Molecular BioSystems

METHOD

View Article Online
View Journal

Cite this: DOI: 10.1039/c3mb70435a

Prediction of outer membrane proteins by combining the position- and composition-based features of sequence profiles†

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

Locating the transmembrane regions of outer membrane proteins (OMPs) is highly important for deciphering their biological functions at both molecular and cellular levels. Here, we propose a novel method to predict the transmembrane regions of OMPs by employing the position- and composition-based features of sequence profiles. Furthermore, a simple probability-based prediction model, which is estimated by the secondary structures of structurally known OMPs, is also developed. Considering that these two methods are both effective and well complementary, we integrate them into a method called TransOMP, which is also capable of identifying OMPs. Furthermore, we develop an OMP identification measure I_CScore by considering transmembrane regions by TransOMP and secondary structural topology by SSEA-OMP. Our methods were benchmarked against state-of-the-art methods and assessed in the genome of *Escherichia coli*. Benchmark results confirmed that our methods were reliable and useful. Meanwhile, we constructed an OMP prediction web server, which can be used for OMP identification, transmembrane region location, and 3D model building.

Received 28th September 2013,
Accepted 26th February 2014

DOI: 10.1039/c3mb70435a

www.rsc.org/molecularbiosystems

1 Introduction

Integral membrane proteins embed in the cellular membranes of diverse organisms and perform a variety of biologically important functions.[1] Depending on the secondary structure in the transmembrane regions as well as physicochemical characteristics and localization, integral membrane proteins are grouped into two main categories, *i.e.* α-helical and β-barrel membrane proteins. The β-barrel membrane proteins are frequently found in the outer membrane of gram-negative bacteria, mitochondria and chloroplasts, therefore, they are also commonly known as outer membrane proteins (OMPs). Currently, except for an *Escherichia coli* OMP (PDB entry: 2J58) containing the α-helical transmembrane region,[2] the remaining OMPs are β-barrel membrane proteins. In fact, OMPs are of great interest in the biological community considering that they play a wide variety of biological roles in cells, including enzymes, transporters and membrane-embedded channels.

Given the difficulty in structural determination of OMPs through wet experiments, computational methods to identify putative OMPs in the sequenced genomes and to locate their transmembrane regions have become increasingly important in recent years. Currently, there exist two major tasks for the computational study of OMPs. One is to identify OMPs from genomes, and the other is to locate transmembrane regions of OMPs. By now, ~20 non-redundant OMP structures have been determined, which allow the development of special computational tools for OMPs. In reality, a few bioinformatics tools have been elegantly designed in the research community. Of them, several methods were developed by statistical analyses based on the amino acid composition.[3–7] C-terminal patterns, hydrophobicity and amphipathicity of β-strands were also used to predict OMPs.[8,9] The machine learning algorithms (*e.g.* Neural Network and Support Vector Machine) were also employed to construct OMP prediction methods.[10–13] Interestingly, hidden Markov models were commonly used by several groups. For example, Bagos and co-workers trained hidden Markov models using structurally known OMPs to locate transmembrane regions.[14,15] The PROFtmb method proposed by the Rost group uses a profile-based hidden Markov model.[16] HHomp identifies OMPs by HMM–HMM matching.[17] In our previous work, we proposed a method called SSEA-OMP to identify OMPs by using secondary structure element alignment.[18]

The past few decades have witnessed a series of studies[19–29] and a few bioinformatics algorithms[3,4,6–9,13–15,17,30–36] for OMP

[a] Institute of Applied Genomics, School of Biological Science and Engineering of Fuzhou University, Fuzhou 350002, China. E-mail: yanrenxiang@fzu.edu.cn; Fax: +86 0591 22866273; Tel: +86 0591 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/c3mb70435a

prediction, however, some critical questions remain to be addressed. For example, are there any new methods for OMP prediction? How reliable are the modelled structures for OMPs by classical fold recognition methods? What is the mechanism by which OMPs insert and fold in a membrane? In this article, we make an attempt to answer one of the questions by developing a novel OMP prediction method based on the position- and composition-based features of sequence profiles.

