and Gromiha 2002); the procedure is repeated for all subsets of data to obtain the average accuracy. This type of test is known as validity-check prediction. Further, we performed a jack-knife test

using a dataset of 35 LMP sequences, which have <35 % sequence identity with each other. We computed the amino acid frequency and dipeptide composition using 34 LMPs and used this information for assigning the type of the left-out protein. In these procedures, the tested proteins contain no information about the training set and hence the prediction accuracy obtained with this method is reliable. We have also examined the reliability of our method with two other datasets in which no information is used to derive the amino acid frequency and dipeptide composition. These datasets include (1) 50 non-redundant globular proteins and (2) 55 OtMPs.

Results and discussion

Amino acid frequencies in GPs, OtMPs and LMPs

The amino acid composition for 20 amino acid residues in LMPs, GPs and OtMPs have been computed using Eq. (1) and the results are displayed in Fig. 1 and values in Table 1 (Values in bold indicates higher frequencies of that amino acid). The residues *Phe*, *Leu*, *Gly*, *Asp*, *Glu* and *Thr* show subtle difference amongst the composition of GPs, LMPs and OtMPs. While the composition of *Phe*, *Leu*, *Cys*, *Val* and *Gly* are higher in LMPs than OtMPs and GPs, an opposite trend is observed for *Tyr*, *Trp*, *Gln*, *Pro* and *Ala*. The formation of disulfide bonds between *Cys* residues requires an oxidative environment and such disulfide bridges are therefore found in membrane proteins (Branden and Tooze 1999). The occurrence of *Cys* is significantly higher in LMPs and OtMPs than in GPs. *Glu* is a strong helix former (Chou and Fasman 1978) and this tendency influences the higher occurrence of it in GPs than LMPs and OtMPs.

On the other hand, the composition of the residues *Trp*, *Arg* and *Thr* is significantly higher in OtMPs than in GPs (Fig. 1). The structural analysis of several OtMPs shows that these residues play an important role in the stability

Fig. 1 Amino acid composition graph of LMP, GP and OtMP

Table 1 Amino acid frequencies for the 20 amino acid residues in LMPs, globular proteins and other membrane proteins

Amino acid residues	LMPs	Globular proteins (GPs)	Other membrane proteins (OtMPs)
Cys	2.364	1.607	2.323
Ile	5.756	5.4958	5.933
Met	2.174	2.371	2.166
Phe	5.4005	4.545	4.596
Trp	1.792	1.333	1.930
Tyr	3.606	3.557	3.615
Val	7.205	6.458	7.166
Leu	11.78	10.77	10.944
Arg	4.488	4.741	5.260
Asn	3.760	4.074	3.769
Asp	3.966	5.085	4.133
Glu	4.916	7.124	4.802
Gln	3.736	3.7815	3.793
His	2.228	2.327	2.139
Lys	3.710	5.400	4.0498
Ser	8.458	7.726	8.025
Thr	5.633	4.942	6.077
Pro	5.027	4.670	4.943
Ala	7.149	7.846	8.103
Gly	6.837	6.131	6.217

and function of OtMPs. We have also analyzed whether any patterns exist between amino acid compositional differences and their characteristics. The difference between the compositions of each amino acid residue in LMPs, GPs and OtMPs is plotted in Fig. 2. The residues with a negative difference indicate their high preference in LMPs and those with a positive value show their dominance in GPs and OtMP. We observed that *Phe*, *Leu*, *Gly*, *Asp*, *Glu* and *Thr* show significant differences between LMPs and GPs as well as between LMPs and OtMPs.

Among the hydrophobic residues, some shows a specific preference for LMPs like *Trp* and *Pro* with respect to OtMP. The presence of sulfur-containing residues *Cys* is higher in LMPs than in GPs. The polar residues have a higher occurrence in GPs than LMPs. The influence of charged residues is interesting; *Arg* has significantly higher occurrence in OtMPs than in GPs and LMPs while *Gln* and *Tyr* has no preference. On the other hand, *Lys* and *Glu* occur more often in GPs than in LMPs while an opposite trend is observed for *Gly*. The occurrence of *Pro* is less in GPs but higher in LMPs.

