Amino Acids (2007) 32: 277–283
DOI 10.1007/s00726-006-0341-y
Printed in The Netherlands

Amino Acids

Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition

Z. Wen, M. Li, Y. Li, Y. Guo, and **K. Wang**

College of Chemistry, Sichuan University, Chengdu, Sichuan, China

Received February 28, 2006
Accepted March 23, 2006
Published online May 26, 2006;

Summary. As an important transmembrane protein family in eukaryon, G-protein coupled receptors (GPCRs) play a significant role in cellular signal transduction and are important targets for drug design. However, it is very difficult to resolve their tertiary structure by X-ray crystallography. In this study, we have developed a Delaunay model, which constructs a series of simplexes with latent variables to classify the families of GPCRs and projects unknown sequences to principle component space (PC-space) to predict their topology. Computational results show that, for the classification of GPCRs, the method achieves the accuracy of 91.0 and 87.6% for Class A, more than 80% for the other three classes in differentiating GPCRs from non-GPCRs and 70% for discriminating between four major classes of GPCR, respectively. When recognizing the structure of GPCRs, all the N-terminals of sequences can be determined correctly. The maximum accuracy of predicting transmembrane segments is achieved in the 7^{th} transmembrane segment of *Rhodopsin*, which is 99.4%, and the average error is 2.1 amino acids, which is the lowest in all of the segments prediction. This method could provide structural information of a novel GPCR as a tool for experiments and other algorithms of structure prediction of GPCRs. Academic users should send their request for the MATLAB program for classifying GPCRs and predicting the topology of them at liml@scu.edu.cn.

Keywords: Delaunay triangulation model – Partial least squares – G-protein coupled receptor – Classification – Structure recognition

1. Introduction

G-protein coupled receptors (GPCRs) are one of the largest superfamilies of membrane proteins in human. According to similarity of sequences and ligands binding data, GPCRs can be divided into five classes (Horn et al., 1998): Class A (receptors related to rhodopsin and the adrenergic receptor), Class B (receptors related to the calcitonin and PTH/PTHrP receptors), Class C (receptors related to the metabotropic receptors), Class D (receptors related to the pheromone receptors), Class E (receptors related to the cAMP receptors). The topology of GPCRs includes five main components, namely, the N-terminal and the C-terminal of the sequences, seven hydrophobic transmembrane α-helixes (TM1–TM7), three extracellular loops and three or four loops in the cytoplasm. This arrangement makes these proteins capable of transducing an extracellular signal into the cell (Attwood et al., 2001) and is helpful to discover the mechanism of signal transduction in or between the cells by annotating the structure and function of GPCRs.

Unfortunately, it is hard to solve the tertiary structures of GPCRs by X-ray crystallography, except for that of one GPCR (bovine rhodopsin) (Palczewski et al., 2000). Meanwhile, a number of computational methods have been developed trying to predict various attributes of proteins based on their sequences (Cai et al., 2003; Chou and Cai, 2002, 2005; Chou, 2005a; Chou and Elrod, 1999; Chou and Zhang, 1994; Feng, 2001, 2002; Lubec et al., 2005; Shen et al., 2005; Xiao et al., 2005, 2006; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). Early methods rely mainly on alignment and similarity-based comparisons (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Altschul et al., 1990, 1997; Pearson and Lipman, 1988; Pearson, 2000). In recent years, basing the analysis on common patterns and profiles, it is taken into account that the structure and function of proteins are determined by the physicochemical properties of their sequence constituents. Based on the concept of pseudo amino acid composition (Chou, 2001), the Fourier transform spectra has been used to predict membrane protein type (Liu et al., 2005a; Wang et al., 2004, 2005), particularly their low-frequency parts (Chou, 1988),

have been used to predict membrane protein types. A number of methods have been also developed to predict GPCRs (Chou, 2005b; Chou and Elrod, 2002; Elrod and Chou, 2002; Inoue et al., 2004), classify GPCRs at family and subfamily level (Bhasin and Raghava, 2004; Karchin et al., 2002; Lapinsh et al., 2002) and recognize the transmembrane segments of the proteins (Cserzö et al., 1997; Hirokawa et al., 1998; Tusnády and Simon, 1998; Pasquier et al., 1999; Lio and Vannucci, 2000; Krogh et al., 2001; Qiu et al., 2004). However, all the transmembrane segments recognizing algorithms, except TMHMM, cannot predict the 7 transmembrane segments of GPCRs exactly (Möller et al., 2001).

