Chapter 4

Identification and analysis of odorant bonding proteins
Chapter 4

Identification and analysis of odorant bonding proteins

4.1 Introduction

Olfaction is an important process to establish behavioural response and involves the binding of small, hydrophobic, volatile molecules to receptors of the nasal neuroepithelia [Buck and Axel 1991]. The olfaction mechanism has been well studied and is generally similar in vertebrates, insects, crustaceans, and nematodes [Ache, 1994; Pelosi, 1996; Hildebrand and Shepherd, 1997].

The first step in olfaction is the solubilization of the hydrophobic odorants in the hydrophilic nasal mucus. Odorant binding proteins (OBPs) are believed to shuttle odorants from the environment to the underlying odorant receptors, for which they could potentially serve as odorant presenters. Previous reports have shown that OBPs are present in large number within a species [Felicioli et al., 1993]. This indicates that OBPs do play an active role in odorant recognition rather than merely serving as passive odorant shuttles [Raming et al., 1990]. Several reports have demonstrated selective binding of odorants to different OBPs derived from a given species [Du and Prestwich, 1995]. OBPs are also suspected to participate in the deactivation of odorants and signal termination [Graham et al., 2001]. Presence of OBPs in non-sensory tissues of insect suggests their non-sensory roles [Kodrik et al., 1995].

OBPs are generally divergent across species and within the same species, with percent of conserved residues as low in some cases as 8% [Dear et al., 1991; Pes et al., 1998]. Although many efforts have been made to study the role of OBPs, their physiological function is still unclear and more sequence data are required for the complete understanding of the odorant transport mechanism. In this work, 3 fold search is employed to identify more OBP members from genome databases. It has been shown in previous work that 3 fold sequence approach can be efficiently employed to identify distant homologous sequences [please see Chapter 2] [Bhaduri et al., 2004c]. In addition, sequence search is performed with multiple structural motifs using
Identification and analysis of odorant bonding proteins

SCANNMOT algorithm [Chakrabarti et al., 2005]. Further, a machine learning method is reported to identify and validate odorant binding proteins from sequence information.

4.2 Types of OBPs

OBPs are small soluble polypeptides, which are thought to act as a carrier for odorants and carries odorant from the environment to the nasal epithelium in vertebrates and sensillar lymph in insects [Vogt and Riddiford, 1981; Pelosi, 1994]. Vertebrate OBPs are members of large family lipocalin and shares eight stranded beta barrel [Baincht et al., 1996]. The OBPs of Lipocalin family includes aphrodisin, odorant-binding protein (OBP), major urinary protein (MUP), alpha-2u-globulin, Lipocalin allergens, probasin (PB) and salivary proteins [Pelosi, 1994, Tegoni et al., 2000; Spinelli et al., 2002]

Insect OBP family is composed of pheromone binding proteins (PBP), which are male-specific and associate with pheromone-sensitive neurons and general-odorant binding proteins (GOBP). Insect OBPs are completely different from their vertebrate counterpart both in sequence and three dimensional folding [Pelosi and Maida, 1995] and contains alpha helical barrel and six highly conserved cysteines [Vogt et al., 1999]. Another class of putative OBPs, named chemosensory proteins (CSPs) has been reported in different Orders of insects, including Lepidoptera [Vogt et al., 1989; Danty et al., 1998; Jacquin-Joly et al., 2001; Wanner et al., 2004]. These polypeptides of about 12 kDa do not exhibit significant homology to PBPs and GOBPs and contain four conserved cysteine residues all involved in intramolecular disulphide bridges. It shares features common to vertebrate odorant-binding proteins, but has a primary structure unlike odorant-binding proteins [McKenna et al., 1994].

4.3 Structural aspects of OBPs

Vertebrate OBPs belong to lipocalin superfamiliy and have an eight stranded anti parallel beta-barrel with a repeated + 1 topology [Cowan et al., 1990; Flower et al., 1993]. The beta-barrel encloses a ligand binding site composed of both an internal cavity and an external loop scaffold. The diversity of cavity and scaffold gives rise to a variety of different binding modes each capable of accommodating ligands of different size, shape, and chemical character. Lipocalin OBPs have two conserved cysteines that form the disulphide bridge between the C-terminal tail and the beta barrel. The
disulfide bridge is conserved in all the lipocalin sequences identified so far, from bacteria to mammals [Flower, 1996], with the exception of bovine OBP. Bovine OBP, an unusual lipocalin, is a domain-swapped dimer, in which the helix near the C-terminal region of each monomer packed against the beta-barrel of the other [Figure 4.1b]. The absence of disulfide bond is very likely mandatory for the domain swapping to occur [Tegoni et al., 1996].

