Amino Acids (2010) 39:777–783
DOI 10.1007/s00726-010-0520-8

ORIGINAL ARTICLE

Identification of functionally diverse lipocalin proteins from sequence information using support vector machine

Ganesan Pugalenthi · Krishna Kumar Kandaswamy · P. N. Suganthan · G. Archunan · R. Sowdhamini

Received: 4 May 2009 / Accepted: 6 February 2010 / Published online: 26 February 2010

Abstract Lipocalins are functionally diverse proteins that are composed of 120–180 amino acid residues. Members of this family have several important biological functions including ligand transport, cryptic coloration, sensory transduction, endonuclease activity, stress response activity in plants, odorant binding, prostaglandin biosynthesis, cellular homeostasis regulation, immunity, immunotherapy and so on. Identification of lipocalins from protein sequence is more challenging due to the poor sequence identity which often falls below the twilight zone. So far, no specific method has been reported to identify lipocalins from primary sequence. In this paper, we report a support vector machine (SVM) approach to predict lipocalins from protein sequence using sequence-derived properties. LipoPred was trained using a dataset consisting of 325 lipocalin proteins and 325 non-lipocalin proteins, and evaluated by an independent set of 140 lipocalin proteins and 21,447 non-lipocalin proteins. LipoPred achieved 88.61% accuracy with 89.26% sensitivity, 85.27% specificity and 0.74 Matthew's correlation coefficient (MCC). When applied on the test dataset, LipoPred achieved 84.25% accuracy with 88.57% sensitivity, 84.22% specificity and MCC of 0.16. LipoPred achieved better performance rate when compared with PSI-BLAST, HMM and SVM-Prot methods. Out of 218 lipocalins, LipoPred correctly predicted 194 proteins including 39 lipocalins that are non-homologous to any protein in the SWISS-PROT database. This result shows that LipoPred is potentially useful for predicting the lipocalin proteins that have no sequence homologs in the sequence databases. Further, successful prediction of nine hypothetical lipocalin proteins and five new members of lipocalin family prove that LipoPred can be efficiently used to identify and annotate the new lipocalin proteins from sequence databases. The LipoPred software and dataset are available at http://www3.ntu.edu.sg/home/EPNSugan/index_files/lipopred.htm.

Keywords Lipocalin · Diverse function · Odorant binding · Support vector machine · Ligand binding · Allergenic proteins · Salivary proteins

G. Pugalenthi · P. N. Suganthan (✉)
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
e-mail: EPNSugan@ntu.edu.sg

K. K. Kandaswamy
Institute for Neuro- and Bioinformatics, University of Lübeck, 23538 Lübeck, Germany

K. K. Kandaswamy
Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, 23538 Lübeck, Germany

G. Archunan
Center for Pheromone Technology, Department of Animal Science, Bharathidasan University, Tiruchirappalli 620024, Tamil Nadu, India

R. Sowdhamini (✉)
National Centre for Biological Sciences, UAS-GKVK Campus, Bellary Road, Bangalore 560065, India
e-mail: mini@ncbs.res.in

Introduction

Lipocalins are functionally diverse family of small ligand binding proteins with 120–180 amino acid residues (Flower 1996). Basically, all lipocalins share a fundamental structure of an eight stranded anti-parallel beta barrel, closed on one side by an N-terminal 3-10 helix and flanked on one

side of barrel by C-terminal alpha-helix. Consequently, they form a characteristic glove shape, encompassing a hydrophobic ligand binding pocket (Flower 1996; Flower et al. 1993, 2000). In addition, they share three structurally conserved regions (SCRs). On the basis of SCRs, lipocalins are grouped into kernel lipocalin that contains all the three SCRs and outlier lipocalins containing one or two SCRs (Flower et al. 1993, 2000).

