Molecular BioSystems

METHOD

GPCRserver: an accurate and novel G protein-coupled receptor predictor†

Cite this: *Mol. BioSyst.*, 2014, **10**, 2495

Renxiang Yan,*[a] Xiaofeng Wang,[b] Lanqing Huang,[a] Jun Lin,[a] Weiwen Cai[a] and Ziding Zhang[b]

Received 5th May 2014, Accepted 4th July 2014

DOI: 10.1039/c4mb00272e

www.rsc.org/molecularbiosystems

G protein coupled receptors (GPCRs), also known as seven-transmembrane domain receptors, pass through the cellular membrane seven times and play diverse biological roles in the cells such as signaling, transporting of molecules and cell–cell communication. In this work, we develop a web server, namely the GPCRserver, which is capable of identifying GPCRs from genomic sequences, and locating their transmembrane regions. The GPCRserver contains three modules: (1) the Trans-GPCR for the transmembrane region prediction by using sequence evolutionary profiles with the assistance of neural network training, (2) the SSEA-GPCR for identifying GPCRs from genomic data by using secondary structure element alignment, and (3) the PPA-GPCR for identifying GPCRs by using profile-to-profile alignment. Our predictor was strictly benchmarked and showed its favorable performance in the real application. The web server and stand-alone programs are publicly available at http://genomics.fzu.edu.cn/GPCR/index.html.

1 Introduction

The G protein-coupled receptor (GPCR) is a major transmembrane (TM) protein type in the cellular membrane and plays critical roles in a wide variety of biological processes, including homeostasis modulation,[1] cell growth[2] and transporting of small molecules.[3] GPCRs are also important to humans. The human genome encodes thousands of GPCRs,[4] and, moreover, it is estimated that a large number of drugs in the market are designed to regulate the mechanism involved in GPCRs.[5] GPCRs are referred to as seven-TM receptors according to the fact that all existing GPCRs contain seven-TM α-helices with loops connecting them. Determination of the three-dimensional (3D) structures is a direct way to decipher their biological functions. Unfortunately, it is very time-consuming, and requires amazing funding and extensive efforts to obtain crystals of GPCRs. Compared with globular proteins, it is much more difficult to determine 3D structures of GPCRs. Due to experimental difficulties, the existing GPCR structures are very limited. For example, although there are more than 90 000 protein structures deposited in the PDB database,[6] the existing 3D structures of GPCRs in the PDB are only ∼100 at the time of March, 2014, and the non-redundant structures of GPCRs are much fewer. Considering the limitations of GPCR structural determination using wet experiments, it is of great need to develop accurate and high-throughput GPCR prediction methods.

Currently, there exist two major tasks to the computational study of GPCRs. One is to identify GPCRs from genome-wide sequences; the other is to locate TM regions of GPCR candidates. The low sequence similarities among some GPCRs, especially the existence of orphan GPCRs, hamper their identification by classical sequence-to-sequence alignments, such as BLAST.[7] Thus, the community needs specific GPCR prediction and identification programs. The past two decades have been witnessing exciting advances of a couple of such bioinformatics methods. In general, a sliding window centered at the target residue is excised and fed into the statistical learning algorithms to train the models. As one of the simplest forms, Gao and Chess developed a hydropathy-curve algorithm to detect proteins with seven hydrophobic stretches to screen potential GPCRs.[8] More sophisticated approaches such as the hidden Markov model (HMM)[9] and the Support Vector Machine (SVM)[10] are also used in the GPCR prediction. To develop the HMM-based methods, their designs of topologies of the HMMs, number of states and their connection need to be fixed in advance by taking insightful knowledge of known GPCRs. Once the topologies of HMMs are fixed, the protein sequence/structural data are used to train the probability of each transition of the HMMs. Phobius,[11] TMHMM,[12] GPCRHMM[13] and HMMTOP[14] are hidden Markov model-based methods for GPCR TM region prediction. PRED-GPCR by Papasaikas and his co-workers is a probabilistic method that uses family-specific HMMs to determine to which GPCR family a target sequence belongs.[15] The Jones group

[a] *Institute of Applied Genomics, School of Biological Sciences and Engineering, Fuzhou University, Fuzhou 350002, China. E-mail: yanrenxiang@fzu.edu.cn; Fax: +86 591 22866273; Tel: +86 591 22866273*

[b] *State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c4mb00272e

proposed a SVM-based method for TM protein topology prediction.[16] Meanwhile, a new set of conformational parameters for TM α helices was developed by Gromiha[17] and the parameters can be used to locate the TM regions of GPCRs. GPCRpred is also a SVM-based GPCR identification method by clustering GPCRs into different families.[18] The TM region prediction programs can be used for GPCR identification by scanning databases for proteins predicted to have seven-TM helices. GPCR identification and TM region prediction have been widely used in biological research. So for example, Nowling *et al.* screened GPCRs in the genomes of three insect vectors using an ensemble procedure;[19] Takeda and his co-workers identified a large number of potential GPCRs when searching the human proteome for proteins predicted to contain 6–8 TM helices.[20] Meanwhile, there are some other bioinformatics studies of GPCRs.[21–23] In general, the performance of statistical learning methods depends on the input features, learning algorithms and optimized parameters. Developers are required to carefully tune the parameters of training algorithms to obtain optimized performance.

In this work, we develop a predictor, which is capable of accurately identifying GPCRs from genomic sequences as well as predicting their TM segments. The TM regions of GPCRs are predicted by using sequence evolutionary profiles with the assistance of neural network learning. Moreover, considering that the secondary structure topologies of GPCRs are conserved, protein secondary structure-based methods for GPCR identification may make sense and we therefore develop such a method. Meanwhile, a novel profile-to-profile alignment algorithm is also developed to detect GPCRs. As that clearly pointed out by Chou in his review[24] as well as that in several closely related studies,[25–27] we can use the following procedure to establish a practical and reliable bioinformatics predictor. Firstly, build a model by using effective mathematical expressions that can truly reflect their intrinsic correlation with the target to be predicted, and then construct or select reliable benchmark datasets to train/test the models. Secondly, objectively evaluate the anticipated accuracy of the new model and compare it with community popular methods. Last but not the least, stand-alone programs and publicly available web servers for the models should be developed to facilitate researchers to use new methods. We will describe the procedure step-by-step in the following sections.