Delaunay triangulation method creates a series of simplexes in PC-space, which ensures that it gives the smallest space in clustering analysis compared with other algorithms (Jin et al., 2003). Following our previous work, which have successfully predicted GPCR subfamilies (Guo et al., 2005) and classified GPCRs and NRs (Guo et al., 2006), it becomes a kernel issue how to resolve the structure features of GPCRs efficiently. In the present study, a fast Delaunay model with PLS has been proposed to predict their topology. By projecting to latent variables and converting the protein sequences into digital signal with three descriptor scales (Sjøstrøm et al., 1995), a set of Delaunay tessellations has been constructed with two main latent variables for unknown protein sequence identification and structure recognition.

2. Materials and methods

Delaunay triangulation method and PLS

In multivariate calibration, the Delaunay triangulation (DT) method has been introduced as a new method with the advantages of less time consuming and the ability of forming the best simplexes (Jin et al., 2003). The DT method (in two dimensions) connects a given set of mesh nodes in such a way that the circle circumscribing any triangular element contains only the nodal points belonging to that triangle (except in the case where four or more nodal points are co-circular). There are two restrictions in this procedure: one is that the circumcircle of every triangle does not contain other data points, the other is that the smallest interior angle of each triangle in the Delaunay mesh must be as large as possible. These two restrictions ensure that the simplex surrounds the point to be predicted well and the error bound can be reduced.

When the Delaunay mesh has been constructed, it is used for prediction of new samples. In two dimensions, if a new sample M falls within a simplex defined by 3 neighbours (M_1, M_2, M_3), its associated property can be calculated through the properties of 3 neighbours according to the following equations:

$$\alpha_{M_1} = \frac{(x_{2M} - x_{2M_2})(x_{1M_3} - x_{1M_2}) + (x_{1M} - x_{1M_2})(x_{2M_2} - x_{2M_3})}{(x_{2M_1} - x_{2M_2})(x_{1M_3} - x_{1M_2}) + (x_{1M_1} - x_{1M_2})(x_{2M_2} - x_{2M_3})} \quad (1)$$