Lipocalins are widely distributed in nature and are found in a variety of organisms such as animals, plants, insects and bacteria (Flower 1996; Flower et al. 1993, 2000; Ganfornina et al. 2000; Grzyb et al. 2006; Frenette Charron et al. 2002; Bishop 2000). In general, lipocalins act as a carrier of small hydrophobic ligands such as retinol, odorant molecules, pheromones, fatty acids, cholesterol, etc. (Flower 1996). In addition to the binding and transport, lipocalins also play many other important physiological functions. For example, several lipocalins have been described as allergenic proteins (Mantyjarvi et al. 2000). Lipocalins are also thought to play a role in chemo-signaling between animals to coordinate social behavior and also function as pheromone stabilizers in the environment, providing a slow release mechanism that extends the effective potency of these volatile molecules in male urine scent marks (Logan et al. 2008). Nitrophorin, a salivary lipocalin from *Rhodnius prolixus*, serves as a vasodilator and platelet aggregation inhibitor (Ribeiro et al. 1993). Some lipocalins act as enzymes (Hieber et al. 2000; Fouchécourt et al. 2002; Yusifov et al. 2000). Few well-known lipocalin enzymes are prostaglandin D2 synthase that is involved in the formation of prostaglandin D2 and violaxanthin de-epoxidases that is involved in the xanthophylls cycle (Hieber et al. 2000; Fouchécourt et al. 2002). In addition, endonuclease activity is reported in tear lipocalins and some forms of the beta-lactoglobulin (Yusifov et al. 2000). Lipocalins have been widely used as biochemical markers of diseases such as inflammatory disease, cancer, lipid disorders, liver and kidney function-related disorders (Xu and Venge 2000). Some of the lipocalins have been used as diagnostic indicators (Devarajan 2007; Xu and Venge 2000). Due to the diverse ligand binding specificities, lipocalins have been exploited for the development of immunotherapies (Schlehuber and Skerra 2005).

Over the past three decades, lipocalin proteins have received much attention due to their remarkable functional diversity and multiple molecular recognition properties (Xu and Venge 2000; Schlehuber and Skerra 2005; Akerstrom et al. 2000). Several efforts have been made to identify new lipocalin members. As of now, prediction of lipocalins is primarily based on the sequence similarity search methods such as PSI-BLAST, HMM, etc. (Altschul et al. 1997; Eddy 1998). Although lipocalins have a well-conserved tertiary structure, their sequence identity is very low and in some cases falls within the twilight zone (Flower et al. 1993; Adam et al. 2008). Assigning sequences to the lipocalin family by sequence search methods is risky, when the pairwise identity is below 20%. With the rapid increase in the newly found protein sequences entering into databanks, an efficient method is needed to identify lipocalins from the sequence databases. Encouraged by the overwhelming success of machine learning methods in an engineering, medical and financial applications, many research groups have been using neural networks (NN), support vector machines (SVMs) and other machine learning algorithms in the biological field especially in the classification and prediction of protein structure and functional profile (Cai et al. 2003; Chou and Shen 2009; Jensen et al. 2003; Tang et al. 2009). So far, SVM and other statistical learning methods have not been explored for the prediction of lipocalins. In this paper, we report a SVM approach to identify lipocalins from sequence information, irrespective of the sequence similarity.

Materials and methods

Dataset

We obtained 196 lipocalin domains from the seed proteins of the Pfam database (Sonnhammer et al. 1997). To enrich the dataset, we performed PSI-BLAST search against non-redundant sequence database with stringent threshold (*E* value 0.01) (Altschul et al. 1997). Redundant sequences that have $\geq$40% sequence similarity were removed from the dataset using CD-HIT (Li et al. 2001). After careful manual examination, a total of 465 lipocalins were selected for the positive dataset.

Training set

325 lipocalin domains were selected from the 465 lipocalin domains for the positive dataset. 325 non-lipocalin domains for the negative set were randomly taken from the seed proteins of the Pfam protein families, which are unrelated to lipocalin (Sonnhammer et al. 1997).

Test set

The remaining 140 lipocalin domains were considered as a positive dataset for testing. The negative dataset was created from the seed proteins of non-lipocalin domains which are selected from the seed proteins of non-lipocalin Pfam protein families (Sonnhammer et al. 1997). The negative sequences which are present in the training dataset were removed. Further, non-lipocalin domains with less

than 40 amino acids were excluded from the negative set. Finally, the test dataset consists of 140 lipocalin domains and 21,447 non-lipocalin domains.

Comparison set

In order to compare LipoPred with other sequence-based methods such as PSI-BLAST and HMM, we created a comparison dataset containing 218 lipocalin proteins that do not overlap with the training and testing dataset.