2 Materials and methods

2.1 Datasets

The benchmark datasets were constructed with the utilization of information in the PDB,[28] and UniProtKB.[29] Firstly, we downloaded 55 structurally known GPCRs from the PDB database with the timestamp of October, 2013. This dataset was named GPCR_PDB55. At the same time, the Swiss-Prot[29] database of UniProtKB (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz) was also downloaded in our local computers. We scanned all the sequences in the Swiss-Prot database and there are 2222 GPCRs showing high sequence similarity with sequences in the GPCR_PDB55 (BLAST *e*-value < 0.01). We further scanned the Swiss-Prot database and obtained 558 potential GPCRs, which were not similar to the 2222 proteins at the sequence level. Among the 558 proteins, 256 ones have already been included in the GPCRDB[30] database (http://www.gpcr.org/7tm/). Moreover, the remaining 302 proteins, which are seven-TM proteins, out of the 558 proteins are probably GPCRs. We found that there are some annotations, such as 'SIMILARITY: belongs to the G-protein coupled receptor 4 family', 'SIMILARITY: belongs to the G-protein coupled receptor Fz/Smo', 'DR Pfam; PF10326; 7TM_GPCR_Str; 1' and so on. Therefore, these 302 proteins are most likely to be GPCRs. Therefore, the 558 proteins were regarded as GPCRs in our benchmark. Further, we randomly selected 721 non-GPCRs membrane proteins from the Swiss-Prot database. These datasets were filtered by removing redundancies at 95% sequence identity. Finally, we obtained 1697 train (GPCR_TRAIN1697), 492 test (GPCR_TEST492) GPCRs and 504 non-GPCR membrane proteins (MEM_504). Details of removing redundancies are available in supplementary file 1 (ESI†). Meanwhile, we collected 2014 non-GPCR proteins, covering 2014 SCOP protein families, from the SCOPe[31] database, and the dataset of the 2014 proteins was named SCOP_2014. We used GPCR_TEST492 to benchmark various GPCR TM location methods. The performance of GPCR identification is assessed by methods' abilities in classification of GPCR/non-GPCR in the GPCR_TEST492, SCOP_2014 and MEM_504 datasets. The datasets are available at http://genomics.fzu.edu.cn/GPCR/dataset/. It should be clearly pointed out that the proteins in the GPCR_TEST492 share low similarity with the proteins of the GPCR_TRAIN1697 dataset at the sequence level (BLAST *e*-value > 0.01).

2.2 Trans-GPCR for TM region prediction

The Trans-GPCR is a neural network based method for TM region prediction. The neural network algorithm used in this work was implemented utilizing the Encog Java neural network framework,[32] which can be downloaded from https://code.google.com/p/encog-java/. The standard back propagation[33] and sigmoid activation function were used. We trained the Trans-GPCR using a similar way to PSIPRED.[34] Briefly, two feed-forward back-propagation neural networks were jointly used. In our work, the first neural network contains two hidden layers, whereas the second neural network only contains a single hidden layer. The nodes in both hidden layers of the first neural network were set to 250; meanwhile, the node number in the hidden layer of the second neural network was set to 70. The architectures and parameters of neural networks were optimized using the training dataset. The input features of the first neural network are evolutionary sequence profiles. The outputs of the first neural network are fed into the second neural network to refine the prediction. To obtain the sequence profiles, the target sequence is iteratively threaded through the NCBI[35] NR database for three repeats with an *e*-value cutoff of 0.001 for collecting multiple sequence alignments (MSAs) using PSI-BLAST.[36] The position specific scoring matrix/profile (PSSM) is generated by the option '-Q'. The position specific frequency matrix/profile (PSFM) is calculated from the generated MSA using Henikoff weight.[37] In the Henikoff weight scheme, a residue in each position is assigned a weight equal to l/(t + s), where t is the number of different residues in the column and s is the number of times the particular residue appears in the column. The position-based weights (*i.e.* Henikoff weights) are then added for each column and divided by the length of the sequence. Then, we use the following equation to calculate the PSFM profile of each residue from a MSA

$$f_{u,r} = \frac{\sum_{i=1}^{N} w_u^i \delta_{u,r}^i}{\sum_{i=1}^{N} w_u^i} \qquad (1)$$

where $f_{u,r}$ is the amino acid frequency of residue r at column u; N is the number of sequences in the MSA; w_u^i is the Henikoff weight for column u of sequence i; $\delta_{u,r}^i$ is set to 1 if sequence i has residue r in column u and 0, otherwise. For unaligned regions, only the target sequence itself is used to calculate the amino acid frequencies.

For each target residue, a sliding window containing $2n + 1$ residues long (*i.e.* window size = $2n + 1$) fragment profiles centered at the target residue is excised from the sequence profiles. The optimal window sizes of two neural networks were determined by performance in the training dataset and were set to 21. There are two sets of generated profiles, including PSFM and PSSM profiles. Using a similar way to Chen *et al.*,[38] we also compute the Shannon entropy for each residue as

$$\text{Entropy} = \sum_{r=1}^{20} -f_{u,r}\log\left(f_{u,r}\right) \qquad (2)$$

where $f_{u,r}$ is calculated using eqn (1); r is the rth residue type. Meanwhile, there are two-dimensional RW-GRMTP (relative weight of gapless real matches to pseudocounts), which are the last two columns in the PSSM profile, of each residue generated by PSI-BLAST. The RW-GRMTP represents the number of aligned residues in that position. The RW-GRMTP information is also used as training features. Considering some elements of the PSSM profile are negatives, we directly scale the values to the range of 0–1 by using the standard logistic function as

$$\frac{1}{1 + \mathrm{e}^{-x}} \qquad (3)$$

where x is the element value of the PSSM profile. Again, we also compute the entropy score for the PSSM profile. For the PSSM profile as well as the PSFM profile, there are 20 residue frequencies and an entropy value. Additionally, an extra unit per amino acid is used to indicate whether the residue spans either the N or C terminus of the protein chain. For a given 21-residue window, input features for the first neural network are window_size_1*(21+21+2+1), where 21 for PSSM, 21 for PSFM, 2 for RW-GRMTP and an additional unit to indicate whether the residue spans either the N or C terminus. The window_size_1 value of 21 is optimized by the performance in the training dataset. Using a similar way to Chou *et al.*,[24] we can denote the input features for position i of a protein as

$$\{[\text{PSSM}(i + s,j)],[E(i + s)],[\text{PSFM}(i + s,j)],[\vec{E}(i + s)], \text{RW-GRMTP}(i + s)\}, j \in [0,20], s \in [-n,n] \qquad (4)$$

where PSSM($i + s,j$) is for the scaled PSSM profile at the position $i + s$; j ranges from 0 to 19, in which [0,19] represents 20 amino acids and one additional bit that is used to indicate whether the residue spans either the N or C terminus of the protein chain; s is a shift value and its value ranges from $-n$ to n (*i.e.* window size). $E(i + s)$ is the Shannon entropy for position $i + s$ calculated using the scaled PSSM profile at position $i + s$. Similarly, PSFM($i + s,j$) is for the PSFM profile at the position $i + s$; [$\vec{E}(i + s)$] is the Shannon entropy for position $i + s$ calculated using the PSFM profile at position $i + s$; RW-GRMTP($i + s$) is the RW-GRMTP values (*i.e.* relative weight of gapless real matches to pseudocounts) at position $i + s$.