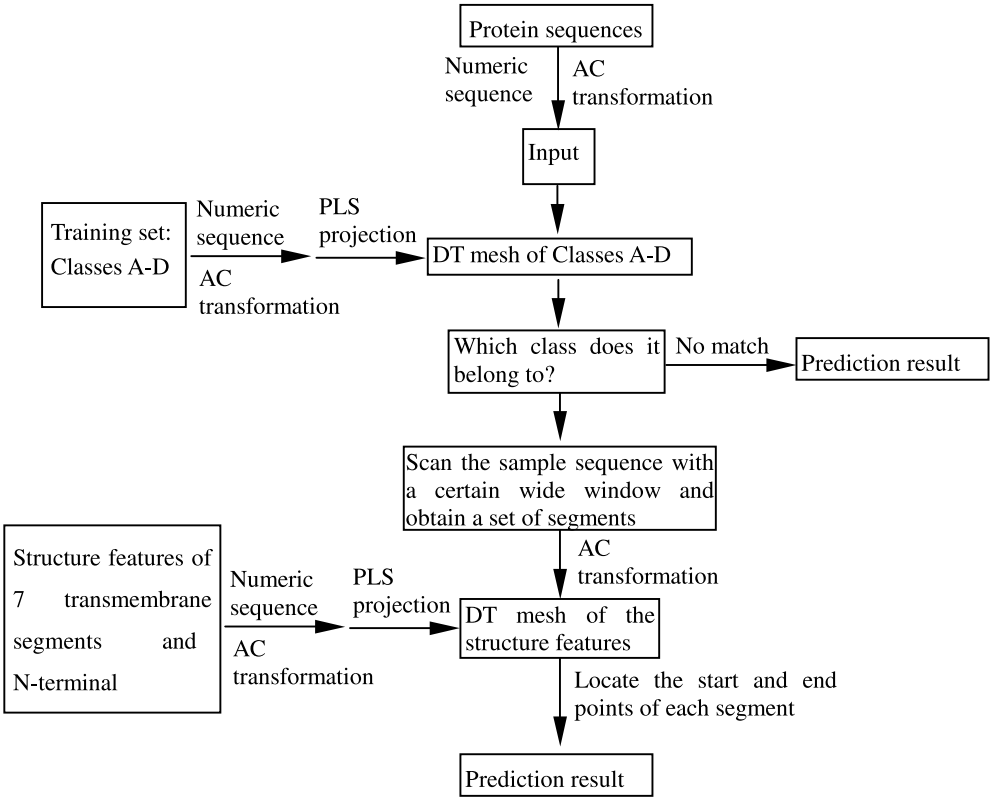

Fig. 1. Diagrammatic view of the classification procedure and topology prediction

$$\alpha_{M_2} = \frac{(x_{2M} - x_{2M_1})(x_{1M_3} - x_{1M_1}) + (x_{1M} - x_{1M_1})(x_{2M_1} - x_{2M_3})}{(x_{2M_2} - x_{2M_1})(x_{1M_3} - x_{1M_1}) + (x_{1M_2} - x_{1M_1})(x_{2M_1} - x_{2M_3})} \quad (2)$$

$$\alpha_{M_3} = \frac{(x_{2M} - x_{2M_2})(x_{1M_1} - x_{1M_2}) + (x_{1M} - x_{1M_2})(x_{2M_2} - x_{2M_1})}{(x_{2M_3} - x_{2M_2})(x_{1M_1} - x_{1M_2}) + (x_{1M_3} - x_{1M_2})(x_{2M_2} - x_{2M_1})} \quad (3)$$

where x_{1i} and x_{2i} are the scores of the objects in PC-space, α_{M_1}, α_{M_2} and α_{M_3} are the contribution of samples M_1, M_2 and M_3, respectively. Then the property of an unknown sample can calculated as follows:

$$y_M = \alpha_{M_1} y_{M_1} + \alpha_{M_2} y_{M_2} + \alpha_{M_3} y_{M_3}, \quad (4)$$

where y_M, y_{M_1}, y_{M_2} and y_{M_3} are the property for samples M, M_1, M_2 and M_3, respectively.

Compared with other multiple calibration methods, the results from DT are better than those from the law of mixtures (LM) method (Jin et al., 2003) and the components used in DT for prediction are fewer than those used in PLS and principle component regression (PCR) (Jin et al., 2003). Therefore, the DT method is our mainly method in this study to identify an unknown protein sequence and locate the positions of 7 transmembrane segments and N-terminal of GPCRs.

Partial least squares projection to latent structures (PLS) as a chemometrics tool for multivariate calibration is widely applied in many fields (Wold et al., 2001). When modeling multidimensional data, PLS is often suggested to reduce original variables. It can save much time in calculation procedure. It provides latent variables that are the combinations of original variables, and relates predictor variables X and response variables y by means of the linear regression model:

$$y = b_0 + Xb + e, \quad (5)$$

where b is the PLS coefficient matrix, e is the residue vector and b_0 is the offset vector:

$$b_0 = \bar{y} - \bar{X}b. \quad (6)$$

In classical PLS modeling, the eigenvector w of the mixed dependent and independent variables matrix $X^T YY^T X$ should be calculated first and PLS component is obtained as follows:

$$b = Xw \quad (7)$$

This can be accomplished by using NIPALS, the power method (Wu et al., 1997).

Datasets

The dataset used for training and testing is four major classes of GPCRs (1526 from Class A, 231 from Class B, 160 from Class C and 58 from class D) obtained from the March 2005 release of the GPCR database (http://www.gpcr.org/7tm/; Horn et al., 1998). When we predict the topology of GPCRs, 4 subfamilies from Class A: *Amine*, *Olfactory*, *Peptide* and *Rhodopsin*, are selected as the training and testing datasets. The proteins with high sequence identity in all classes were not removed in order to provide enough sequences to construct DT mesh that can be applied to GPCR families.