Features

In this work, each sequence is encoded by 119 features (Table 1) (please see the help file in LipoPred software package). We categorized 20 amino acids into 10 functional groups based on the presence of side chain chemical groups such as phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido (Q/N), hydroxyl (S/T) and non-polar (A/G/I/L/V/P) (Pugalenthi et al. 2008). Further, we categorized 20 amino acids into hydrophobic, hydrophilic and neutral amino acids.

Frequency of amino acids and groups

The frequency of 20 amino acids (number of occurrences of amino acid "X" divided by length of the protein), frequency of 10 amino acid groups and frequency of hydrophobic, hydrophilic and neutral amino acids were computed for each sequence. We utilized tripeptide information for the classification. In general, a large number of tripeptides can be obtained from all possible combinations of 20 amino acids. To reduce the feature dimension, we derived 27 tripeptides from all possible combinations of three amino acid groups (hydrophobic, hydrophilic and neutral amino acid groups). The frequency of 27 tripeptides was calculated for each sequence (please see the help file in LipoPred software package).

Table 1 List of 119 features

Features	No of features
Frequency of 20 amino acids	20
Frequency of 10 functional group	10
Frequency of hydrophobic, hydrophilic and neutral amino acids	3
Overall composition of helix (H), stand (E) and coil (C)	3
Frequency of 10 functional groups in helix, strand and coil regions	30
Frequency of hydrophobic, hydrophilic and neutral amino acids in helix, strand and coil regions	9
Frequency of 27 tripeptides	27
Frequency of hydrophobic, hydrophilic and neutral amino acid-rich short peptides	3
Physicochemical properties	14
Total	119

Frequency of short peptides

We incorporated the frequency of short peptides (10 residue length, in this case) which are rich in hydrophobic or hydrophilic or neutral amino acids. For example, if a short peptide with more than six hydrophobic residues, then we consider this peptide as hydrophobic-rich short peptide. Similarly, we calculated hydrophilic, neutral-rich short peptides. The frequency of hydrophobic-rich peptides (contains at least 6 hydrophobic amino acids) or hydrophilic-rich peptides (contains at least 6 hydrophilic amino acids) or neutral amino acid-rich peptides (contains at least 6 neutral amino acids) was calculated for each sequence.

Content of secondary structural element (SSE)

Secondary structural element (helix: H, strand: E and coil: C) information was assigned to all the sequences in the alignment using a secondary structure prediction program, PSIPRED (McGuffin et al. 2000). The overall composition of helix (H), beta-sheet (E), coil (C) and the frequencies of 10 amino acid groups, hydrophobic, hydrophilic and neutral amino acids at helix, sheet, and coil regions were calculated.

Physicochemical properties

We obtained 14 physicochemical properties from UMBC AAIndex database (Kawashima et al. 1999). The computed physicochemical properties include molecular weight, hydrophobicity, hydrophilicity, refractivity, average accessible surface area, flexibility, melting point, side chain volume, side chain hydrophobicity, normalized frequency of beta-sheet and alpha-helix, polarity, heat capacity, and isoelectric points.

Classification protocol

In this study, we used a SVM to identify lipocalin proteins. The SVM and its applications to bioinformatics are extensively described in the literature (Pugalenthi et al. 2008; Cortes and Vapnik 1995; Burges 1998; Cai et al. 2002; Chou 2001, 2005; Chou and Cai 2005; Muller et al. 2001). Thus, only a brief description is given here. A linear SVM constructs a hyperplane that separates two different classes of feature vectors with a maximum margin. One class represents lipocalins and the other

non-lipocalins. For the cases where the data are linearly non-separable, the SVM uses non-linear kernel functions to transform the input data into a high-dimensional space, while allowing us to solve the problem in the input space itself. Maximization of the margin can be reduced to a convex quadratic optimization problem, which has a unique global minimum (Muller et al. 2001). For our case of binary classification problem, a training dataset $\{\boldsymbol{x}_i, y_i\}$, $i = 1, 2, \ldots, L$, where $\boldsymbol{x}_i \in \Re^d$ is the input vector consisting of d features representing the ith sample, and $y_i \in (+1, -1)$ denotes the class label. The SVM discriminant function is given by

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x) + b \tag{1}$$

where m ($<L$) is the number of input patterns having non-zero values of the Lagrangian multipliers α_i obtained by solving a quadratic optimization problem, $K(\boldsymbol{x}_i, \boldsymbol{x})$ is the kernel matrix and b is the bias term. Simulations were preformed using LIBSVM version 2.81 (Chang and Lin 2001). The SVM training was performed on the training dataset by tuning the values of regularization parameter and RBF kernel parameter using grid search (C: 0.1–10000, gamma: 0.0001–10). The final SVM model was developed on optimized regularization parameter C and gamma value of RBF kernel function.