The feature numbers for the second neural network are window_size_2*(2+1), where 2 denotes the outputs (*e.g.* prediction scores of TM/non-TM) of the first neural network and an additional unit to indicate whether the residue spans either the N or C terminus. The window_size_2 value of 21 is optimized using the same way as that of window_size_1. The average length of the TM regions is 22 in our training dataset. Meanwhile, lengths of the loops connecting the TM segments are diverse. Based on this observation, we transform the prediction of orphan residues, assigning a TM (non-TM) residue to the non-TM (TM) region if its neighbor six residues (*i.e.* ±3 positions) are non-TM (TM).

2.3 GPCR identification

2.3.1 Trans-GPCR for GPCR identification. Furthermore, the Trans-GPCR not only predicts the TM regions of GPCRs but can also identify GPCRs. For a target sequence, the Trans-GPCR determines whether it is a GPCR by the following equation

$$\text{TransGPCR\_Score} = \sum_{i=1}^{N} \max(NN(M)) - NN(-), 0) \qquad (5)$$

where $NN(M)$ and $NN(-)$ are the TM and non-TM prediction scores of residue i by two output nodes of the second neural network in the Trans-GPCR method; N is the length of target protein. We use max($NN(M)$ − $NN(-)$,0) to ensure that only predicted TM regions are summed (*i.e.* positive values). Here, we use a reliable parameter for position i of target protein as

$$\text{residue\_reliable}(i) = \text{abs}(NN(M) - NN(-)) \qquad (6)$$

where residue_reliable(i) is a reliable index; abs is the absolute mathematic function; $NN(M)$ and $NN(-)$ are defined in eqn (5). residue_reliable(i) ranges [0–1], where a higher score corresponds to a more reliable prediction for residue i. It should be clearly pointed out that the parameter TransGPCR_Score is to determine whether a protein is a GPCR, whereas residue_reliable(i) is a position-specific reliability index of prediction for position i of target protein.

2.3.2 SSEA-GPCR for GPCR identification. Here, we also develop a GPCR identification algorithm by using secondary structure element alignment (SSEA). Since protein secondary structural topologies of GPCRs are more conserved than single sequences, SSEA is therefore able to identify GPCRs.

The SSEA-GPCR method searches a target sequence against GPCR and non-GPCR databases. In this process, the top *i* SSEA similarity scores between GPCRs (non-GPCRs) are recorded (*i.e.* $\text{SSEA}_{\text{max\_gpcr\_}i}$ and $\text{SSEA}_{\text{max\_non\_gpcr\_}i}$). In the SSEA algorithm, the secondary structural string for each sequence is converted into secondary structure elements such that 'H' represents a helix element, 'E' denotes a strand element, and 'C' stands for a coil element. Meanwhile, the predicted secondary structural string was shortened and the length of each element was retained for the scoring of SSEA. Here, the Needleman–Wunsch global alignment algorithm[39] was used with the gap penalties set to zeros. The alignment score of SSEA between two secondary structure elements with lengths L_i and L_j is defined as

$$\text{Score}(i,j) = \begin{cases} \min\left(L_i, L_j\right) & \text{Match between two identical elements} \\ 0.5 \times \min\left(L_i, L_j\right) & \text{Match between }\alpha\text{-helix}/\beta\text{-strand and coil} \\ 0 & \text{Match between }\alpha\text{-helix and }\beta\text{-strand} \end{cases} \tag{7}$$

where $\min(L_i, L_j)$ stands for the minimal length between L_i and L_j. The normalized SSEA alignment score is obtained by dividing by the length of the target sequence. Details of the SSEA algorithm can be referred from its original developer[40] or our previous work.[41] For a target sequence, the SSEA_gpcr prediction score is calculated using a simple *K*-nearest algorithm as

$$\text{SSEA\_gpcr} = \frac{\sum_{i=1}^{K} \text{SSEA}_{\text{gpcr\_top\_}i} - \sum_{i=1}^{K} \text{SSEA}_{\text{non\_gpcr\_top\_}i}}{K} \tag{8}$$

where $\text{SSEA}_{\text{gpcr\_top\_}i}$ and $\text{SSEA}_{\text{non\_gpcr\_top\_}i}$ are the top *i* prediction scores of searching target protein against GPCR and non-GPCR databases; the value of *K* is primarily optimized and set to 10. Here, the GPCRs in the training dataset are used as the GPCR database to calculate $\text{SSEA}_{\text{max\_gpcr}}$. Meanwhile, we collected 3836 non-GPCR proteins, which cover 1061 folds, 1713 superfamilies and 3836 families, as a non-GPCR database from the SCOPe database.[31] This dataset was named non-GPCRlib_3836. The nonGPCRlib_3836 is used when calculating $\text{SSEA}_{\text{max\_non\_gpcr}}$. The proteins in the SCOP_2014 dataset, which has been described in the Datasets section, share low similarity with proteins in the non-GPCR database (*i.e.* nonGPCRlib_3836) at the sequence level (BLAST *e*-value $>$ 0.01).

2.3.3 PPA-GPCR for GPCR identification. GPCRs constitute a large superfamily of proteins.[13] Therefore, profile-to-profile alignment, which represents one of the useful methods to detect distant homologs, should be effective to identify potential GPCRs. Similar to the SSEA-GPCR method, the Needleman–Wunsch global alignment algorithm is also used and the penalties for ending gaps are set as zeros. The scoring function for profile-to-profile alignment is

$$\text{Score}(i,j) = \text{PF}(i,j) + w_1\text{SS}(i,j) + \text{shift} \tag{9}$$

where PF(i,j) is an evolutionary profiles-based term. Evolutionary profiles are generated from MSAs, which represent the divergence of proteins in the same family, and contain important information to infer the protein features. The MSAs are obtained using the same way as that in the Trans-GPCR. The values of gap openning, gap extension, w_1 and shift were obtained by maximum of the sequence alignments to structural alignments[42] of all-to-all pair-wises for the 55 structurally known GPCRs in the GPCR_PDB55 dataset. The values of gap openning, gap extension, w_1 and shift were set to −7.1, −0.56, 0.7 and −0.9. The profile similarity score is

$$\text{PF}(i,j) = \frac{1}{2}\sum_{k=1}^{20}\left(\text{PSFM}(i,k)_q\text{PSSM}(j,k)_t + \text{PSFM}(j,k)_t\text{PSSM}(i,k)_q\right) \tag{10}$$

where PSFM$(i,k)_q$ represents the frequency of the *k*th amino acid at the *i*th position of the PSFM profile for target protein; PSSM$(j,k)_t$ denotes the *k*th amino acid at the *j*th position of the PSSM profile for the template. Similarly, PSSM$(j,k)_t$ represents the frequency of the *k*th amino acid at the *j*th position of the PSFM profile for the template; PSSM$(i,k)_q$ denotes the *k*th amino acid at the *i*th position of the PSSM profile for target protein. In our method, the similarity score for each pair of secondary structure profile columns is defined as Pearson's correlation coefficient between them as