2 Materials and methods

2.1 Benchmark dataset

The benchmark datasets were constructed with the utilization of information in the PDB,[37] UniProtKB[38] and OPM[39] databases. Firstly, we extracted OMPs, which are structurally determined and contain transmembrane region annotation in UniProtKB. A set of protein entries was collected by scanning the PDB database with the keyword 'outer membrane protein'. The PDB entries were mapped to UniProtKB accession names and the corresponding annotations were extracted. Only the proteins containing the 'transmembrane β-strand' annotation by UniProtKB were retained. Meanwhile, the obtained proteins were split into chains. We manually checked each chain and removed non-typical β-barrel chains (*e.g.* 1EK9A, 1IMOA, 3LDTA and 1WP1A). The remaining proteins were further filtered by 30% identity. In this procedure, we got 14 non-redundant OMPs. We named these 14 proteins the OMP14 dataset. Secondly, we collected OMPs from the OPM database. We BLASTed all β-barrel transmembrane proteins of the OPM database against the sequences in the OMP14 dataset, and there exist 14 β-barrel transmembrane proteins of the OPM database that were not homologous with the OMP14 dataset. Therefore, we compiled these 14 OMPs into a new dataset called ADD14. It should be pointed out that the transmembrane regions of the ADD14 proteins were not annotated in the UniProtKB dataset. The transmembrane regions of these proteins were annotated based on the membrane boundaries obtained from the OPM database. Both OMP14 and ADD14 datasets are available in the ESI,† S1.

2.2 Profile composition-based features (PCF)

In this work, composition-based features of sequence profiles are calculated by the procedure consisting of three main steps: (i) sequence profile generation; (ii) fragment profile extraction, and (iii) *k*-spaced residue pair composition construction.

(i) Sequence profile generation. To obtain the sequence profiles, the query sequence is iteratively threaded through the NCBI NR database for 3 repeats with an *e*-value cutoff of 0.001 for collecting multiple sequence alignments (MSAs). The Henikoff weight scheme[40] is used to reduce the redundancy of MSAs in the position-specific frequency matrix (PSFM profile) and position-specific scoring matrix (PSSM profile) constructions.

(ii) Fragment profile extraction. To get the fragment profiles of the 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. Both PSFM and PSSM fragment profiles are excised for the encoding construction. Based on the OMP14 dataset, different window sizes in the range of 1 to 41 were tested. The optimal window size was determined by the Leave-One-Out (LOO) procedure. In each step of the LOO test, one protein is the test target and the remaining 13 proteins are used as the training set. We tested different window sizes in the 13 training proteins. The obtained optimized window size of the 13 training proteins was applied to the target protein. According to our preliminary computational experiments, the sizes of windows for PCF-based encodings were consistently set as 11 in this study.

(iii) *k*-Spaced residue pair composition construction. The encoding of PCF for the target residue is constructed by using the excised fragment profiles. Taking a fragment profile with N residues as an example, it is an $N \times 20$ matrix of twenty amino acid occurrence probabilities and it can be represented as $a[i,j]$, where i denotes the position of the target residue in the fragment and j stands for the occurrence probability of the jth amino acid. To calculate the amino acid pair of A_mA_n with k-spaced (*i.e.* pairs that are separated by k any other amino acids), we use a similar way as Chen *et al.*,[41] in which the encoding was employed for the classification of 5 integral membrane protein types. The equation used here is

$$A_mA_n = \frac{1}{N-k-1}\sum_{i=1}^{N-k-1}\min(a[i,m],\ a[i+k+1,n]) \qquad (1)$$

where N denotes the window size of the fragment profile and k represents that the k-spaced residue pairs are taken into account. There are 20 amino acids and 400 amino acid pairs in total. Therefore, the dimensional number of features for each specific k is 400. Amino acid pairs for $k = 0, 1, \ldots, \frac{N-1}{2}$ are jointly considered. It is very time-consuming and the performance cannot be improved if k is larger than 5. The maximum value of k is set to not larger than 5. We use two sets of profiles (*i.e.* PSFM and PSSM) to build the features and train the models. Considering that some elements of the PSSM profile are negatives, we employ two strategies for the transformations. The first strategy is very simple and we directly set the element value of the PSSM profile to 0 if it is negative; in the second strategy, all PSSM profile elements are scaled to the range of 0–1 by using the standard logistic function as