Figure 4.1: Three dimensional structure of vertebrate odorant binding proteins. a) 3d structure of boar salivary lipocalin (SAL), a pheromone binding protein specifically expressed in the sub maxillary glands of the boar (pdbcode 1gm6) [Loebel et al., 2000]. b) 3d structure of domain swapped odorant binding protein from bovine nasal mucosa [pdbcode 1obp] [Bianchet et al., 19996]

In contrast to the vertebrate PBP, the overall fold of insect OBP is roughly conical, with four anti parallel helices converging to a point and enclosing the hydrophobic pocket [Figure 4.2]. The hydrophobic pocket is flat and roughly triangular in shape. The residues lining the pocket are nearly all hydrophobic. Insect OBPs have been suggested to possess intrinsic conformational flexibility that would allow them to modify their conformation on ligand binding [Tegoni et al., 2004].
4.4 Analysis of OBP primary sequences

Lipocalin is a diversified superfamily and is known to have very low sequence identity among the members. However, few characteristic conserved signatures allow the identification of lipocalins. Figure 4.3 shows multiple sequence alignment of lipocalin odorant binding proteins obtained from GenDiS [Chapter 2]. The structural members (pdbcode 1dzka, 1jv4a and 1e5pa) were placed in the top of alignment. The alignment was created by multiple sequence alignment program T-COFFEE [Notredame et al., 2000]. As seen in Figure 4.3, vertebrate OBPs contain conserved GxW motif, located about 15-20 residues from the amino terminal (marked with red color box). The side chain of tryptophan (W), present in all members, has been proposed as a part of the ligand binding site. In addition, this residue is reported as structurally conserved residues (SCR) in MegaMotifBase. Two cysteine residues at 67 and 160th position in the alignment are strictly conserved in all members (marked with pink color box). These two cysteines form conserved disulfide bond, which is absent bovine OBP. Generally, vertebrate OBPs have large proportion of conserved hydrophobic residues (blue color). Residues having non polar side chains are also systematically found in few positions [marked with orange color box].
Identification and analysis of odorant bonding proteins