$$\text{SS}(i,j) = \frac{3\sum_{k=1}^{3} Q_{i,k}T_{j,k} - \sum_{k=1}^{3} Q_{i,k}\sum_{k=1}^{3} T_{j,k}}{\sqrt{3\sum_{k=1}^{3} Q_{i,k}{}^2 - \left(\sum_{k=1}^{3} Q_{i,k}\right)^2}\sqrt{3\sum_{k=1}^{3} T_{j,k}{}^2 - \left(\sum_{k=1}^{3} T_{j,k}\right)^2}} \tag{11}$$

where $Q_{i,k}$ is the possibility of *k*th (*i.e.* k = 1, 2, 3 corresponding to α-helix (H), β-strand (E), and coil (C), respectively) secondary structure type at *i*th position of the target sequence. $T_{j,k}$ is the possibility of *k*th secondary structure type at *j*th position of the template sequence. The prediction possibilities of protein secondary structure are obtained by using PSIPRED. Similar to the SSEA-GPCR, the normalized PPA-GPCR alignment score is also obtained by being divided by length of the target sequence. Moreover, the estimated significant Zscore of PPA-GPCR alignment scores should be calculated. We use the SCOPe_1187 dataset, which is constructed by randomly selecting one protein of each fold from the SCOPe database, as a reference database to calculate mean and standard deviation of random scores. The Zscore is calculated as

$$\text{Zscore} = \frac{\text{raw} - \text{mean}}{\text{std}} \tag{12}$$

where raw is the alignment score between a target and a specific template; mean and std are the average and standard deviation of scores aligning the target sequence to the 1187 proteins in the SCOPe_1187 dataset. There are two Zscores for any pair of target-template alignments. Here, we use a symmetrical Zscore similar to FFAS-3D[43] as

$$\mathrm{Zscore}(q,t) = \mathrm{ave}(\mathrm{Zscore}_q,\mathrm{Zscore}_t) \tag{13}$$

where Zscore_q and Zscore_t are the Zscores of the target and template proteins by searching the SCOPe_1187 database using eqn (12). Here, we use the average of Zscore_q and Zscore_t as the final value of the calibrated score. Note that Zscore(q,t) is symmetrical with respect to two proteins. We also tested the minimum and maximum of the two Zscores, but the performance cannot be improved. For each target, we search it against a GPCR database, which is GPCR_TRAIN1697 in our benchmark. The maximum Zscore(q,t) of the target and the templates (*i.e.* 1697 pair-wise alignment scores) in the GPCR_TRAIN1697 database is recorded and is named PPA_gpcr in this paper. Confidence intervals (CI) of PPA_gpcr are computed using the common assumption of a normal distribution by the following as

$$\left[\mu - Z\frac{\mathrm{SD}}{\sqrt{n}},\quad \mu + Z\frac{\mathrm{SD}}{\sqrt{n}}\right] \tag{14}$$

where μ and SD are the mean and standard deviation of PPA_gpcr scores; n is the sample size; Z is the critical value and the value of Z is 1.96 in a 95% confidence level.

2.3.4 Combined methods. The combined methods can be constructed by using complementary algorithms with improved performance. When combining the top four methods (HMMTOP, TMHMM, Phobius and Trans-GPCR) for TM/non-TM region prediction, we use two bits to denote their prediction for each residue (*i.e.* [1, 0] for TM and [0, 1] for non-TM predictions). To combine the four methods, the corresponding bit values are simply added. For example, [1, 0], [1, 0], [1, 0] and [0, 1] are added and the result is [3, 1]. The combined prediction for a residue is TM if the value of the first bit is bigger than that of the second bit, and non-TM, otherwise. The combined method for TM/non-TM prediction is named TM-Combined in this paper. Similarly, we also combined the methods (Trans-GPCR, SSEA-GPCR and PPA-GPCR) for GPCR identification (Iden-Combined) using a weighted score as

$$\text{Iden-Combined} = w_1\text{PPA\_gpcr} + w_2\text{Trans\_gpcr} + w_3\text{SSEA\_gpcr} \tag{15}$$

where Iden-Combined is the combined prediction score; w_1, w_2 and w_3 are weighted to balance the three terms. Considering the value ranges of the three terms, the values of w_1, w_2 and w_3 are primarily optimized and set to 0.1, 0.0067 and 1, respectively.

2.3.5 Amino acid distribution of TM/non-TM regions. It is also interesting to mine the amino acid distribution of TM/non-TM regions in the GPCRs. The formula for calculating the composition of ith residue is defined as

$$\mathrm{composition}(i) = \frac{\sum_{k=1}^{N}\delta_k}{N} \tag{16}$$

where i stands for composition of ith residue; δ_k is set to 1 if the position k of the sequence is ith residue and 0, otherwise; N is the total number of residues in the TM/non-TM regions.

2.4 Performance assessment

When the benchmark is performed over the test dataset in the TM/non-TM region prediction, the overall performance of different methods is evaluated with respect to four parameters: accuracy (Ac), sensitivity (Sn), specificity (Sp) and Matthew correlation coefficient (Mcc). The TM (non-TM) residues of GPCRs are considered positives (negatives). The equations for these parameters are as follows

$$\mathrm{Ac} = \frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{fn}+\mathrm{tn}+\mathrm{fp}} \tag{17}$$

$$\mathrm{Sn} = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}} \tag{18}$$

$$\mathrm{Sp} = \frac{\mathrm{tn}}{\mathrm{tn}+\mathrm{fp}} \tag{19}$$

$$\mathrm{Mcc} = \frac{\mathrm{tp}\times\mathrm{tn}-\mathrm{fp}\times\mathrm{fn}}{\sqrt{(\mathrm{tp}+\mathrm{fp})(\mathrm{tp}+\mathrm{fn})(\mathrm{tn}+\mathrm{fn})(\mathrm{tn}+\mathrm{fp})}} \tag{20}$$

where tp, fp, fn and tn are the numbers of true positives, false positives, false negatives and true negatives, respectively. The performance of GPCR identification can be measured by receiver operating characteristic (ROC) curves.[44] The ROC curves plot the true-positive rate (instances) as a function of false-positive rate (instances) for all possible thresholds of prediction scores by various methods. The set of four equations (eqn (17)–(20)) is used for single-label systems. For multi-label systems, which are more frequent in system biology,[45,46] a completely different set of metrics as defined in ref. 47 is needed.