Dipeptide composition in GPs, OtMPs and LMPs

The dipeptide compositions for all possible 400 (20×20) pairs in LMPs, GPs and OtMPs have been computed using

Fig. 2 Differences in the Amino acid frequencies of LMPs, GPs and OtMPs

Eq. (2) and the dipeptide compositions of LMPs is presented in Table 2, while rest of matrices are available in supplementary material. In the Table S4 of supplementary material positive values indicate higher occurrence in LMPs than GPs and negative values show the preference in GPs. Figure 3 shows the higher value of dipeptide as white square, we observed that the occurrence of dipeptides, GQ, WD, LG, FT, WW, SH, TK, RD, RH, MD, WE and KK are significantly higher in LMPs. Interestingly, most of the dipeptides involve the residue *Ser*, which is one of the most favored residues in OtMPs (Gromiha and Suwa 2005) and plays an important role to the structure and stability (Pautsch and Schulz 2000; Zeth et al. 2000; Vandeputte-Rutten et al. 2001). On the other hand, the dipeptides, EE, HH, HP, RE and YP have higher occurrence in GP. It is noteworthy that most of these dipeptides involve the charged residues, Glu and His, which have significantly higher occurrence in GPs compared. Further, we have classified the residues into six groups, aliphatic (*Ala*, *Gly*, *Ile*, *Leu* and *Val*), aromatic (*Phe*, *Trp* and *Tyr*), sulfur containing (*Cys*, *Met*), polar (*Pro*, *Thr*, *Ser*, *Asn* and *Gln*), positively charged (*His*, *Lys* and *Arg*) and negatively charged (*Asp*, *Glu*) residues and analyzed the distribution. The distribution of amino acid residues in α -helical membrane proteins is entirely different due to the presence of the stretch of hydrophobic residues in this class of proteins. As expected that α -helical membrane proteins have higher occurrence of hydrophobic neighbors (AI, IF, IL, LI, LL, LV etc.) than LMPs and OtMPs.

We have analyzed the influence of amino acid residues (*Glu*, *Leu*, *Cys*, *Gln*, *Val* and *Ser*) that show significantly different abundance between LMPs and GPs, i.e., the residues which are rich in LMPs/GPs (Table 1; Fig. 1). *Leu*, *Ser* and *Thr* are significantly present in large amount in LMPs. Interestingly, only using these amino acid residues, the LMPs discriminate with the same accuracy as when using all the 20 amino acid residues (95 and 91 % in the dataset of well-annotated sequences and known three-dimensional structures, respectively). From Table 1, the

Table 2 Amino acid frequencies (First column) and dipeptide composition of Lysosomal membrane proteins