Feature selection

In this work, we used feature selection to remove possible redundant features from the original feature set. By redundant, we mean that the feature has negligible influence on the final classification performance. In this method, we utilized information gain algorithm with the ranker method to identify the important features that distinguish positive and negative classes (Mitchell 1997). This method was implemented using Weka 3.5 (Frank et al. 2004). The information gain for each feature was calculated and the features were ranked according to this measure. Feature selection was performed by fivefold cross-validation on the training dataset.

Performance measures

To measure the performance, we used four different parameters derived from the four scalar values: TP (true positives: number of correctly classified lipocalins), TN (true negatives: number of correctly classified non-lipocalins), FP (false positives: number of non-lipocalins incorrectly classified as lipocalins) and FN (false negatives: number of lipocalins incorrectly classified as non-lipocalins). Using the following formulas, we calculated accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC):

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{4}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{5}$$

Results and discussion

Classification result

In this work, an SVM method is reported for the prediction of lipocalins from the protein sequence using 119 sequence-derived properties. We trained our SVM classifier on the training dataset containing 325 lipocalin domains and 325 non-lipocalin domains. To verify the training performance, we carried out fivefold cross-validation. LipoPred method achieved 88.61% accuracy with 89.26% sensitivity and 85.27% specificity and MCC of 0.74 from fivefold cross-validation. The performance of LipoPred was evaluated using the test dataset containing 140 lipocalin and 21,447 non-lipocalin domains. As shown in Table 2, LipoPred achieved 84.25% accuracy with 88.57% sensitivity, 84.22% specificity and MCC of 0.16.

Since LipoPred classifier was trained using the domain regions, it is important to test the performance of the classifier on full length sequences that contain lipocalin domains and/or non-lipocalin domains. We created a dataset containing full length sequence for each domain in the testing dataset that consists of 140 lipocalin domains and 21,447 non-lipocalin domains. Using all features, LipoPred classifier achieved 82.50% accuracy with 82.14%

Table 2 Classification result achieved on test dataset (test-domain) containing 140 lipocalin and 21,447 non-lipocalin domains

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
10	89.29	77.10	0.13	77.18
25	89.29	82.17	0.15	82.22
50	88.57	81.59	0.14	81.63
75	87.14	84.35	0.16	84.37
100	88.57	82.85	0.15	82.89
119	88.57	84.22	0.16	84.25

MCC Matthew's correlation coefficient

sensitivity, 82.86% specificity and MCC of 0.65. This result indicates that LipoPred can be used to predict lipocalin proteins from the full length protein sequences that contain lipocalin and/or non-lipocalin domains.

In order to identify the most prominent features that distinguish lipocalins from non-lipocalins, we applied a feature reduction protocol to eliminate the redundant features. As shown in Table 2, LipoPred achieved an accuracy of 84.37% with 87.14% sensitivity, 84.35% specificity and MCC of 0.16 and using 75 features on the test dataset. When the features were further reduced from 75 to 10, we observed a significant drop in the specificities and accuracies although there is a slight increase in the sensitivities.