3 Results and discussion

3.1 The performance of TM region prediction

Among the resulting measures, Ac and Mcc are the most comprehensive parameters to assess the prediction performance. The neural network model of the Trans-GPCR was intensively trained on the GPCR_TRAIN1697 dataset and generated the results of Ac = 0.940 and Mcc = 0.877. Further, the performance of the TM region location was tested on the GPCR_TEST492 dataset. HMMTOP, TMHMM, Memast and Phobius programs were installed in our local computers and the proteins were directly fed into them. The prediction results of the TM regions for various methods were summarized in Table 1. HMMTOP, TMHMM, Memast and Phobius generated Ac (Mcc) scores of 0.927 (0.804), 0.934 (0.823), 0.912 (0.766) and 0.935 (0.826), respectively. The Trans-GPCR generated slightly lower Ac and Mcc values than that of HMMTOP, TMHMM and Phobius. Although these methods were benchmarked on the same dataset, it should be pointed out that proteins in the test dataset of the Trans-GPCR share low similarity with the proteins in the training dataset (BLAST e-value $>$ 0.01). Meanwhile, the TM regions of

Table 1 Performance of TM region prediction of various methods on datasets

Method[a]	Ac	Sn	Sp	Mcc
Benchmark result on GPCR_TRAIN1697				
HMMTOP	0.910	0.896	0.919	0.814
TMHMM	0.907	0.890	0.920	0.809
Memsat	0.892	0.906	0.882	0.780
Phobius	0.903	0.894	0.909	0.801
Trans-GPCR[b]	0.940	0.930	0.948	0.877
TM-Combined	0.935	0.943	0.930	0.867
Benchmark result on GPCR_TEST492				
HMMTOP	0.927	0.865	0.947	0.804
TMHMM	0.934	0.874	0.954	0.823
Memsat	0.912	0.848	0.932	0.766
Phobius	0.935	0.884	0.951	0.826
Trans-GPCR	0.923	0.833	0.952	0.791
TM-Combined	0.935	0.901	0.946	0.828

[a] All residues of the test dataset were used to count measures of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). [b] Trans-GPCR was intensively trained on the GPCR_TRAIN1697 dataset. Proteins in the GPCR_TEST492 dataset share low similarity with proteins in GPCR_TRAIN1697 (BLAST *e*-value > 0.01). Therefore, benchmark of the Trans-GPCR on the GPCR_TRAIN1697 dataset does not make a lot of sense. We just want to know how much performance decreases when tested the Trans-GPCR on the GPCR_TEST492 compared with that of GPCR_TRAIN1697.

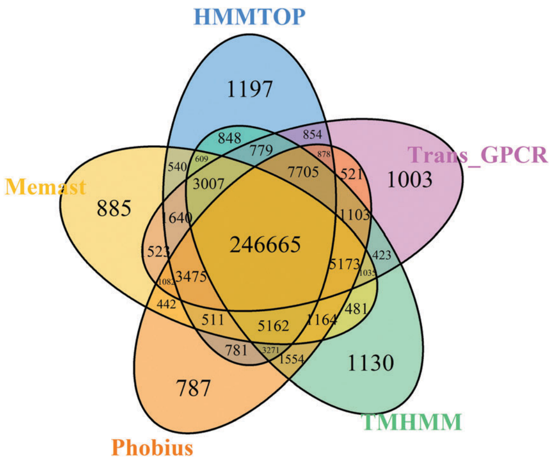

Fig. 1 Venn diagram showing the complementarity of various methods. For the samples correctly distinguished by two or more methods, they correspond to the number in the overlapped regions.

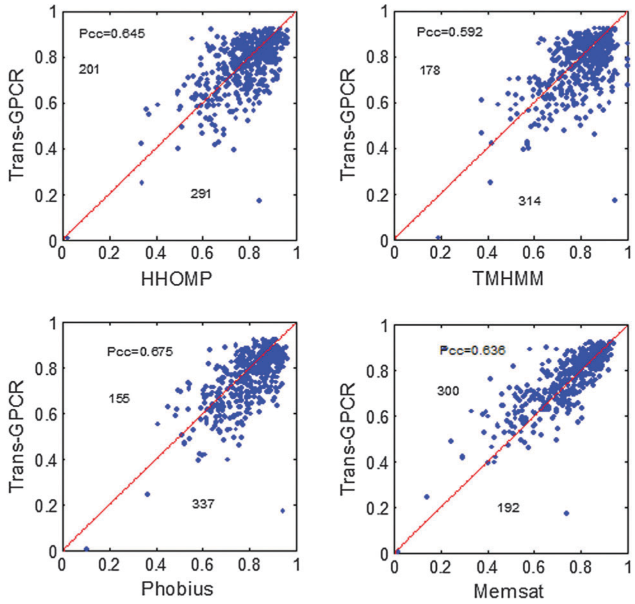

Fig. 2 All-to-all comparisons of Mcc scores between methods on the GPCR_TEST492 dataset. The number in each panel denotes the number of proteins/points in upper and lower triangles, respectively. Meanwhile, Pearson's correlation coefficient (Pcc) values are also given.

some proteins in the Swiss-Prot database are annotated by using TMHMM, Memsat and Phobius (see http://www.uniprot.org/manual/transmem for details). The complementarity of these methods is given in Fig. 1 using the VennDiagram package.[48] For example, HMMTOP, TMHMM, Phobius, Memast and Trans-GPCR methods correctly distinguish 1197, 1130, 787, 885 and 1003 residues that cannot be correctly distinguished by other methods. In Fig. 2, two Mcc values of each protein by two methods correspond to a point. We calculated their statistical significances using Student's *t*-test (Table 2). The *p*-values of Mcc scores for the methods were 0.01 lower although both HMMTOP, TMHMM and Phobius were HMM-based algorithms. The different and complementary methods can be combined to generate improved performance. This is demonstrated by the TM-combined method, which generated the highest Ac (0.935) and Mcc (0.828) values in the GPCR_TEST492 dataset. The increase in sensitivity using TM-combined may be ascribed to that TM-combined measure is a consensus method by considering scores of the top four methods. But the TM-combined method did not generate higher Mcc values in the GPCR_TRAIN1697, and this may be because the proteins of GPCR_TRAIN1697 were used to train the Trans-GPCR method. Therefore, it is very difficult for the TM-combined method to generate better performance. Meanwhile, we also calculated Pearson's correlation coefficient (Pcc) between them (Fig. 2). As seen from the data above, we can know that the benchmarked five methods were significantly different (p-value < 0.01). The most significant methods were TMHMM and Memsat (p-value $< 2.2 \times 10^{-16}$). To better understand the prediction error generation, it is important to know the misclassification rates between TM/non-TM. As can be seen from Table 3, the largest misclassification state is TM to non-TM, which is consistent for the five predictors.