$$\frac{1}{1+e^{-x}} \qquad (2)$$

where x is the element value of the PSSM profile. In total, there are three encodings, *i.e.* PCF^F for the calculation of k-spaced residue pairs using the PSFM profile, PCF^L for the calculation of k-spaced residue pairs using the PSSM profile by setting negative values to zeros, and PCF^E for the calculation of k-spaced residue pairs using the PSSM profile by scaling the element values to the range of 0–1 using the standard logistic function. The differences between the encodings used here and those by Chen *et al.*[41] are that we use two sets of profiles (*i.e.* PSFM and PSSM) and two strategies for PSSM transformations whereas Chen *et al.* only used the PSSM profile by setting negative values

to zeros. In this work, we will show that the PSFM is more informative in the OMP transmembrane region prediction than PSSM and scaling the PSSM profile by using the standard logistic function (*i.e.* eqn (2)) also generates higher Mcc values than that by simply setting negative values to zeros.

2.3 Profile position-based features (PPF)

The procedure to generate PPF-based features is similar to that of PCF, including sequence profile generation, fragment profile extraction and encoding construction. For a sliding window containing the $(2n + 1)$ fragment profile with the target residue in the center, there are $(2n + 1) \times 20$ features. The optimal window sizes for PPF-based encodings were assigned as 35. The optimization process of the window sizes in PPF-based methods is the same as that in PCF-based methods. Again, the two transformations are also used to scale the elements of the PSSM profile. Similarly, there are also three types of resulting features, *i.e.* PPF^F for building features using the PSFM profile, PPF^L for building features using the PSSM profile by setting negative values to zeros, and PPF^E for building features using the PSSM profile by scaling the element values to the range of 0–1 using the standard logistic function. The difference between the encoding construction of PPF and that used in PCF is that the profile values are directly used as feature vectors in PPF whereas the amino acid pairs are counted in PCF.

2.4 Secondary structure-based prediction

The transmembrane regions of OMPs are strongly relevant to the secondary structure. Therefore, we also analyze and use the predicted secondary structure by PSIPRED, which is a Neural Network-based method.[42] The binary checkpoint profile generated by PSI-BLAST[43] is fed to PSIPRED[42] and the output possibilities for secondary structures are employed to predict the transmembrane regions of OMPs. The possibility for the target residue in the transmembrane region is defined as

$$P(M|SS) = P(E)P(M|E) + P(C)P(M|C) + P(H)P(M|H) \quad (3)$$

where P(M|SS) is the possibility of the target residue in the transmembrane region by using secondary structure information; P(E), P(C) and P(H) are the possibilities of the target residue that are predicted as strand, helix and coil by PSIPRED; P(M|E), P(M|C) and P(M|H) are the conditional possibilities of the target residue in the transmembrane region if the secondary structures are strand, coil and helix, respectively. P(M|E), P(M|C) and P(M|H) were estimated by the structurally known OMP14 dataset and the corresponding values are shown in Table 1. When the OMP14 dataset was tested, 13 training proteins were used to estimate the probabilities and 1 target protein was tested in each run of cross-validation. In Table 1, the probabilities in any row (*i.e.* any secondary structure type) sum up to 1. Similarly, the equation for the target residue in the non-transmembrane region is defined as

$$\begin{aligned} P(\sim M|SS) = {} & P(E)(1 - P(M|E)) + P(C)(1 - P(M|C)) \\ & + P(H)(1 - P(M|H)) \end{aligned} \quad (4)$$

Table 1 The observed probabilities of a residue to be located in transmembrane (non-transmembrane) in the three types of secondary structures

Type[a]	Transmembrane	Non-transmembrane
β-Strand (E)	0.700	0.300
Coil (C)	0.098	0.902
α-Helix (H)	0.000	1.000

[a] 'E' denotes a strand element, 'C' stands for a coil element, and 'H' represents a helix element.

where P(∼M|SS) is the possibility of the target residue in the non-transmembrane region by using secondary structure information; other terms are defined in eqn (3). Finally, the prediction score for the target residue located in the transmembrane region by using the secondary structure information is

$$\begin{aligned} SS\_pred = {} & P(M|SS) - P(\sim M|SS) = P(E)P(M|E) + P(C)P(M|C) \\ & + P(H)P(M|H) - P(E)(1 - P(M|E)) \\ & + P(C)(1 - P(M|C)) + P(H)(1 - P(M|H)) \\ = {} & P(E)(2P(M|E) - 1) + P(C)(2P(M|C) - 1) \\ & + P(H)(2P(M|H) - 1) \end{aligned} \quad (5)$$

where SS_pred is the prediction score and other terms are defined in eqn (3) and (4).