Prediction of lipocalins

To test the capability, LipoPred was evaluated by an independent dataset obtained from the NCBI database using keyword search. The sequences that are present in the positive training dataset were removed from the list. After careful manual inspection, 218 odorant binding proteins were selected for the testing. The efficiency of LipoPred in the identification of lipocalins was compared with the other methods such as HMM, PSI-BLAST and SVM-Prot (Altschul et al. 1997; Eddy 1998, Cai et al. 2003) (Table 3). BLAST search for each sequence was carried out against the database of 465 lipocalins that were used for the training and testing. HMM search for each query sequence was performed against the HMM profile obtained from Pfam database (Pfam release 23) (Sonnhammer et al. 1997). BLAST and HMM searches were carried out using four different *E* value thresholds (*E* value 0.001, 0.01, 0.1 and 1). SVM-Prot search was performed using online SVM-Prot webserver (Cai et al. 2003. Since SVM-Prot has no separate class for lipocalin family, only those proteins which are predicted as "All-lipid binding proteins" and "Lipid binding" classes by SVM-Prot are considered as correct prediction. Out of 218 proteins, our approach correctly predicted 194 proteins whereas PSI-BLAST (*E* value 1), HMM (*E* value 1) and SVM-Prot methods predicted 183, 174 and 106 proteins, respectively. Further analysis of the 218 lipocalins shows that 43 proteins have no single homologous protein in the SWISSPROT (Bairoch and Apweiler 2000) database according to PSI-BLAST search results. Our method correctly predicted 39 out of 43 proteins. This result shows the capability of our prediction system for recognizing novel lipocalins that are non-homologous to other proteins.

Table 3 Prediction result of 218 lipocalins by LipoPred, HMM, PSI-BLAST and SVM-Prot methods

Methods	Number of proteins predicted as lipocalin
LipoPred	194
HMM (*E* value 0.001)	126
HMM (*E* value 0.01)	143
HMM (*E* value 0.1)	163
HMM (*E* value 1)	174
PSI-BLAST (*E* value 0.001)	157
PSI-BLAST (*E* value 0.01)	162
PSI-BLAST (*E* value 0.1)	176
PSI-BLAST (*E* value 1)	183
SVM-Prot	106

Prediction of new lipocalin members and hypothetical protein annotation by LipoPred

Recently, Williford et al. (2004) reported water-soluble proteins, a major component of nutritive "Milk" in the cockroach, *Diploptera punctata*. These proteins do not have significant sequence similarity with lipocalin proteins. Due to the presence of structurally conserved lipocalin signatures, these proteins were proposed as lipocalins. Five cockroach milk proteins (GI accession code: 38455985, 38455983, 38455949, 38455947 and 38455945) were obtained from the NCBI sequence database. No hits were found for these sequences from PSI-BLAST search against non-redundant sequence database. Similarly, HMM search against Pfam family profiles returned no hits with significant match to the lipocalin families. We applied LipoPred to identify whether these proteins belong to the lipocalin family or not. Surprisingly, all the sequences were predicted as lipocalin by LipoPred. This result shows the usefulness of LipoPred in the identification of new lipocalin members that have negligible sequence similarity with other lipocalin members.

To test the efficiency of LipoPred, we selected nine hypothetical proteins containing lipocalin family-specific signature (GxW) at the N-terminal region (Flower 1996; Flower et al. 1993, 2000). LipoPred predicted all the nine proteins as lipocalins (Table 4). Out of nine proteins, three proteins (GI code: 6518066, 67083284 and 109472117) were predicted as lipocalins by HMM, PSI-BLAST and SVM-Prot methods (Altschul et al. 1997; Eddy 1998; Cai et al. 2003). Four hypothetical proteins (GI code: 17532811, 84997868, 12847113, 94377397 and 74206886) are annotated as lipocalin in INTERPRO database (Hunter et al. 2009) and one protein (GI code: 99078160) is annotated as lipocalin in KEGG database (Kanehisa and Goto 2000). This result shows that LipoPred method is useful to identify new lipocalin members which are difficult to annotate using sequence similarity.

Table 4 Prediction result for nine lipocalins by LipoPred, HMM, PSI-BLAST and SVM-Prot

GI code	GxW signature	Source of annotation	LipoPred	HMM	PSI-BLAST	SVM-Prot
6518066	Yes	–	Lipocalin	Lipocalin	Lipocalin	Lipocalin
67083284	Yes	–	Lipocalin	Lipocalin	–	–
109472117	Yes	–	Lipocalin	–	Lipocalin	–
94377397	Yes	Interpro	Lipocalin	–	–	–
17532811	Yes	Interpro	Lipocalin	–	–	–
12847113	Yes	Interpro	Lipocalin	–	–	–
99078160	Yes	KEGG	Lipocalin	–	–	–
74206886	Yes	Interpro	Lipocalin	–	–	–
84496314	Yes	Interpro	Lipocalin	–	–	–