3.2 Benchmark of GPCR identification

The performance of GPCR identification was compared *via* ROC analysis. As can be seen from Fig. 3, the PPA-GPCR generated the best performance, resulting in an AUC score of 0.990. The Trans-GPCR and the SSEA-GPCR generated AUC scores of 0.978 and 0.955. Because the performance at low false positive rates is more important in real-world application, therefore, we paid more attention to the comparison of different methods' performance at <1% false positive rates (Fig. 3B and Table 4). As shown in Table 4, the SSEA-GPCR correctly recognized 193 GPCRs before including 36 false positives, whereas the Trans-GPCR can detect 306 GPCRs. The distribution of profile-to-profile alignment

Table 2 Student's t-test p-values of the five methods of Mcc scores

Method	HMMTOP	TMHMM	Memsat	Phobius	Trans-GPCR
HMMTOP		3.594×10^{-05}	9.885×10^{-13}	1.443×10^{-06}	3.000×10^{-4}
TMHMM			2.200×10^{-16}	0.697	4.947×10^{-13}
Memsat				2.200×10^{-16}	1.588×10^{-08}
Phobius					2.200×10^{-16}
Trans-GPCR					

Table 3 Misclassification rates in the benchmark dataset

Native	Predicted	HMMTOP	TMHMM	Memsat	Phobius	Trans-GPCR
M[a]	—[a]	0.134	0.126	0.151	0.115	0.166
—	M	0.052	0.045	0.067	0.048	0.047

[a] Here 'M' and '—' represent transmembrane and non-transmembrane residues. The misclassification rate is calculated using equation $E(i)/N(i)$, where $E(i)$ is the number of misclassified state i and $N(i)$ is the total number of state i in the benchmark dataset.

Table 4 ROC table (≤36 false positives) for different methods

Receiver operator characteristics (≤36 false positives[a])								
Methods	12	16	20	24	28	32	36	Auc[b]
Trans-GPCR	133	165	233	266	289	293	306	0.978
SSEA-GPCR	120	139	160	173	188	192	193	0.955
PPA-GPCR	319	346	354	356	374	382	385	0.990
Iden-Combined	343	381	388	411	431	444	461	0.993

[a] Here, false positives correspond to those non-GPCRs predicted as GPCRs. [b] The Auc score represents the corresponding area under a ROC curve.

scores (*i.e.* PPA_gpcr measure) in the three types of proteins (*i.e.* GPCR, non-GPCR membrane proteins, and globular proteins) is presented in Fig. 4. The confidence interval (CI) values of PPA_gpcr for GPCRs, non-GPCR membrane proteins and globular proteins were [13.53, 14.47], [6.77, 7.27] and [2.48, 2.61] (Table 5), respectively. There is no overlap among these intervals, suggesting that the PPA-GPCR method can be used to distinguish GPCRs in a reasonable result. The PPA-GPCR detects more GPCRs (385 hits) than the Trans-GPCR and SSEA-GPCR methods at the same cutoff of false positives. When

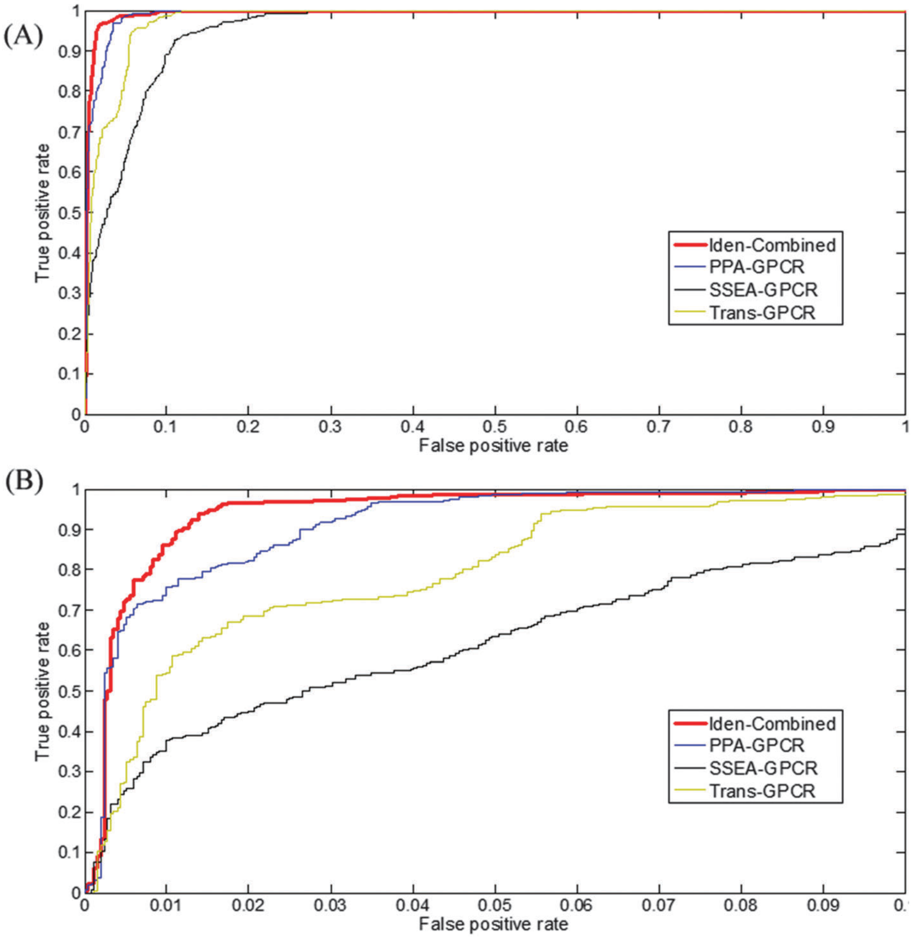

Fig. 3 Comparison of ROC curves for different methods.

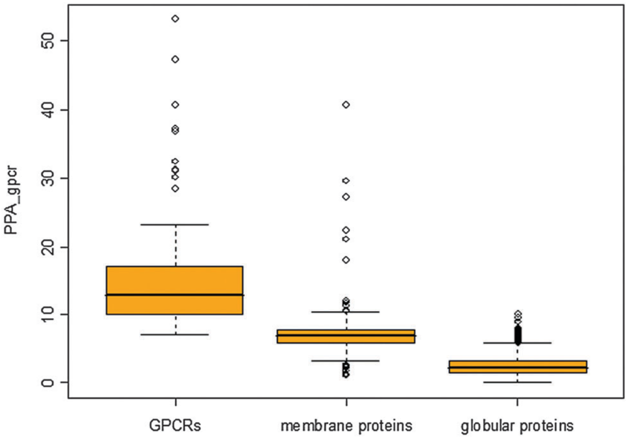

Fig. 4 Boxplot of PPA_gpcr scores in the three types of proteins. Here, membrane proteins denote the non-GPCR membrane proteins.

Table 5 Mean, standard deviation and confidence intervals (CI) at a 95% level

Methods	Mean	Standard deviation	CI
GPCRs	14.00	5.32	[13.53, 14.47]
Membrane proteins	7.02	2.88	[6.77, 7.27]
Globular proteins	2.54	1.44	[2.48, 2.61]

we used the Iden-Combined measure to identify GPCRs it identifies the most GPCRs at the 1% false positive rate (Fig. 3 and Table 4). Despite the lack of sequence homology between some GPCRs, all GPCRs share similar conserved secondary structural topologies and have the homologous relationships. Therefore, the SSEA-GPCR and the PPA-GPCR should be effective to detect them. Our benchmark results also support this point of view.