2.5 Support vector machine learning and the TransOMP method

The encodings can be transformed to a prediction score *via* support vector machine learning (SVM). The SVM package used in this work is svmlight (svmlight.joachims.org/). The overall performance of various encoding-based methods is assessed by the Leave-One-Out (LOO) procedure. In each step of the LOO test, 1 protein is selected as the test target and the remaining 13 proteins are used as the training set. This process is iteratively performed 14 times for all OMPs in the OMP14 benchmark. When we applied SVM learning, linear, polynomial, RBF, and sigmoid kernel functions of SVM were tested and we selected the linear kernel to optimize the prediction. The top and complementary encodings are selected to construct the TransOMP method, which is a weighted average model by combining PCF^F, PCF^E, PPF^L and SS_pred as

$$T\_CScore(i) = \frac{w_1 PCF^F + w_2 PCF^E + w_3 PPF^L + w_4 SS\_pred}{w_1 + w_2 + w_3 + w_4} \quad (6)$$

where T_CScore(i) is the predicted confident score of TransOMP for the residue in the position i of the sequence. The encodings of PCF^F, PCF^E, PPF^L and SS_pred are calculated for the target residue i. w_1, w_2, w_3 and w_4 are weights of PCF^F, PCF^E, PPF^L and SS_pred. Here, w_1, w_2, w_3 and w_4 are empirically set to 0.3, 0.25, 0.25 and 0.2 to balance the terms. The flowchart for constructing TransOMP is depicted in Fig. 1. Furthermore, we combine SSEA-OMP and TransOMP to identify OMPs. SSEA-OMP discriminates OMPs by using secondary structure alignment and we have shown that its performance is comparable to other existing methods in our previous work.[18] Briefly, SSEA-OMP discriminates

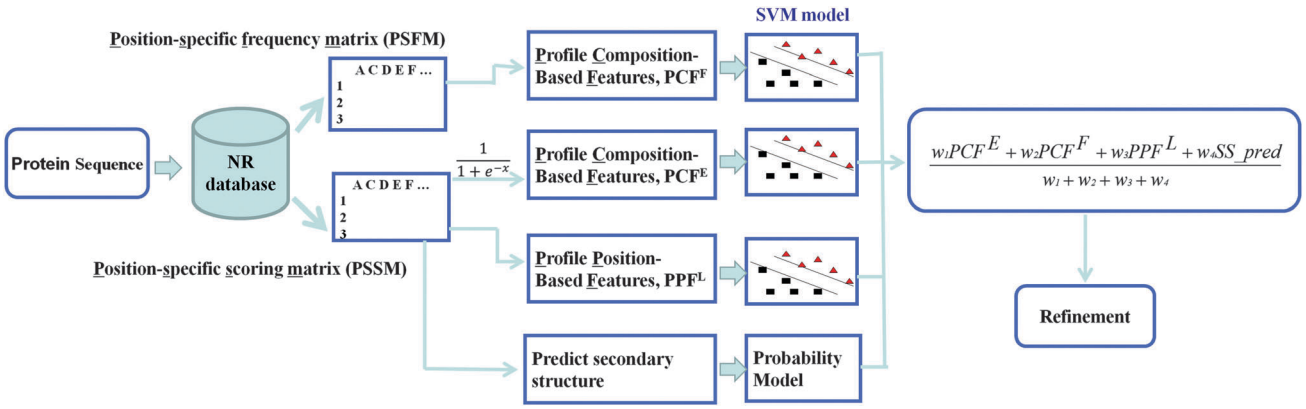

Fig. 1 Flowchart of the TransOMP method.