Table 5 List of top ten selected features

Features
Frequency of cysteine
Frequency of strand
Frequency of helix
Frequency of cysteine in helix
Frequency of hydrophilic residues (RKNDEP)
Frequency of amino acids containing non-polar side chain (AGILVP)
Frequency of tryptophan (W)
Frequency of histidine (H) in helix
Frequency of cysteine in strand
Frequency of amino acids containing phenyl group [FWY]

Overview of the selected features

We analyzed the top ten features to understand their role in the classification. Feature selection protocol was employed to reduce the features from 119 to 10 (Table 2). The list of top ten features and their names are given in Table 5.

Lipocalin proteins have a well-conserved tertiary structure containing anti-parallel beta barrel and an alpha-helix (Flower 1996; Flower et al. 1993, 2000). The use of secondary structure information can help us to improve the prediction performance. As expected, frequencies of helix and strand were selected in the top ten features. Another important feature of lipocalin family proteins is the presence of a conserved sequence signature "GxW" (Flower 1996, 1993, 2000). The conserved tryptophan (W) in the GxW motif has been shown to play a role in maintaining structure, stability and ligand affinity (Gasymov et al. 1999). The selected features include two such related features namely, frequency of Tryptophan (W) and frequency of amino acids containing phenyl groups (FWY).

It has been reported that most lipocalin proteins are presumed to have a one or two conserved disulfide bonds (Flower 1996; Flower et al. 1993, 2000; Glasgow et al. 1998; Yang et al. 1994). It should be noted that LipoPred selected features such as frequency of cysteine and frequency of cysteine in secondary structures. This brief analysis shows that the selected features play a significant role in the classification. However, the detailed analyses of all the features are required to provide more information about their role in the classification.

Conclusion

Identification of lipocalin members from sequence databases is difficult due to poor sequence similarity. We reported an SVM-based method, LipoPred, for the prediction of lipocalins from sequence using sequence-derived properties. Very high prediction accuracies on the training and testing datasets show that LipoPred is potentially useful tool for the prediction of lipocalins from protein primary sequence. Further, the capability of our method was tested by an independent dataset consisting of 218 members and this method correctly predicted 194 lipocalin proteins. LipoPred is shown to be useful in the identification of new lipocalin members that have no sequence homologs in the sequence database. LipoPred is also useful for identifying lipocalins from the full length sequence containing lipocalin and non-lipocalin domains. The LipoPred program and dataset are available at http://www3.ntu.edu.sg/home/EPNSugan/index_files/lipopred.htm.

Acknowledgments GP and PNS acknowledge the financial support offered by the A*Star (Agency for Science, Technology and Research). RS acknowledges the support provided by the National Center for Biological Sciences (NCBS). KKK acknowledges the support by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]. KKK acknowledges Mr. Rajeev Gangal, CEO, Insilico division, Systems Biology India Pvt Ltd, Maharashtra, India and Prof. Thomas Martinetz and Dr. Stefen Moller, Institute for Neuro- and Bioinformatics, University of Luebeck, Germany for their support.

References

Adam B, Charloteaux B, Beaufays J, Vanhamme L, Godfroid E, Brasseur R, Lins L (2008) Distantly related lipocalins share two conserved clusters of hydrophobic residues: use in homology modeling. BMC Struct Biol 8:1

Akerstrom B, Flower DR, Salier JP (2000) Lipocalins: unity in diversity. Biochim Biophys Acta 1482:1–8

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database, its supplement TrEMBL in 2000. Nucleic Acids Res 28(1):45–48

Bishop RE (2000) The bacterial lipocalins. Biochim Biophys Acta 1482:73–83

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167

Cai YD, Liu XJ, Xu XP, Chou KC (2002) Prediction of protein structural classes by support vector machines. Comput Chem 26:293–296

Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31(13):3692–3697

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/_cjlin/libsvm

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet 43:246–255

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19

Chou KC, Cai YD (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inform Model 45:407–413

Chou KC, Shen HB (2009) Recent advances in developing web-servers for predicting protein attributes. Nat Sci 1:63–92

Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20:273–297

Devarajan P (2007) Neutrophil gelatinase-associated lipocalin: new paths for an old shuttle. Cancer Ther 5(B):463–470