	–	Cys	Ile	Met	Phe	Trp	Tyr	Val	Leu	Arg	
Cys	2.3640	2.3199	1.4504	1.0499	1.1502	0.6633	1.1093	1.6129	1.7052	1.7450	
Ile	5.7563	2.0040	3.3112	1.5493	3.0515	1.5245	2.2342	3.1584	4.1895	2.5918	
Met	2.1743	0.7660	1.5909	0.9648	1.0563	0.9142	1.3219	1.7765	2.4090	1.1469	
Phe	5.4005	1.6674	3.6075	1.4085	3.4045	2.0958	2.3091	3.0811	4.5727	2.1668	
Trp	1.7921	0.7355	1.9216	0.9142	1.1458	1.3461	1.8097	1.2290	1.6622	1.4887	
Tyr	3.6065	1.2799	2.0453	1.5604	3.6567	1.3767	1.7409	2.8855	2.3367	2.4952	
Val	7.2053	2.1327	3.3573	1.9554	3.7848	1.4489	2.6582	4.4359	4.5389	2.1399	
Leu	11.784	2.2263	3.6427	1.5179	3.1880	1.4260	2.3717	4.4095	7.6881	2.5636	
Arg	4.4885	1.4388	2.2203	0.9580	2.3759	2.2999	1.7659	2.4834	2.5305	3.6556	
Asn	3.7609	1.3602	2.6420	1.1664	2.4567	1.0308	2.2983	2.2163	2.3231	1.4677	
Asp	3.9662	2.0499	3.2705	1.1323	2.4156	1.0304	1.7809	2.3607	1.9199	2.4104	
Glu	4.9162	1.2185	1.9544	1.1285	1.7313	1.1302	2.3173	2.4056	2.6128	2.8775	
Gln	3.7363	1.1839	1.7711	1.6174	1.7286	1.0787	2.0856	2.3692	2.2650	2.3757	
His	2.2285	1.4291	1.4938	0.8917	1.3042	1.2596	1.1761	1.4740	1.6532	1.9185	
Lys	3.7102	1.2285	2.4472	1.4616	1.4309	0.8496	1.9456	2.1305	2.4584	2.1237	
Ser	8.4586	1.5451	3.1266	1.5840	3.1139	1.4647	2.8441	3.8380	4.6147	2.9790	
Thr	5.6337	1.6413	2.5812	1.1054	3.4656	1.7515	2.1081	3.7278	3.4928	1.7942	
Pro	5.0271	1.2123	1.7926	0.9405	2.0808	0.9536	1.8119	2.7244	3.7383	3.3479	
Ala	7.1496	1.6223	3.1581	1.7228	3.3432	0.8747	2.4825	3.8978	4.3877	2.2120	
Gly	6.837	1.5830	3.6006	1.4001	3.1273	1.1356	1.9169	3.5236	5.0826	2.5033	
	Asn	Asp	Glu	Gln	His	Lys	Ser	Thr	Pro	Ala	Gly
Cys	2.6080	1.2593	1.2967	1.7194	1.4487	1.4900	1.7030	1.3115	1.6300	1.5152	2.5373
Ile	2.7490	1.5073	1.6652	1.9353	1.5951	2.3205	3.7022	2.8443	3.1850	2.8818	2.5392
Met	1.2169	1.3519	1.7075	1.0749	0.7419	1.5227	1.5896	1.4201	1.0612	2.3817	1.7094
Phe	2.1950	1.8013	2.0334	1.8368	1.5281	1.8190	3.0620	2.4743	2.1009	3.0805	2.4808
Trp	1.3870	1.0461	1.3134	0.7751	0.4695	1.5686	1.7366	1.5618	0.6460	1.3674	0.9585
Tyr	1.5295	1.5632	1.8179	2.6896	1.3918	1.6793	2.4268	2.1892	1.9334	1.8194	1.8624
Val	2.5634	2.3822	2.1955	2.0351	1.5566	2.0756	3.5759	3.4430	2.7734	3.4970	3.3272
Leu	2.1766	2.4089	2.7026	3.0104	1.8414	2.3752	5.2987	3.8645	3.7044	4.6362	4.9039
Arg	2.7101	2.3111	2.9412	2.4376	1.6240	2.4052	2.7082	2.3419	2.3306	2.4849	2.0799
Asn	2.8249	1.4234	2.2104	2.5623	1.1708	2.0417	2.7665	2.7402	2.0666	2.0519	2.2764
Asp	1.5979	2.5275	2.9387	1.7898	1.2917	2.5259	2.7931	2.5349	2.3360	1.8441	2.2859
Glu	2.2899	3.2896	4.9042	2.4384	1.9546	3.3560	2.3214	2.6674	2.1670	2.9831	2.5779
Gln	1.7308	2.2139	2.6497	2.4103	1.6429	2.3745	1.9512	2.1173	2.5956	2.6319	2.6530
His	2.7320	2.1673	1.1031	1.4520	1.1565	1.4836	1.9125	1.7191	1.4869	1.2207	1.2793
Lys	2.6755	2.9983	2.7202	2.0002	1.5643	2.6332	2.1549	2.2964	2.1300	2.0449	2.0487
Ser	2.0135	2.9258	2.6956	2.2683	2.0919	2.3692	5.4209	3.3005	3.4445	3.5483	4.1205
Thr	1.7290	2.2165	2.5055	2.0470	1.2807	2.0366	3.7854	2.9311	2.6059	3.5213	3.3596
Pro	1.8927	2.1927	3.1617	2.8863	1.6522	1.9928	3.0801	2.6453	3.1328	2.9682	2.9302
Ala	1.7442	2.5829	2.6701	2.18046	1.7512	1.8821	3.6539	2.6773	3.1405	5.0364	4.1312
Gly	2.2679	2.5022	2.8584	1.6014	1.4413	1.8043	3.6796	3.2058	2.5614	3.8569	4.0503