3.3 Significances of prediction scores and decision making

It is very necessary to estimate the significances of predictions when developing new probabilistic models. We estimated the significant scores of the Trans-GPCR, the SSEA-GPCR and the PPA-GPCR from the test dataset. In the Trans-GPCR method, we designed two output nodes in two neural networks to represent the prediction scores of TM/non-TM regions. The difference of the two nodes of the second neural network for target residue is represented by the measure residue_reliable(i). The larger the residue_reliable(i) score is, the more significant and reliable the target residue is. In our benchmark result, if the residue_reliable(i) > 0.911, it can generate a prediction result with less than a 1% false positive rate. Meanwhile, we also tested the Trans-GPCR_Score, SSEA_gpcr and PPA_gpcr scores, which are the parameters to identify GPCRs, in the benchmark dataset to obtain their reliable cutoffs. In our benchmark, if Trans-GPCR_Score is larger than 84.834, the prediction result is at less than a 5% false positive rate. At the same false positive rate control, SSEA_gpcr and PPA_gpcr should be larger than 0.094 and 7.545, respectively. The prediction scores and corresponding false positive rates were summarized in Table 6. A question should be discussed here is that how to determine whether a protein is a GPCR using these methods. We suggest combining the three methods to make decisions. If proteins are predicted to have less than 1% false positive rates by the three methods, the proteins should be regarded as candidates for being GPCRs with high confidences. It is easy to distinguish GPCRs and globular proteins. However, it may be difficult to distinguish GPCRs from some non-GPCR membrane proteins according to the fact that some of them have similar topologies and exist in the similar biological environments. For such cases, maybe researchers can use the TM helices number and PPA_gpcr scores to determine whether a protein is a GPCR or not. Alternatively, users can resort to combined methods (*i.e.* TM-Combined and Iden-Combined) to make decisions. If the Iden-Combined score is higher than 1.589 by the combined method, the prediction is at less than a 1% false positive rate. Some hard targets may need further literature survey. In our web server (see supplementary file 2 (ESI†) for details), we provide the prediction scores by the Trans-GPCR, the SSEA-GPCR, the PPA-GPCR and Iden-Combined for each job. To provide a real application example, we conduct our method on the proteome of Homo sapiens (see supplementary file 3 (ESI†) for details).

Table 6 Cutoffs of different methods at 95% and 99% confidence levels

Methods	95% level	99% level
TransGPCR_Score[a]	84.834	112.295
residue_reliable[a]	0.000	0.911
SSEA_gpcr	0.094	0.139
PPA_gpcr	7.545	9.664
Iden-Combined	1.354	1.589

[a] TransGPCR_Score is a measure to determine whether a protein is GPCR, whereas residue_reliable(i) is a parameter to describe the reliability in position i of a protein (*i.e.* TM or non-TM residue).

3.4 Lengths and amino acid distribution of TM/non-TM regions

We calculated the mean lengths for TM/non-TM regions, but did not find significant differences between TM segments of different GPCRs in the training dataset. The lengths of TM helices are in the range of 6 to 30 amino acids and the average length of TM helices is 22. The Beta-1 adrenergic receptor (Swiss-Prot entry: Q9TT96) contains the longest (30 residues) TM segment in the sixth TM in our training dataset. Although Q7P0A1, Q6BKW6 and Q60880 proteins contain TM segments longer than 30, their segments were annotated as two independent parts. For example, the 220–261 of protein Q7P0A1 is the TM region. But 220–240 and 241–261 of this long region were annotated as two independent parts in the Swiss-Prot database. For such regions, we also counted them as two segments. Meanwhile, putative olfactory receptor 10J6 protein (Swiss-Prot entry: Q8NGY7) contains the minimum length TM regions (6 residues) in our training dataset. The length of loops connecting TM helices is more diverse. The protein Q4LBB6 contains the longest loop (843 residues), which connects the fifth and sixth TMs.

The amino acid compositions in the TM, non-TM regions and their differences are shown in Fig. 5, in which the similarities and differences of the 20 amino acid residues in the

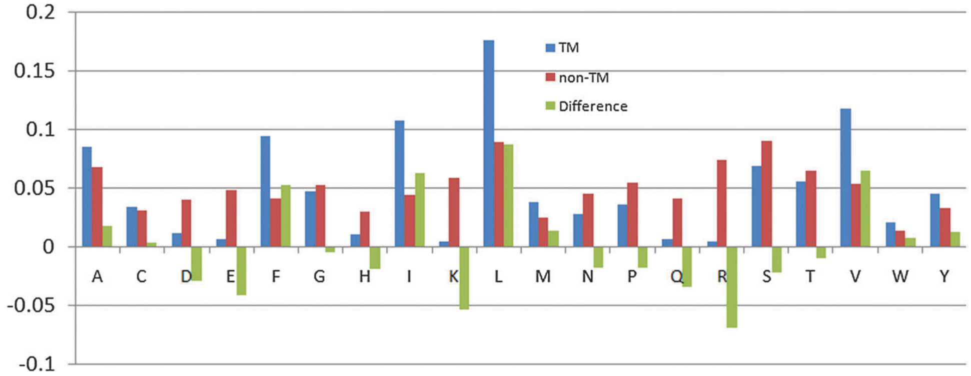

Fig. 5 Amino acid composition of the 20 amino acid residues in TM regions (blue bars), non-TM regions (red bars) and differences between them (green bars).

TM/non-TM regions were brought out. Residues with positive scores of differences suggest their preference in the TM regions while those with negative scores show their dominance in the non-TM regions. As can be seen from Fig. 5, the most differences of amino acids are R (ARG), K (LYS), E (GLU), L (LEU), V (VAL) and I (ILE). Among them, L, V and I are aliphatic amino acids; R, K, and E are charged amino acids. Interestingly, L, V and I are enriched in the TM regions whereas R, K, and E are enriched in non-TM regions. Meanwhile, C (CYS) and G (GLY) show subtle difference in the amino acid composition. The amino acid composition differences in the TM and non-TM regions can be regarded as conformational parameters of amino acids in TM regions. Similarly, Gromiha developed a set of such conformational parameters in a different way in 1999. Pearson's correlation coefficient between our parameters and those developed by Gromiha is 0.932 (see supplementary file 4 (ESI†) for details), suggesting both sets of parameters can be used to represent the preferences of amino acids in the TM regions although they are calculated using different ways. Meanwhile, we also tested the performance of secondary structure prediction by PSIPRED on GPCRs, and PSIPRED shows an overall Q3 accuracy of 76.6% (see supplementary file 5 (ESI†) for details).

4 Conclusions

In this work, we developed a practical predictor for GPCR TM region prediction (Trans-GPCR), and GPCR identification (Trans-GPCR, SSEA-GPCR and PPA-GPCR). Our predictor has been intensively benchmarked and its favorable performance has been demonstrated in the real application.