Classification and structure recognition

In this paper, three steps have been taken to identify whether an unknown protein sequence is GPCR and which Class or family it belongs to. Then,

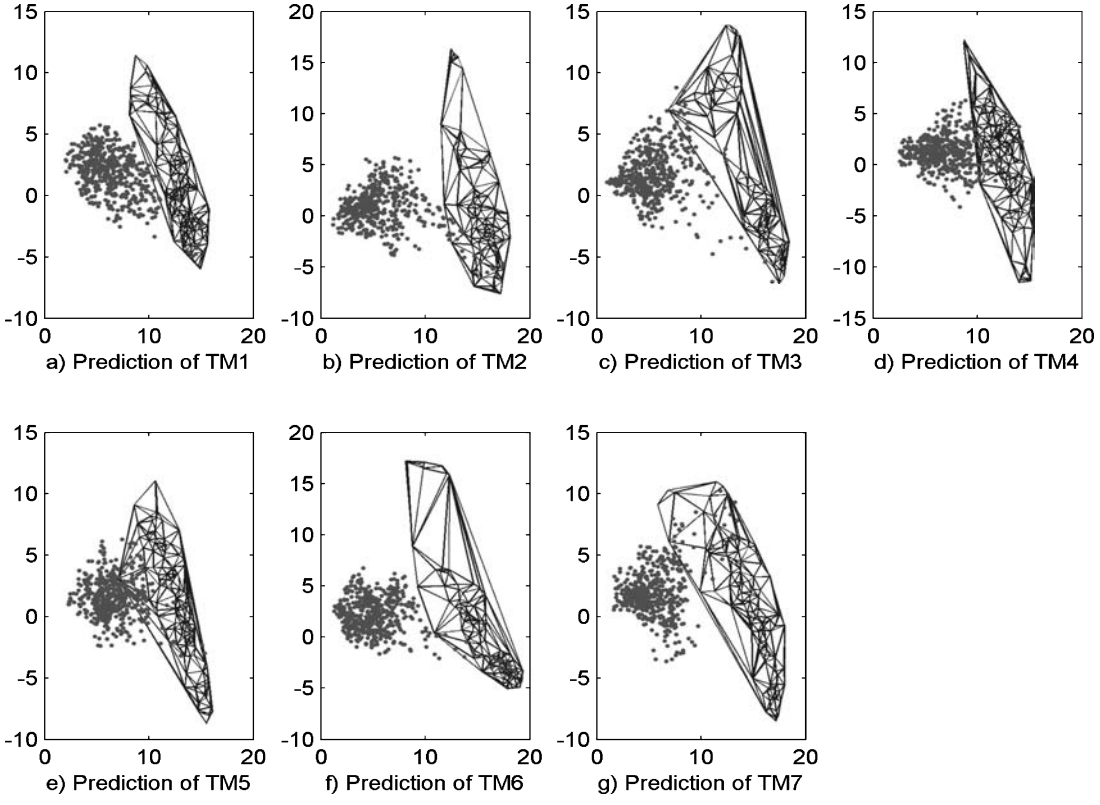

Fig. 2. Projected results of 7 transmembrane segments of *Alpha-1B adrenergic receptor* (P35368) by PLS. Subgraphs **a**–**g** show the distribution of the sample sequence projected to PC-space. The DT-meshes are constructed by 242 corresponding transmembrane segments of training sequences of *Amine* and the points indicate the segments of *Alpha-1B adrenergic receptor*, which are obtained by scanning the protein with a window size of 20 amino acids. When the points fall into the DT-mesh, we consider that these segments are the transmembrane segments and locate the start and the end position

according to the structure features of the specific family, we predict the topology of the sequence. The whole procedure is shown in Fig. 1.