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763

Flower DR (1996) The lipocalin protein family: structure and function. Biochem J 318:1–14

Flower DR, North AC, Attwood TK (1993) Structure and sequence relationships in the lipocalins and related proteins. Protein Sci 2:753–761

Flower DR, North AC, Sansom CE (2000) The lipocalin protein family: structural and sequence overview. Biochim Biophys Acta 1482:9–24

Fouchécourt S, Charpigny G, Reinaud P, Dumont P, Dacheux JL (2002) Mammalian lipocalin-type prostaglandin D2 synthase in the fluids of the male genital tract: putative biochemical and physiological functions. Biol Reprod 66:458–467

Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20:2479–2481

Frenette Charron JB, Breton G, Badawi M, Sarhan F (2002) Molecular and structural analyses of a novel temperature stress-induced lipocalin from wheat and Arabidopsis. FEBS Lett 517:129–132

Ganfornina MD, Gutiérrez G, Bastiani M, Diego S (2000) A phylogenetic analysis of the lipocalin protein family. Mol Biol Evol 17:114–126

Gasymov OK, Abduragimov AR, Yusifov TN, Glasgow BJ (1999) Binding studies of tear lipocalin: the role of the conserved tryptophan in maintaining structure, stability and ligand affinity. Biochim Biophys Acta 1433:307–320

Glasgow BJ, Abduragimov AR, Yusifov TN, Gasymov OK, Horwitz J, Hubbell WL, Faull KF (1998) A conserved disulfide motif in human tear lipocalins influences ligand binding. Biochemistry 37:2215–3325

Grzyb J, Latowski D, Strzalka K (2006) Lipocalins—a family portrait. J Plant Physiol 163:895–915

Hieber AD, Bugos RC, Yamamoto HY (2000) Plant lipocalins: violaxanthin de-epoxidase and zeaxanthin epoxidase. Biochim Biophys Acta 1482:84–91

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(Database Issue):224–228

Jensen LJ, Gupta R, Staerfeldt HH, Brunak S (2003) Prediction of human protein function according to gene ontology categories. Bioinformatics 19(5):635–642

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Kawashima S, Ogata H, Kanehisa M (1999) AAindex: amino acid index database. Nucleic Acids Res 27:368–369

Li W, Jaroszewski L, Odzik GA (2001) Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics 17:282–283

Logan DW, Marton TF, Stowers L (2008) Species specificity in major urinary proteins by parallel evolution. PLoS ONE 3(9):e3280

Mantyjarvi R, Rautiainen J, Virtanen T (2000) Lipocalins as allergens. Biochim Biophys Acta 1482:308–317

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16(4):404–405

Mitchell TM (1997) Machine learning. McGraw-Hill, New York

Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 2:181–201

Pugalenthi G, Kumar KK, Suganthan PN, Gangal R (2008) Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. Biochem Biophys Res Commun 367:630–634

Ribeiro JM, Hazzard JM, Nussenzveig RH, Champagne DE, Walker FA (1993) Reversible binding of nitric oxide by a salivary heme protein from a bloodsucking insect. Science 260:539–541

Schlehuber S, Skerra A (2005) Lipocalins in drug discovery: from natural ligand-binding proteins to anticalins. Drug Discov Today 10:23–33

Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28(3):405–420

Tang K, Pugalenthi G, Suganthan PN, Lanczycki CJ, Chakrabarti S (2009) Prediction of functionally important sites from protein sequences using sparse kernel least squares classifiers. Biochem Biophys Res Commun 384(2):155–159

Williford A, Stay B, Bhattacharya D (2004) Evolution of a novel function: nutritive milk in the viviparous cockroach, *Diploptera punctata*. Evol Dev 6:67–77

Xu S, Venge P (2000) Lipocalins as biochemical markers of disease. Biochim Biophys Acta 1482:298–307

Yang CY, Gu ZW, Blanco-Vaca F, Gaskell SJ, Yang M, Massey JB, Gotto AM, Pownall HJ (1994) Structure of human apolipoprotein D: locations of the intermolecular and intramolecular disulfide links. Biochemistry 33:12451–12455

Yusifov TN, Abduragimov AR, Gasymov OK, Glasgow BJ (2000) Endonuclease activity in lipocalins. J Biochem 347:815–819