amino acid compositions of *Ser*, *Asn*, *Gln*, *Thr*, *Gly*, *Tyr*, *Ala*, *Arg* and *Leu* are higher in OtMPs than in GPs. These nine residues are used to discriminate OtMPs. As expected, these residues could successfully omit globular and membrane proteins (89 and 90.8 %, respectively). On the other

hand the residues, which have higher amino acid composition in LMPs than GPs, picked up the LMPs with high accuracy whereas the exclusion of GPs is poor. We have also examined the accuracy of discriminating OtMPs using the amino acid composition of each of the 20 amino acid

Fig. 3 Dipeptide composition of LMPs, GPs and OtMPs. On the X-axis and Y-axis there are 20 amino acids “CIMFWYVLRNDEQHKST-PAG”, so X_1Y_1 is the dipeptide CC, X_1Y_2 is the dipeptide CI, $X_{12}Y_{12}$ is the dipeptide EE and so on

Table 3 Discrimination results of Lysosomal membrane proteins

PDB ID	LMP-OtMP	LMP-GP	Assignment
<i>Lysosomal membrane proteins (L)</i>			
4DDJ	5.004992	−32.165	L
3ZR5	6.0587	11.0695	L
3GRO	−24.9499	−37.6438	–
3FGR	2.1133	21.9376	L
2VT0	−4.8720	11.2879	L
2DOB	5.4	−32.165	L
2AGC	16.97469	14.28982	L
2AG4	2.259317	−1.45062	L
1TJJ	2.259317	−1.45062	L
1SN6	47.28758	15.34255	L
1PUB	0.564446	15.34255	L
1PJA	1.6415	41.6140	L
1N69	−9.83254	12.1583	L
1M6D	−27.3957	70.31871	L
1LYB	1.7614	−0.0593	L
1EH5	−10.2901	−21.9565	–
PDB ID	GP-LMP	GP-OtMP	Assignment
<i>Globular proteins (G)</i>			
2WB0	71.4957	4.1204	G
1AGR	54.7	27.96	G
1BBT	47.49	61.52	G
1AKO	57.49	25.88	G
1AYL	−47.49	−47.49	–
1BDO	61.52	61.52	G
1BNC	57.49	57.49	G
1BPL	25.88	25.88	G
1AOR	15.86	15.86	G

Table 3 continued

PDB ID	GP-LMP	GP-OtMP	Assignment
1APY	−46.55	−46.55	–
1ATL	47.76	−47.76	G
1BIA	40.60	−40.60	G
1BAM	47.49	47.49	G
1BCO	−61.52	−61.52	–
1BKS	57.49	57.49	G
1BLE	−25.88	−25.88	–
PDB codes	OtMP-GP	OtMP-LMP	Assignment
<i>Other membrane proteins (O)</i>			
1QD6	71.35	61.52	O
1I78	−52.84	−57.49	–
1QFG	80.49	25.88	O
1FEP	119.54	15.86	O
1KMO	37.22	46.55	O
1NQW	76.22	47.76	O
1A0S	53.35	40.60	O
1K24	−88.99	−47.49	–
2MPR	53.46	61.52	O
1HXX	52.84	61.52	O
1E54	58.33	57.49	O
3PRN	−58.33	−25.88	–
2POR	125.45	80.49	O
7AHL	87.23	119.54	O
1EK9	67.87	37.22	O