Initially, all of the sequences, including the sample and the training sets (Class A–D), are converted into numeric array by substituting each amino acid with three descriptor scales (Sjøstrøm et al., 1995). All auto-covariances (ACs) (Sjøstrøm et al., 1995) of these sequences are calculated according to Eq. (8) with all lags from $-lg$ to lg, which form a new multivariate data matrix with dimensionality m (objects) times $(2 \times lg + 1) \times 3 \times 3$ (variables).

$$AC_{x(j,k),\mathrm{lag}} = \sum_{i=1}^{N-|\mathrm{lag}|} \left(x_j(i+\mathrm{lag}) - \frac{1}{N}\sum_{i=1}^{N} x_j(i) \right) \left(x_k(i) - \frac{1}{N}\sum_{i=1}^{N} x_k(i) \right), \tag{8}$$

where index i, lag are the amino acids positions and the lags $(\mathrm{lag} = [-lg, lg])$, respectively and N is the length of a sequence. Index j and k are the descriptor scales ($j = 1, 2, 3$ and $k = 1, 2, 3$).

Secondly, two latent variables (PLS1 and PLS2) of the ACs matrix of the training sets (Class A–D) are extracted to form a new PC-space with PLS1 versus PLS2. All of the coefficients are projected to the PC-space and used to construct the DT mesh (Wold et al., 1993). When the sample sequence is projected to the same space, we can determine whether this sample belongs to the Class or family according to the projected point falling in the DT mesh or out of it. If the sample is classified as the Class A, an additional DT mesh should be constructed with the sequences of the subfamilies of the Class A to determine which subfamilies it belongs to.

Finally, 8 structure features of the specific Class or subfamily, namely, 7 transmembrane segments (TM1–TM7) and N-terminal, are selected to calculate the ACs and constitute 8 characteristic DT meshes. To determine the positions of the corresponding region in the sample sequence, the sliding window analysis with the window size of L amino acids as used by Chou (2001a, b, c; 2002) in predicting protein signal peptides is carried out. According to the experimental data, the transmembrane segments in GPCRs usually include 20–25 amino acids, and the length of N-terminal in each class of GPCRs is different. In order to achieve the best-predicted results, a proper length of the segments should be taken to project to the PC-space. In this paper, L is 20 for transmembrane segments and is decided for N-terminal by specific class or subfamily of GPCRs. Therefore, we can obtain $n - L$ segments of the sample sequence in this way. These segments are projected to the same space of the characteristic DT meshes in turn. Then, the start point of each characteristic region can be located by selecting the first point falling in the DT mesh and the length of it can be calculated by $m + L$, where m is the points in the DT mesh. All the results of our experiments are detailed in Section 3.

As an example, consider the subgraghs a–g of projecting results of 7 transmembrane segments of *Alpha-1B adrenergic receptor* (P35368) (Fig. 2). Each DT-mesh in the subgragh is constructed with projecting points of the corresponding transmembrane segments of 242 sequences in *Amine*. Then, all the segments of *Alpha-1B adrenergic receptor* are obtained by scanning it with a fixed window size (20 amino acids) and are projected to the PC-space (points in the subgragh). When the points fall into the DT-mesh, we consider that these segments are the transmembrane segments. By finding the 1^{st} point falling into the mesh and counting the number of points, we can locate the start and the end position of the transmembrane segment. Table 1 shows the predicted results of 7 transmembrane segments. Predicted results of three other algorithms, PRED-TMR (Pasquier et al., 1999), HMMTOP (Tusnády and Simon, 1998, 2001) and TMHMM (Krogh et al., 2001), are also listed in the table. It can be seen that the predicted results of our method and TMHMM are very close to the measured data, whereas two other algorithms make more prediction errors.