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>ljv4a</td>
<td>EAASSTGRNFVNEKINGEWHTILADBKREKTEDNG-NFRLFLEQIVHLE</td>
<td></td>
</tr>
<tr>
<td>1dzka</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1e5pa</td>
<td>FAEIQGKWYTVIAADNLIEKEEGG-PLRFYRHDICYK</td>
<td></td>
</tr>
<tr>
<td>E.maximus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>M.musculus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>B.taurus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.coturnix</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>E.caballus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>H.sapiens</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>P.troglodytes</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>R.norvegicus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.familiaris</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</th>
<th>51</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ljv4a</td>
<td>K-S-LVLLKFRTVEDEESGELSMDKAKDKEKALEYSVTYDFTNFTIPKTDY</td>
<td></td>
</tr>
<tr>
<td>1dzka</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1e5pa</td>
<td>KESKVYLMFFSKENGICEFSLTGQEQGNT-DVNYAGNKFVSYASE</td>
<td></td>
</tr>
<tr>
<td>E.maximus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>M.musculus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>B.taurus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.coturnix</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>E.caballus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>H.sapiens</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>P.troglodytes</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>R.norvegicus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.familiaris</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</th>
<th>101</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>ljv4a</td>
<td>DNFLMAHINEKDEGTFQLOMGLNGREPDLRSSFERFQLCEDEHGLIEN</td>
<td></td>
</tr>
<tr>
<td>1dzka</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1e5pa</td>
<td>KDF - FTNKRDRAGGETNVAGKGNALTPEENL1VQFAHKEP7VENS</td>
<td></td>
</tr>
<tr>
<td>E.maximus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>M.musculus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>B.taurus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.coturnix</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>E.caballus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>H.sapiens</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>P.troglodytes</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>R.norvegicus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.familiaris</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</th>
<th>151</th>
<th>168</th>
</tr>
</thead>
<tbody>
<tr>
<td>ljv4a</td>
<td>ITDL4EHANNCDD</td>
<td></td>
</tr>
<tr>
<td>1dzka</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1e5pa</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>E.maximus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>M.musculus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>B.taurus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.coturnix</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>E.caballus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>H.sapiens</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>P.troglodytes</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>R.norvegicus</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>C.familiaris</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Figure 4.3 Multiple sequence alignment of lipocalin odorant binding proteins obtained from GenDis. 1dzka, ljv4a and 1e5a are structural members. E.maximus - Elephas maximus; M.musculus - Mus musculus; B.taurus - Bos Taurus; E.caballus - Equus caballus; C.coturnix - Coturnix coturnix; H.sapiens - Homo sapiens; P.troglodytes - Pan troglodytes; R.norvegicus - Rattus norvegicus; C.familiaris - Canis familiaris.
The primary sequences of insect OBPs are not usually well conserved, but most members show a similar patterning of hydrophobic and hydrophilic residues that defines the helical regions. Figure 4.3 shows multiple sequence alignment of insect OBPs. Structural (1c3za and 1dzka) and sequence members were obtained from GenDiS database and the alignment was created using T-COFFEE [Notredame et al., 2000]. The most striking conservation is six cysteine residues that are present in characteristic positions in all known insect OBPs [Pelosi and Maida, 1995]. The six conserved cysteines (C1-C6) are connected by three interlocking disulphide bridges [Scaloni et al., 1999; Leal et al., 1999]. Interestingly, 1c3za has only four conserved cysteines connected by two disulfide bonds. Interlocking disulfide pattern is missing 1c3za due to the absence of cysteines C2 and C5. The disulfide bonding pattern in 1dqe and 1cz3a is shown in Figure 4.4. In addition to the conserved cysteines, the insect OBPs show conserved hydrophobic residues in few positions in the alignment. Conserved hydrophilic residues are found in one position which is marked with orange color box.
Identification and analysis of odorant bonding proteins


C1

-----------SQE-----------VRNLNSLNF-----------GKALDECKKEMLT-----------TDAINEDF
---E---ETPREEKL----------QKHSDAACKAESGV----------SEESLNKV
A. octiescostata
-----------MSEEELA----------KQLRHaCVAQTGV----------DEAHTTV
L. lineolaris
-----------A-----ALTVMAGE----------LPEEMREMA----------QLHDGVCVERTG----------DNGLIGPC
T. castaneum
-----------CLVEIFPAE----------MDDDMMKEL----------NMLNCTETGRTG----------TDDQIEN
B. mori
-----------Y----------YMALAAHGQ----------LHDEIAELA----------AMVRACADESSV----------DLNVEKV
L. migratoria
-----------LLLMAAAAARDVNRNLRGTGRMDAA----------KEVDTHCRSSSTG----------PRDMLHR
R. palmarum
-----------KFDDSI----------ISDDIIKL----------KGHLDCVKGK----------EAEIANL
A. aegypti
----------P-ALVGAAALSVFQQAANL----------EDIKGRINGETAYEACLLSAGL----------DVSSLKSL
A. gambiae
----------S--LVFPGECL----------IDISKVTLDA----------FYFLFGCARDLV----------PPEDILEY
D. melanogaster
-----------L--LLFDPAVAM----------TMEQFLTSL----------DMIRSCGAFKFR----------KTEDLDRL
M. sexta
-----------L--ILFDASYM----------SRQQLKNSG----------KMMDKSCIKNPD----------TEDEVQGI
S. guani
-----------V----------LLLIAILASA----------LAADNDDPF----------ESTRFKQREYGG----------SDEELLAL
H. picea
-----------TLE----------HHQKVMEQA----------VDAGAKCAGEL----------TPEDDLKL
L. dispar
-----------VIPVEPSKD----------VMIQMAKLF----------GKPILLOQELGA----------DDSVKVEF
A. ipsilon
-----------L--TGSVIGTAE----------VMSHVTAHF----------GKALEECRESGSL----------SAEVELFE
L. maderae
-----------AVATATLAD----------STQSFKDAM----------GPLVRECMVSGSA----------TEYDDFKTV
P. regina
-----------FVITALACNI----------PAELTKEEAM----------ITITATECCEAAG----------SDADEFAM
A. ceranaal
-----------VRVQGIDQRT----------VVAKMYEKL----------MPDIMPACDELM----------SEDIATNI
R. prolixus
-----------VIAVYNNGA----------EILKSIAM----------MKDRXCTRHMGL----------FSDKFIFIK
D. yakuba
-----------V--VT----------VLTVGFL----------MIQIAHCAAAKV----------QK------
C. sonorensis
-----------ALLLL----------NLALSLAG----------TRGTSRCEKRNKL----------DTRIQU
D. mauritiana
-----------HCACI----------FILEQFPR----------SESSNCFPHNEG----------DEDIAEAI
Z. nevadensis
-----------VLYNAVFAS----------EAVEVD----------CNEVKECHESNP----------DARYLQD
P. dominulus
-----------SDHAY----------KRYLHAP----------EPVLALCKRESLG----------EADKDK
P. ariasi
-----------PVILLTL----------GLLVVKCERPEFWKCRDFFK----------