OMPs by searching the query sequence against OMP and non-OMP databases. Meanwhile, the maximal SSEA (secondary structure element alignment) similarity score between OMPs (non-OMPs) are recorded (*i.e.* $SSEA_{max\_omp}$ and $SSEA_{max\_non\_omp}$). 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. Thus, the predicted secondary structural string was shortened and the length of each element was retained for the scoring of SSEA. The alignment score of SSEA between two secondary structure elements with lengths L_i and L_j is defined as

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

where $\min(L_i, L_j)$ stands for the minimal length between L_i and L_j. The total alignment score is further divided by the average length of these two sequences to obtain a normalized similarity score. For details of the SSEA algorithm one can refer to our previous work.[18] It should be mentioned here that the SSEA-OMP method is for OMP identification. For a query sequence, the prediction score ΔSSEA is calculated as

$$\Delta\text{SSEA} = SSEA_{max\_omp} - SSEA_{max\_non\_omp} \quad (8)$$

Here, we use the same scheme as mentioned previously to construct the library of the SSEA-OMP method, *i.e.* 486 cluster consensus sequences, which were derived from 23 structurally solved OMPs collected by the Söding group, are used as the OMP database, and 941 non-OMPs collected by Gromiha *et al.*[31] are used as the non-OMP database. For a query sequence, TransOMP determines whether it is OMP using the following equation:

$$\text{TransOMP\_Score} = \sum_{i=1}^{N} \max(\text{T\_CScore}(i), 0) \quad (9)$$

where T_CScore(i) is defined in eqn (6) and we use max(T_CScore(i),0) to ensure that only transmembrane regions are calculated (*i.e.* positive values). N is the length of the target protein. Furthermore, we propose an OMP identification confident score (I_CScore) by combining TransOMP_Score and ΔSSEA as

$$\text{I\_CScore} = \frac{\alpha\text{TransOMP\_Score}}{N} + \Delta\text{SSEA} \quad (10)$$

ΔSSEA is calculated using eqn (8). To balance the two terms, TransOMP_Score is normalized by the length of the target protein (N). α is empirically set to 0.3 to optimize prediction.

2.6 Refinement of transmembrane region prediction

The refinement was developed by the observation that transmembrane regions are segments, and the length of each segment is more than 2 residues. We designed the following steps to refine the prediction. Firstly, for one or two residues predicted to be located in the transmembrane region, we transform their states to non-transmembrane if the six residues around the ±3 positions (*i.e.* neighbouring 6 residues) locate in the non-transmembrane region (Fig. 2A and B). Meanwhile, the prediction scores for the (two) residues are reset to the average of the neighbouring 6 residues. Secondly, for one (two) residue(s) predicted to be located in the non-transmembrane region, if its 5 neighbouring residues locate in the transmembrane regions, the states of the (two) residues will be transformed into the transmembrane region (Fig. 2C and D). Similarly, the prediction scores of the (two) residues are reset to the average of the neighbouring 5 transmembrane residues. This refinement process can correct a few obvious errors by prediction. For example, when we applied the refinement process to the TransOMP method in the OMP14 dataset, 2 false positives and 4 false negatives were corrected. Note that the refinement is quite simple and further optimization may result in an improved performance.

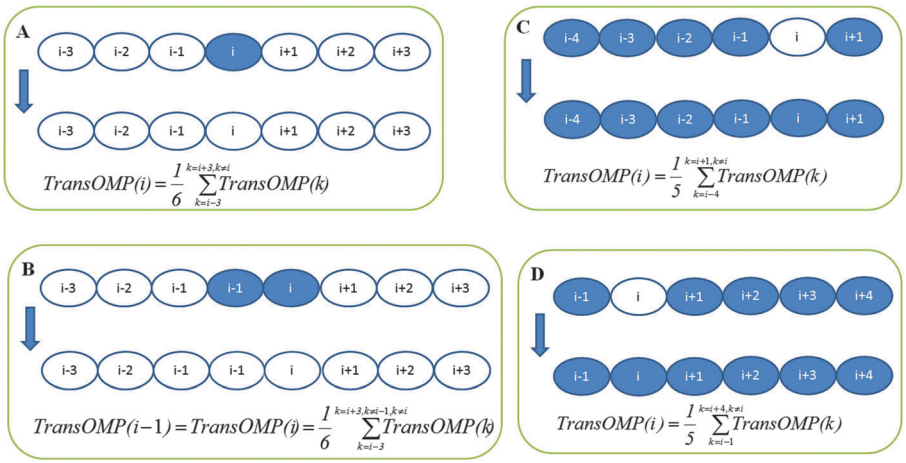

Fig. 2 Refinement of TransOMP prediction. Transmembrane and non-transmembrane regions are coloured in blue and white, respectively.