Table 1. Comparing predicted results of our method with measured data and the results of three other algorithms in recognizing 7 transmembrane segments of *Alpha-1B adrenergic receptor* (P35368)

	Measured (AA)	Predicted (AA)	PRED-TMR (AA)	HMMTOP (AA)	TMHMM (AA)
TM1	46–70	45–74	69–89	68–90	48–70
TM2	84–105	81–107	–	103–125	82–104
TM3	116–141	120–141	113–134	140–161	119–141
TM4	162–182	161–182	141–162	182–203	162–184
TM5	202–224	194–220	183–203	222–241	204–226
TM6	296–319	294–320	223–243	315–334	295–317
TM7	327–340	325–349	317–335	349–368	332–351

3. Results and discussion

As is well known, the jackknife (leave-one-out) test is an objective and rigorous testing procedure and has been used widely for cross-validation in statistical prediction (Chou and Zhang, 1995; Cai and Chou, 2005; Chou, 2005c; Chou, 1995; Chou and Cai, 2004; Fielding and Bell, 1997; Gao et al., 2005; Liu et al., 2005b; Shen and Chou, 2005a; Shen and Chou, 2005b; Xiao et al., 2005; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). However, it is time-consuming for large samples. In this paper, 5-fold cross-validation is adopted to test Class A and other Classes (Class B, Class C and Class D) are validated by jackknife test. In the 5-fold cross-validation, the Class A is separated randomly into five equal-size sets. One of them is used for testing and the others are used for constructing the DT mesh in turn. During the process of the jackknife, n samples of 1 case in the Class of GPCR are tested sequentially. The remaining $n - 1$ cases form the training set and construct the DT mesh.

To measure the performance of our method, we use specificity (Sp), sensitivity (Se), total accuracy (Acc) and Matthew's correlation coefficient (MCC), which can provide a better summary of performance in our study (Matthews, 1975; Baldi et al., 2000).

The testing results on Class A–D are shown in Table 2. The module performs the best classification on Class A with the total accuracy of 91.0% in differentiating GPCRs from non-GPCRs and can classify the four classes with more than 80% accuracy. The method performs the best in Class A among the four classes. In distinguishing the GPCRs from non-GPCRs, all the Sp of our method in the four families is over 90%; the best Sp is 98.9% in Class A. In identifying the classes of the GPCR, the highest Sp, Se and Acc are still in Class A, namely 95.2%, 91.4% and 87.6%. The lowest Se is 58.6% in Class D. It may be because there are sufficient training sequences in Class A to form the DT-mesh, whereas the less training

Table 2. The performance of the method in differentiating GPCRs from non-GPCRs and discriminating between four major classes of GPCR

GPCR classes	Number of sequences	Non-GPCRs				GPCRs			
		Sp (%)	Se (%)	Acc (%)	MCC	Sp (%)	Se (%)	Acc (%)	MCC
Class A	1526	98.9	91.4	91.0	0.54	95.2	91.4	87.6	0.11
Class B	231	95.0	83.6	85.5	0.69	79.1	83.6	72.3	0.28
Class C	160	94.3	83.1	85.6	0.73	78.2	83.1	74.4	0.43
Class D	58	97.1	58.6	84.2	0.67	64.2	58.6	71.0	0.38

Sp, Se, Acc, MCC: specificity, sensitivity, total accuracy and Matthew's correlation coefficient, respectively. Class A, Class B, Class C and Class D of GPCR are selected from GPCR database after removing the fragmental sequences. The method performs better in Class A than in the other three classes. In distinguishing the GPCRs from non-GPCRs, all the Sp of our method in the four families is over 90%, best Sp is 98.9% in Class A. In identifying the classes of the GPCR, the highest Sp, Se and Acc are still in Class A, namely 95.2%, 91.4% and 87.6%

set of Class D results in more errors. To calculate the true negative (TN) and the false positive (FP) of our method in discriminating between the four major classes of GPCR, the DT mesh is constructed by one class and 90 negative samples selected randomly from the other three classes. The total accuracy decreases slightly owing to the increasing identity of the sequences between the test set and training set. In this case, the method can also achieve the accuracy of more than 70%. Poor results of MCC were obtained for Class A, as the numbers of the negative samples are much less than those of the positive samples.

When one GPCR protein's family is identified, we try to predict its topology. Four subfamilies of Class A, namely, Amine, Olfactory, Peptide and Rhodopsin, have been selected as the test sets. All the segmental sequences and the sequences without the annotation of transmembrane segments are removed. To assess the performance of method in predicting the start and end positions of 7 transmembrane segments and locating the N-terminal, the jackknife/Leave-one-out test has been used in this case (Fielding and Bell, 1997).