C2  C3

Idqea

YNFWKE--GY----------EIK----------NRETOQAINCMCLSTKLNMDDEPNLH----------
---RN--RE----------EVD----------DKPLKEHACICLRLRFGIDIASFEQ----------
A. octiescostata
---KD--QKG----------FPP----------DEKFKCVKLMTMEAMIAVGDGGVVD----------
L. lineolaris
-----------AK----------GAD----------DQKLCYKFCVFVGNLGVISDEGELD----------
T. castaneum
-----------RK----------GAD----------DDSKVFKVCVDQMGCMTDGGAVD----------
B. mori
-----------NA--GTDLAT----------DGKLCYKCTMEAGMSM----------DGVD----------
L. migratoria
-----------AE----------GQ--TVD----------DDSFKLYCIMFNSLDDDVGFV----------
R. palmarum
-----------KN-AE----------FTE----------DDKLKYHVCHLIQVAMGLHAFD----------
A. aegypti
-----------QT---GD----------FSN----------GDVRKLCVKCFEFEKFMIAGEGNLH----------
A. gambiae
-----------KK-Rl----------FPP----------DQLTGCYKGLCMRLYDDYVGKFD----------
D. melanogaster
-----------RV-GD----------FNPFP----------SLQDXTKVCVLAMGTVNKKGEFN----------
M. sexta
-----------EQ----------GPFIE----------DRRVCMCYCITYMTVQVKN----------NLKLS----------
S. guani
-----------EE----------N----------LEPIMKCLEPFLKDLKTEDTGNF----------
H. picea
-----------AK--RE----------LPE----------TKAGKVCITVKIFQMGNADGITK----------
L. dispar
-----------LDFWKG-GY----------VMK----------DRTQOCMCLMAMKELDSSAMH----------
A. ipsilon
-----------QHFWKE--DF----------EUV----------HRLEOCACISMNSKFLQDDSMMH----------
L. maderae
-----------IN--RN----------PLE----------SRTAQAILLACALDKVLGELSGPEAIY----------
P. regina
-----------VHK----------FPE----------SKEGKMNATLKFVGDMSDDGGM----------
A. ceranaal
-----------QAARAN----------GAD----------MSQGLLCLKAVVVRKMTDDVL-------------
R. prolixus
-----------PS--QK----------VPS----------DEVENCLCYVTGVVQVGEGF----------
A. melliferaal
-----------QEL----------QEDISEGNKYLTHNMSCFFCERKESHIQND--EQ----------
D. yakuba
-----------YVD----------YVD----------RSLKANGNQ2LREMKTVNY-GSVH----------
C. sonorensis
-----------RN--QIQTPFK----------MNKCIKLCHYFQVQGMRMGYHMDAOI----------
D. mauritiana
-----------LSW----------PAVHDLLS----------VRSKHCVYVCLTYNYNIVYSGRIS----------
Z. nevadensis
-----------HMTG--SFPEDN----------VRSAKCFIYMNMEYMVWDSDGGAVD----------
F. dominulus
-----------LSD-ES----------TVD----------QKIFSCLAILKLDNAGLY----------GELK----------
P. ariasi
-----------IDQ----------REC---PFAYHYPDVKFRIARMA----------