Objectively speaking, our predictor has strengths and limitations compared to some other methods. The most obvious strength is its potential application to identify GPCRs that show little sequence similarity to known GPCRs but with similar topologies or homologous relationships. However, the qualities of both GPCR identification and their TM region location are relied on the input profiles, and it may create problems if there are false homologous sequences imbedded in the MSAs that are used to calculate sequence profiles. This is one obvious limitation/disadvantage of our predictor.

Anyway, our server should be useful based on its performance in the benchmark. Although our predictor is a solely computational tool, we also hope that the development of such novel methods will be helpful to accelerate the exploration of the sequence-structure-function landscape in GPCRs.

Acknowledgements

We would like to thank Shaoyu Su in Fujian Science and Technology Information Institute for critical reading of the manuscript. This work was supported by Start-Up Fund of Fuzhou University (510046), National Natural Science Foundation of China (31301537) and Science Development Foundation of Fuzhou University (2013-XY-17).

References

1 G. G. Hazell, C. C. Hindmarch, G. R. Pope, J. A. Roper, S. L. Lightman, D. Murphy, A. M. O'Carroll and S. J. Lolait, *Front. Neuroendocrinol.*, 2012, **33**, 45–66.
2 R. T. Dorsam and J. S. Gutkind, *Nat. Rev. Cancer*, 2007, **7**, 79–94.
3 F. Giordano, S. Simoes and G. Raposo, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 11906–11911.
4 D. K. Vassilatis, J. G. Hohmann, H. Zeng, F. Li, J. E. Ranchalis, M. T. Mortrud, A. Brown, S. S. Rodriguez, J. R. Weller, A. C. Wright, J. E. Bergmann and G. A. Gaitanaris, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 4903–4908.
5 J. P. Overington, B. Al-Lazikani and A. L. Hopkins, *Nat. Rev. Drug Discovery*, 2006, **5**, 993–996.
6 H. M. Berman, *Acta Crystallogr.*, 2008, **64**, 88–95.
7 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.
8 Q. Gao and A. Chess, *Genomics*, 1999, **60**, 31–39.
9 L. Rabiner, *Proc. IEEE*, 1989, **77**, 257–286.
10 L. C. Chang CC, *Computer Program*, 2001.
11 L. Kall, A. Krogh and E. L. Sonnhammer, *J. Mol. Biol.*, 2004, **338**, 1027–1036.

12 A. Krogh, B. Larsson, G. von Heijne and E. L. Sonnhammer, *J. Mol. Biol.*, 2001, **305**, 567–580.
13 M. Wistrand, L. Kall and E. L. Sonnhammer, *Protein Sci.*, 2006, **15**, 509–521.
14 G. E. Tusnady and I. Simon, *J. Mol. Biol.*, 1998, **283**, 489–506.
15 P. K. Papasaikas, P. G. Bagos, Z. I. Litou and S. J. Hamodrakas, *SAR QSAR Environ. Res.*, 2003, **14**, 413–420.
16 T. Nugent and D. T. Jones, *BMC Bioinf.*, 2009, **10**, 159.
17 M. M. Gromiha, *Protein Eng.*, 1999, **12**, 557–561.
18 M. Bhasin and G. P. Raghava, *Nucleic Acids Res.*, 2004, **32**, W383–W389.
19 R. J. Nowling, J. L. Abrudan, D. A. Shoue, B. Abdul-Wahid, M. Wadsworth, G. Stayback, F. H. Collins, M. A. McDowell and J. A. Izaguirre, *Parasites Vectors*, 2013, **6**, 150.
20 S. Takeda, S. Kadowaki, T. Haga, H. Takaesu and S. Mitaku, *FEBS Lett.*, 2002, **520**, 97–101.
21 D. W. Elrod and K. C. Chou, *Protein Eng.*, 2002, **15**, 713–715.
22 K. C. Chou, *J. Proteome Res.*, 2005, **4**, 1413–1418.
23 X. Xiao, J. L. Min, P. Wang and K. C. Chou, *PLoS One*, 2013, **8**, e72234.
24 K. C. Chou, *J. Theor. Biol.*, 2011, **273**, 236–247.
25 W. Chen, P. M. Feng, H. Lin and K. C. Chou, *Nucleic Acids Res.*, 2013, **41**, e68.
26 Y. Xu, X. J. Shao, L. Y. Wu, N. Y. Deng and K. C. Chou, *PeerJ*, 2013, **1**, e171.
27 X. Xiao, J. L. Min, P. Wang and K. C. Chou, *J. Theor. Biol.*, 2013, **337**, 71–79.
28 J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter and E. E. Abola, *Acta Crystallogr.*, 1998, **54**, 1078–1084.
29 E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider and A. Bairoch, *Methods Mol. Biol.*, 2007, **406**, 89–112.
30 F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. E. Cohen and G. Vriend, *Nucleic Acids Res.*, 2003, **31**, 294–297.
31 N. K. Fox, S. E. Brenner and J. M. Chandonia, *Nucleic Acids Res.*, 2014, **42**, D304–D309.
32 J. Heaton, 2008, 1–429.
33 E. R. David, E. H. Geoffrey and J. W. Ronald, in *Neurocomputing: foundations of research*, ed. A. A. James and R. Edward, MIT Press, 1988, pp. 696–699.
34 D. T. Jones, *J. Mol. Biol.*, 1999, **292**, 195–202.
35 K. D. Pruitt, T. Tatusova, W. Klimke and D. R. Maglott, *Nucleic Acids Res.*, 2009, **37**, D32–D36.
36 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
37 S. Henikoff and J. G. Henikoff, *J. Mol. Biol.*, 1994, **243**, 574–578.
38 Z. Chen, Y. Wang, Y. F. Zhai, J. Song and Z. Zhang, *Mol. BioSyst.*, 2013, **9**, 2213–2222.
39 S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.*, 1970, **48**, 443–453.
40 T. Przytycka, R. Aurora and G. D. Rose, *Nat. Struct. Biol.*, 1999, **6**, 672–682.
41 R. X. Yan, Z. Chen and Z. Zhang, *BMC Bioinf.*, 2011, **12**, 76.
42 Y. Zhang and J. Skolnick, *Nucleic Acids Res.*, 2005, **33**, 2302–2309.
43 D. Xu, L. Jaroszewski, Z. Li and A. Godzik, *Bioinformatics*, 2014, **30**, 660–667.
44 T. Fawcett, *Pattern Recogn. Lett.*, 2006, **27**, 861–874.
45 K. C. Chou, Z. C. Wu and X. Xiao, *PLoS One*, 2011, **6**, e18258.
46 K. C. Chou, Z. C. Wu and X. Xiao, *Mol. BioSyst.*, 2012, **8**, 629–641.
47 K. C. Chou, *Mol. BioSyst.*, 2013, **9**, 1092–1100.
48 H. Chen and P. C. Boutros, *BMC Bioinf.*, 2011, **12**, 35.