During the procedure of the jackknife test, one sequence of the subfamily has been singled out as the testing data in turn and the 8 feature structures (7 transmembrane segments and N-terminal) of the rest are projected to the PC-space separately. Then, 8 DT-meshes are constructed with the PLS coefficients. After scanning the test sequence with a fixed window size, we project the segments of the test sequence to the PC-space and locate the positions of the feature structures according to the points falling into the DT-mesh.

The performance of our method in predicting the topology of four subfamilies of Class A is shown in Table 3. More than 90% accuracy has been achieved in predicting the topology of *Rhodopsin* and the average error is within 8 amino acid residues. In addition, all the N-terminals of the sequences are correctly predicted. The prediction accuracy of Peptide in our method is lower than that of the other three subfamilies because of the much discrepancy of the structures in these four subfamilies. In *Peptide*, the N-terminal often includes some structures, which are similar to transmembrane segments, such as signal peptide,

Table 3. The results of predicting 7 transmembrane segments and N-terminal of four subfamilies of Class A

Subfamilies of Class A	Seq		Prediction results of all segments							
			TM1	TM2	TM3	TM4	TM5	TM6	TM7	N-term
Amine	243	PN	183	231	197	207	175	223	230	243
		Err	4.1	4.3	3.0	2.9	6.2	4.5	3.7	–
Olfactory	395	PN	328	364	378	347	340	380	383	395
		Err	8.0	5.0	4.6	6.6	10.1	3.5	3.5	–
Peptide	486	PN	338	337	338	339	308	305	306	486
		Err	3.6	6.4	6.0	5.1	5.6	3.9	5.7	–
Rhodopsin	168	PN	152	161	158	162	157	164	167	168
		Err	2.7	5.6	4.0	5.4	7.4	2.7	2.1	–

Seq, TM1–TM7, PN, Err: total number of sequences, 7 transmembrane segments, number of segments predicted correctly in our method and average errors (*n* amino acid residues) in predicting the start and the end positions of 7 transmembrane segments, respectively. The prediction of N-term is to determine which side is the N-terminal of the sequence

and the third loop regions consist of 10–50 amino acids residues, which are much less than those of the other three subfamilies.

4. Conclusions

The present study illustrates a Delaunay model, which based on PLS, for classification of GPCRs and structure recognition. For classification, the method can achieve more than 80% and 70% accuracy in differentiating an unknown sequence from non-GPCRs and between the four major classes, respectively. When predicting the topology of the GPCR protein sequence, the model also gets a high accuracy, especially in predicting the N-terminal of the sequence. Accordingly, this model can be used to identify a novel GPCR as well as its topology. Moreover, the standard DT-mesh can be constructed when the GPCRs database is established. It should save much time in the procedure of classification and prediction. For our further study, the model will be extended to solve issues of the tertiary structure of proteins and the interaction between specific proteins.

Acknowledgement

This work was supported by the foundation of the State Key Laboratory of Chemo/Biosensing and Chemometrics.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410

Altschul SF, Madden TL, Schäffer 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: 3389–3402

Attwood TK, Croning MDR, Gaulton A (2002) Deriving structural and functional insights from a ligand-based hierarchical classification of G protein coupled receptors. Protein Eng 15: 7–12

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16: 412–424

Bhasin M, Raghava GPS (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res 32: W383–W389

Cai YD, Chou KC (2005) Using functional domain composition to predict enzyme family classes. J Proteome Res 4: 109–111

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

Chou KC (1988) Review: Low-frequency collective motion in biomacromolecules and its biological functions. Biophys Chem 30: 3–48

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21: 319–344

Chou KC (2001a) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43: 246–255 (Erratum: ibid., 2001, 44: 60)

Chou KC (2001b) Using subsite coupling to predict signal peptides. Protein Eng 14: 75–79

Chou KC (2001c) Prediction of signal peptides using scaled window. Peptides 22: 1973–1979

Chou KC (2002) Review: Prediction of protein signal sequences. Curr Prot Peptide Sci 3: 615–622

Chou KC (2005a) Review: Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6: 423–436