48
Identification and analysis of odorant bonding proteins

**Fig. 4.4** Multiple sequence alignment of lipocalin odorant binding proteins obtained from GenDiS. 1dqea and 1c3za are structural members. *A. aegypti - Aedes aegypti; A. ipsilon - Agrotis ipsilon; A. octiescostata - Anomala octiescostata;**
Identification and analysis of odorant bonding proteins

A. gambiae - Anopheles gambiae; A. mellifera - Apis mellifera; B. mori - Bombyx mori; 
D. yakuba - Drosophila yakuba; C. sonorensis - Culicoides sonorensis; D. mauritiana - 
Drosophila mauritiana; D. melanogaster - Drosophila melanogaster; H. picea - 
Heptophylla picea; L. maderae - Leucophaea maderae; L. migratoria - Locusta 
migratoria; L. lineolaris - Lygus lineolaris; L. dispar - Lymantria dispar; M. sexta - 
Manduca sexta; P. ariasi - Phlebotomus ariasi; P. regina - Phormia regina; 
P. dominulus - Polistes dominulus; R. prolixus - Rhodnius prolixus; R. palmarum - 
Rhynchophorus palmarum; S. guani - Scleroderma guani; T. castaneum - Tribolium 
castaneum; Z. nevadensis - Zootermopsis nevadensis; A. cerana - Apis cerana

![Diagram of disulfide bonding pattern](image)

Figure 4.5: Disulfide bonding pattern in pheromone-binding protein from *Bombyx 
mori* (1DQEA) and THP12 carrier protein from *Tenebrio molitor* (1C3ZA)

4.5 Identification of odorant binding proteins from sequence information using 
machine learning approach

4.5.1 Overview

At present, prediction of odorant binding proteins is based on the sequence 
similarity search methods such as BLAST [Altschul *et al.*, 1997], HMM [Eddy, 1998] 
etc. Since OBPs show very low or even insignificant sequence similarity, these 
methods may not be employed efficiently to identify odorant binding proteins from 
genome databases. With the rapid increase in newly found protein sequences entering 
into databanks, an efficient method is needed to identify OBPs from the sequence 
databases. So far, SVM and other statistical learning methods have not been explored 
for predicting odorant binding proteins. In this section, a machine learning method 
based on regularized least squares classifier (RLSC) method is reported for the 
identification of odorant binding proteins from sequence-derived properties irrespective 
of sequence similarity.
Identification and analysis of odorant bonding proteins

4.5.2 Datasets

All odorant binding proteins were obtained from GenDiS [see chapter 2] and Pfam [Bateman et al., 2004] databases. Sequences having more than 40% sequence identity were removed from the dataset. After careful manual examination, a total of 476 odorant binding proteins were considered for the construction of positive dataset which includes 40 vertebrate odorant binding proteins, 282 insect general odorant binding proteins, 46 pheromone binding proteins and 108 chemosensory proteins. Due to the limitation in the number of known odorant binding proteins, the positive dataset could not be enhanced any further. The negative samples were taken from seed proteins of Pfam protein families [Bateman et al., 2004], which are unrelated to odorant binding proteins. The final negative dataset consists of 2157 non-odorant binding domains.

4.5.3 Derivation of physicochemical properties from protein sequence

In this work, amino acid composition and nine physicochemical properties are employed to describe each protein. Given the sequence of a protein, its amino acid composition and the properties of every constituent amino acid are computed and then used to generate feature vector. The computed amino acid properties include molecular weight, hydrophobicity, hydrophilicity, hydration potential, refractivity, average and total accessible surface area, secondary structural content and propensity of amino acids at secondary structures [Kawashima et al., 1999]. Secondary structure for each sequence is predicted using PSIPRED [McGuffin et al., 2000]. Additionally, frequencies of dipeptides and tripeptides are used to represent protein sequences for classification [Smialowski et al., 2006]. To reduce the dimensionality of feature space, the amino acids are clustered into 11 groups with similar physicochemical or structural properties. All possible pairwise and triplet combinations are computed from the 11 groups and this gives rise to 66 dipeptide and 1331 triplet combinations. The dipeptide and tripeptide frequencies are computed from each sequence and are represented by one or more pairwise and triplet combinations respectively. As a feature space, 1463 feature vectors represent each protein sequence.