residues. In this method, the composition of only one amino acid residue used at a time for discrimination and the calculation has been repeated for 20 times. Most of the

residues have a predictive ability of 40–65 %. *Leu* is the best predictor for LMPs and *Glu* is the best predictor of GPs. Further, we noticed that *Ala* has the ability of predicting OtMPs and excluding GPs at an accuracy level of about 79. Further for validation of our method we have used sequences of structured proteins from the PDB, the result is shown in Table 3.

Conclusion

We have systematically analyzed the amino acid sequences of LMPs, GPs, and OtMPs and developed the amino acid frequencies and dipeptide composition parameters for these classes of proteins. The similarities and differences of the 20×21 elements between LMPs, GPs and OtMPs have been calculated. This method correctly identified 95 % of the LMPs and excluded up to 88 % of the globular proteins. These accuracy levels are comparable to or better than other previous reported methods. This method could be effectively used to discriminate LMPs and for detecting LMPs in genomic sequences. The preferences of weighted sum of amino acid frequencies and dipeptide composition together between LMPs, GPs and OtMPs have been applied, and this result is further used to identify the LMPs and to exclude GPs and OtMPs. This method correctly identified 95 % of the LMPs and excluded up to 88 % of the globular proteins.

Acknowledgments VT and PT are thankful to University Grants Commission, New Delhi for a research fellowship. The work has been supported by a DBT-BIF Grant to DKG under its BTISNet scheme.

Conflict of interest Not declared.

References

- Ahmad S, Gromiha MM (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 18:819–824
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28:235–242
- Branden C, Tooze C (1999) Introduction to protein structure. Garland Publishing Inc., New York
- Cai YD, Ricardo PW, Jen CH, Chou V (2004) Application of SVM to predict membrane protein types. *J Theor Biol* 226:373–376
- Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol* 47:45–148
- Chou KC, Shen HB (2007) MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun* 360:339–345
- Chou KC, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. *Nat Sci* 2009(2):63–92
- Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 34:508–519
- Gromiha MM, Suwa M (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics* 21(7):961–968
- Gromiha MM, Ahmad S, Suwa M (2005) Application of residue distribution along the sequence for discriminating outer membrane proteins. *Comput Biol Chem* 29:135–142
- Liu H, Wang M, Chou KC (2005) Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem Biophys Res Commun* 336:737–739
- Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: orientations of Proteins in Membranes database. *Bioinformatics* 22:623–625
- Pautsch A, Schulz GE (2000) High-resolution structure of the OmpA membrane domain. *J Mol Biol* 298:273–282
- Sonnhammer ELL, Heijne GV and Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In: *Proceedings of Sixth International Conference on Intelligent Systems for Molecular Biology*, AAAI/MIT Press: Menlo Park, CA, 6, 175–182
- Tripathi V, Gupta D (2014) Discriminating lysosomal membrane protein types using dynamic neural network. *J Biomol Struct Dyn* 32(10):1575–1582
- Vandeputte-Rutten L, Kramer RA, Kroon J, Dekker N, Egmond MR, Gros P (2001) Crystal structure of the outer membrane protease OmpT from *Escherichia coli* suggests a novel catalytic site. *EMBO J* 20:5033–5039
- Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. *Protein Eng Des Sel* 17:509–516
- Wang SQ, Yang J, Chou KC (2006) Using stacking generalization to predict membrane protein types based on pseudo-amino acid. *J Theor Biol* 242:941–946
- Zeth K, Diederichs K, Welte W, Engelhardt H (2000) Crystal structure of Omp32, the anion-selective porin from *Comamonas acidovorans*, in complex with a periplasmic peptide at 2.1 Å resolution. *Structure* 8:981–992