Chou KC (2005b) Prediction of G-protein-coupled receptor classes. J Proteome Res 4: 1413–1418

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

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321: 1007–1009 (Corrigendum: ibid., 2005, 329: 1362)

Chou KC, Cai YD (2005) Predicting protein localization in budding yeast. Bioinformatics 21: 944–950

Chou KC, Elrod DW (1999a) Protein subcellular location prediction. Protein Eng 12: 107–118

Chou KC, Elrod DW (1999b) Prediction of membrane protein types and subcellular locations. Proteins 34: 137–153

Chou KC, Elrod DW (2002) Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 1: 429–433

Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269: 22014–22020

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Cserzö M, Wallin E, Simon I, von Heijne G, Elofsson A (1997) Prediction of transmembrane α-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 10: 673–676

Elrod DW, Chou KC (2002) A study on the correlation of G-protein coupled receptor types with amino acid composition. Protein Eng 15: 713–715

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2: 291–303

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environm Conserv 24: 38–49

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guo YZ, Li ML, Lu MC, Wen ZN, Wang KL, Li GB, Wu J (2006) Classifying GPCRs and NRs based on protein power spectrum from fast Fourier transform. Amino Acids (in press)

Guo YZ, Li ML, Wang KL, Wen ZN, Lu ML, Liu LX, Jiang L (2005) Fast Fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies. Acta Biochim Biophys Sin 37: 759–766

Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14: 378–379

Horn F, Weare J, Beukers MW, Horsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G (1998) GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res 26: 275–279

Inoue Y, Ikeda M, Shimizu T (2004) Proteome-wide functional classification and identification of mammalian-type GPCRs by binary topology pattern. Comp Biol Chem 28: 39–49

Jin L, Pierna JAF, Xu Q, Wahl F, de Noord OE, Saby CA, Massart DL (2003a) Delaunay triangulation method for multivariate calibration. Anal Chim Acta 488: 1–14

Jin L, Pierna JAF, Wahl F, Dardenne P, Massart DL (2003b) The law of mixtures method for multivariate calibration. Anal Chim Acta 476: 73–84

Karchin R, Karplus K, Haussler D (2002) Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18: 147–159

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. J Mol Biol 305: 567–580

Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JES (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci 11: 795–805

Lio P, Vannucci M (2000) Wavelet change-point prediction of transmembrane proteins. Bioinformatics 16: 376–382

Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737–739

Liu H, Yang J, Ling JG, Chou KC (2005b) Prediction of protein signal sequences and their cleavage sites by statistical rulers. Biochem Biophys Res Commun 338: 1005–1011

Lubec G, Afjehi-Sadat L, Yang JW, John JP (2005) Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol 77: 90–127

Matthews BW (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451

Möller S, Croning MDR, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17: 646–653

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443–453

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Crystal structure of rhodopsin: a G-protein coupled receptor. Science 289: 739–745

Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ (1999) A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng 12: 381–385

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132: 185–219

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444–2448

Qiu J, Liang R, Zou X, Mo J (2004) Prediction of transmembrane proteins based on the continuous wavelet transform. J Chem Inf Comput Sci 44: 741–747

Shen HB, Chou KC (2005a) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334: 288–292

Shen HB, Chou KC (2005b) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334: 577–581

Sjøstrøm M, Rännar S, Wieslander Å (1995) Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances. Chemom Intell Lab Syst 29: 295–305

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197

Tusnády GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283: 489–506

Tusnády GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17: 849–850

Wang M, Yang J, Xu ZJ, Chou KC (2005) SLLE for predicting membrane protein types. J Theor Biol 232: 7–15

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

Wold S, Jonsson J, Sjørstrøm M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277: 239–253

Wold S, Sjøstrøm M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58: 109–130

Wu W, Massart D, de Jong S (1997) The kernel PCA algorithms for wide data. Part I: theory and algorithms. Chemom Intell Lab Syst 36: 165–172

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50: 44–48

Authors' address: Prof. Menglong Li, College of Chemistry, Sichuan University, Chengdu 610064, China,
Fax: +86-28-85412356, E-mail: liml@scu.edu.cn