4.5.4 Classification Models

All results presented here are acquired through a leave-one-out cross-validation (LOOCV) procedure. A regularized least squares classifier (RLSC) is used as the classification model. From the machine learning viewpoint, RLSC belongs to the large
family of kernel methods and is closely related to the well-known support vector machines (SVM) [Cortes and Vapnik, 1995; Burges, 1998]. The difference between RLSC and SVM is that they formulate the classification in different ways. However, both of them can achieve comparable classification performance [Zhang and Peng, 2004]. Recall that the dataset is now represented as $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i$ denotes the instance (i.e. the protein sequences) and $y_i$ is the corresponding class label.

An RLSC (denoted as $f$) typically classifies a data points $x$ by

$$f(x) = \text{sign}\left[\sum_{i=1}^{n} a_i k(x_i, x)\right]$$

(4.1)

where $k$ is the so-called kernel function that models the dissimilarity between data points $x_i$ and $x$, and the coefficients $a_i$'s are to be computed by training. In practice, the kernel function is usually defined before training the RLSC. And the $a_i$'s are computed through the training process, which involves solving a system of linear equations:

$$(K + \lambda n I)a = Y$$

(4.2)

where $a = [a_1, a_2, \ldots, a_n]^T$, $Y = [y_1, y_2, \ldots, y_n]^T$ and $\lambda$ is a predefined positive constant called the regularization parameter. $I$ is an identity matrix of size $n$. $K$ is the kernel matrix, whose components can be computed as $K_{ij} = k(x_i, x_j)$.

In our experiment, a Gaussian kernel $k(x_i, x_j) = \exp(-\sigma^2 \|x_i - x_j\|^2)$ is used for the RLSC since the Gaussian kernel is suggested as the first choice for most kernel methods. It is obvious that the values of the kernel-parameter $\sigma$ and the regularization parameter $\lambda$ are crucial to the RLSC's performance. Thus, both parameters are optimized to maximize the balanced leave-one-out accuracy.

4.5.5 Feature Selection

The main purpose of conducting feature selection is to remove possible redundant features from the original feature set. A wrapper approach [Kohavi and John, 1997] is designed to conduct feature selection. In this method, balanced leave-one-out
Identification and analysis of odorant bonding proteins

accuracy (BLOOA) of RLSC is utilized as the selection criterion. The sequential backward elimination (or the recursive feature elimination) scheme is employed as the search scheme. The feature selection procedure is started form the whole feature subset (i.e. with all the 1463 features) and BLOOA is calculated. Then, features are iteratively pruned from the feature set. At each iteration, the feature whose omission leads to the largest BLOOA is pruned.

4.5.6 Leave-one-out cross-validation

Among the independent test dataset, sub-sampling (e.g., 5 or 10-fold sub-sampling) test and jackknife test, which are often used for examining the accuracy of a statistical prediction method, the jackknife test is deemed the most rigorous and objective as analyzed by a comprehensive review [Chou and Zhang, 1995] and has been increasingly adopted by leading investigators to test the power of various prediction methods [Chou and Shen, 2006a, b; 2007a,b; Shen and Chou, 2007].

In this work, Leave-one-out (i.e., jackknife) cross-validation approach has been employed to estimate generalization performance of a classifier. It involves removing one protein from the training set, training the classifier (in our case, the RLSC) on the remaining proteins and then predicting class label of the removed (left out) protein using the trained classifier. This process was repeated until all proteins had been left out. Then the leave-one-out accuracy is computed by counting the total number of correct predictions and divided it by $n$ (i.e. the number of samples in the original dataset).

Although LOOA has been commonly used in the literature, it is also known that LOOA may not provide a precise evaluation on the performance of a classifier if a large unbalance in the population of different classes exists in the data of interest. To be specific, a good classifier is usually expected to provide high accuracy on both the positive and negative data. But LOOA will bias more to the True Positive rate if there is much more positive samples in the dataset and vice versa. Since our dataset contains much more negative instances than positive instances, alternative metrics needs to be used in addition to the LOOA. We resort to the balanced LOOA (BLOOA) [Cawley, 2006], which can be computed as:

53
Identification and analysis of odorant bonding proteins

\[ BLOOA = \frac{1}{2} (TP + TN) \]  \hspace{1cm} (4.3)

where \( TP \) and \( TN \) denote the true positive and true negative rate, respectively.

4.5.7 Discussion

The confusion matrix achieved by RLSC is given in Table 4.1. To analyze the impact of the feature selection procedure on the classification performance, we selected eight feature subsets by decreasing the number of features. The performance of the method for discriminating between odorant binding proteins and non-odorant binding proteins is summarized in Table 4.2. In this Table, TP and TN stand for true positive (correctly predicted OBPs), and true negative (correctly predicted non-class-members). The results show that our method can distinguish odorant binding proteins from other protein sequences with an accuracy of >90% and Matthews Correlation Coefficient (MCC) of 0.922, when evaluated through leave one out cross validation. Using all the 1463 features, the RLSC achieved the TP rate of 94.5% and the TN rate of 98.4%. The overall Leave-one-out accuracy (LOOA), Balanced LOOA and MCC were 97.7%, 96.5% and 0.922 respectively. As seen in Table 4.2, feature selection generally does not deteriorate the classification performance much. The usage of smaller number of features only leads to a decrease of the TN rate. The TP rate is less influenced by the feature selection. In some cases, feature selection even leads to slight increase of the TP rates.

<table>
<thead>
<tr>
<th>Original class</th>
<th>Predicted class</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>451</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>35</td>
<td>2122</td>
</tr>
</tbody>
</table>

Table 4.1: Confusion matrix for RLSC on the training dataset

54
Identification and analysis of odorant bonding proteins

Table 4.2 Classification results achieved on different feature subsets. The optimal values of $\sigma$ and $\lambda$ are also given.

<table>
<thead>
<tr>
<th>Features</th>
<th>$\sigma$</th>
<th>$\lambda$</th>
<th>LOOA</th>
<th>BLOOA</th>
<th>TP rates</th>
<th>TN rates</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1463</td>
<td>2.614e-5</td>
<td>1e-009</td>
<td>0.977</td>
<td>0.965</td>
<td>0.945</td>
<td>0.984</td>
<td>0.922</td>
</tr>
<tr>
<td>450</td>
<td>4.714e-5</td>
<td>1e-009</td>
<td>0.975</td>
<td>0.963</td>
<td>0.945</td>
<td>0.981</td>
<td>0.915</td>
</tr>
<tr>
<td>250</td>
<td>6.325e-5</td>
<td>1e-008</td>
<td>0.970</td>
<td>0.961</td>
<td>0.948</td>
<td>0.975</td>
<td>0.901</td>
</tr>
<tr>
<td>100</td>
<td>1e-4</td>
<td>1e-008</td>
<td>0.970</td>
<td>0.962</td>
<td>0.950</td>
<td>0.975</td>
<td>0.903</td>
</tr>
<tr>
<td>50</td>
<td>1.414e-4</td>
<td>1e-008</td>
<td>0.967</td>
<td>0.958</td>
<td>0.945</td>
<td>0.971</td>
<td>0.891</td>
</tr>
</tbody>
</table>

LOOA - Leave-one-out accuracy (LOOA); BLOOA – Balanced LOOA
MCC- Matthews Correlation Coefficient; $\sigma$ - Kernel-parameter
$\lambda$ - Regularization parameter; TN – True negative; TP-True positive

To test the capability, our algorithm was evaluated by independent dataset obtained from NCBI database using keyword search. The keywords used for the search includes “odorant binding proteins”, “pheromone binding proteins”, “chemosensory proteins”, “antennal protein” and “juvenile hormone binding proteins”. The sequences that are present in the positive training dataset were removed from the list. After careful manual inspection, 414 odorant binding proteins were selected for independent testing. The performance of our algorithm was compared with PSI-BLAST [Altschul et al., 1997] and HMM [Eddy, 1998]. PSI-BLAST search for each sequence was carried out against the database of positive training dataset. HMM analysis for each query sequence was performed against the HMM profile obtained from the positive training dataset. Our approach correctly predicts 402 proteins as odorant binding proteins whereas PSI-BLAST and HMM methods predict 369 and 360 proteins respectively. The overall prediction accuracy for our approach, PSI-BLAST and HMM method is 97.1%, 89.1% and 86.9% respectively (Table 4.3).
Table 4.3 Prediction result of 414 odorant binding proteins by RLSC, PSI-BLAST and HMM methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Correctly predicted as odorant binding proteins</th>
<th>Incorrectly predicted as non odorant binding proteins</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLSC</td>
<td>402</td>
<td>12</td>
<td>97.1%</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>369</td>
<td>45</td>
<td>89.1%</td>
</tr>
<tr>
<td>HMM</td>
<td>360</td>
<td>54</td>
<td>86.9%</td>
</tr>
</tbody>
</table>

Further analysis of 414 odorant binding proteins shows that 56 proteins have no single homologous protein in the SWISSPROT [Bairoch and Apweiler, 1996] database based on PSI-BLAST search result. A similarity E-value threshold of 0.01 was used for homologue search to ensure maximum exclusion of proteins that have a homologue. Our method correctly predicts 52 proteins as odorant binding proteins. This result shows the capability of our prediction systems for recognizing novel odorant binding proteins that are non-homologous to other proteins.

4.6 Genomic distribution of odorant binding proteins: OBDD

4.6.1 Odorant Binding Domain Database (OBDD)

The OBDD is an odorant binding domain database containing comprehensive collection of Odorant Binding Proteins (OBP) from Bacteria, animal and Insects. This database reports odorant binding proteins from different species. Information about the odorant binding proteins at the family level and taxonomic level provided. The database can be accessed from http://caps.ncbs.res.in/obdd/

4.6.2 Searching for OBP sequences in genome databases

The genome-wide survey of OBPs stems from the three-fold searches starting from sequence and structural members from SCOP [Murzin et al., 1995], PASS2 [Bhaduri et al., 2004a], GenDiS [Chapter 2] and PFAM [Bateman et al., 2004] databases and additionally consulted the literature. Multiple approaches such as PSI-BLAST [Altschul et al., 1997], HMMsearch of the HMMer suite [Eddy, 1998] and structural motif constrained PHI-BLAST [Bhaduri et al., 2004] have been employed to
identify homologues in the sequence databases. Structural motif definitions for structural members are obtained from MegaMotifbase and imotdb databases. In addition, other resources like INTERPRO [Mulder et al., 2003] have also been consulted to obtain sequence family members and derive an enriched database of putative members after careful validations using multiple bioinformatic approaches and machine learning method described in section 4.5.

4.6.3 Result and Statistics

OBDD consists of totally 1226 odorant binding domains from 184 species distributed into 4 phyla. The most highly populated phyla are Arthropoda which contains 914 domains from 124 genomes (Table 4.4). Further, 1226 odorant domains are assigned to four protein families (Table 4.4). The family definition is taken from PFAM database [Bateman et al., 2004]. Our analysis shows that Drosophila melanogaster, Anopheles gambiae and Mus musculus contains approximately 51% of 1226 OBPs.

Table 4.4 Number of genomes and number of OBPs in 4 phyla

<table>
<thead>
<tr>
<th>Phyla</th>
<th>Number of genomes</th>
<th>Number of OBPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthropoda</td>
<td>124</td>
<td>914</td>
</tr>
<tr>
<td>Chordata</td>
<td>44</td>
<td>282</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>Nematoda</td>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>

4.6.4 Other features

Multiple sequence alignment is provided for each genome and family. The alignment is created using CLUSTALW [Thompson et al., 1994]. The conserved residues are marked on the alignment. Phylogenetic information is provided for each family and phyla. Phylogenetic analysis performed using PHYLIP program [Felsenstein, 1985]. The complete taxonomic hierarchy is provided for each genome.
4.7 Summary

Odorant Binding Proteins (OBPs) play major role in the olfaction. The physiological function of odorant binding protein is still unclear due to the lack of more sequence data. OBPs are identified from genome databases using multiple sequence search approaches. In addition, a SVM method is proposed for the identification of OBP using sequence derived properties. Overall prediction accuracy of 97.7% (94.5% and 98.4% for positive and negative classes respectively) shows that this method is potentially useful tool for the prediction of odorant-binding proteins that have insignificant sequence similarity. The OBPs obtained from 3 fold search approach and SVM method are compiled into a database OBDD.