2.7 Performance assessment

When the test is performed over all proteins in the 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 transmembrane (non-transmembrane) residues are considered positives (negatives). These parameters are defined as

$$\mathrm{Ac} = \frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{fn}+\mathrm{tn}+\mathrm{fp}} \quad (11)$$

$$\mathrm{Sn} = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}} \quad (12)$$

$$\mathrm{Sp} = \frac{\mathrm{tn}}{\mathrm{tn}+\mathrm{fp}} \quad (13)$$

$$\mathrm{Mcc} = \frac{\mathrm{tp}\times\mathrm{tn}-\mathrm{fp}\times\mathrm{fn}}{\sqrt{(\mathrm{tp}+\mathrm{fp})(\mathrm{tp}+\mathrm{fn})(\mathrm{tn}+\mathrm{fn})(\mathrm{tn}+\mathrm{fp})}} \quad (14)$$

where tp, fp, fn and tn denote the number of true positives, false positives, false negatives and true negatives, respectively. The performance of OMP identification can be measured using a receiver operating characteristic (ROC) curve. The ROC curve plots the true-positive rate (instances) as a function of false-positive rate (instances) for all possible thresholds of prediction scores by various methods.

3 Results and discussion

3.1 The performance of transmembrane region prediction

Based on the OMP14 dataset, the overall performance of different OMP transmembrane region prediction methods is assessed using the LOO procedure as described in the Materials and methods section. The prediction results of the transmembrane regions for various methods are summarized in Table 2. Mcc is the most comprehensive parameter to measure the prediction performance. Of the 6 encoding-based methods, PCF^F obtains the highest Mcc value (Mcc = 0.641). PCF^E and PPF^L are ranked 2nd and 3rd. The top methods, PCF^F, PCF^E, PPF^L, and SS_pred were selected to construct the TransOMP method according to the fact that these methods can generate Mcc values higher than 0.600 although the benchmark set has already been filtered by 30% sequence identity. We also tried to include other encodings (*e.g.* PPF^E), but the performance was not improved. Although both PCF- and PPF-based methods were inferred from sequence profiles, PCF-based methods should not be considered redundant to PPF-based methods, since PPF encoding counts the 20 amino acid occurrence probabilities whereas PCF counts the residue pair composition. In order to examine the complementary encoding among the four encodings, we generated a Venn diagram using R package[44] based on their prediction results (Fig. 3). As shown in Fig. 3, the results predicted by the three encoding-based methods are well complementary. For example, PCF^F, PCF^E, PPF^L and SS_pred methods correctly distinguish 17, 25, 95 and 186 residues that cannot be identified by other three methods, respectively. Although the SS_pred predictor generated the lowest Mcc value among the four encoding-based methods, it was most complementary to other methods according to Fig. 3.

Table 2 Performance of transmembrane region prediction of various methods on the OMP14 dataset

Method	TP[a]	TN	FP	FN	Ac	Sn	Sp	Mcc
Encoding-based methods								
PCF^F	1607	2977	349	590	0.829	0.731	0.895	0.641
PCF^E	1585	2991	335	612	0.829	0.721	0.899	0.638
PPF^L	1595	2982	344	602	0.828	0.725	0.896	0.638
PPF^E	1591	2918	408	606	0.816	0.724	0.877	0.613
SS_pred	1603	2887	439	594	0.812	0.729	0.868	0.605
PCF^L	1564	2919	407	633	0.812	0.712	0.877	0.602
PPF^F	1530	2874	452	667	0.797	0.696	0.864	0.572
Comparison with well-established methods								
TransOMP	1691	2982	344	506	0.846	0.769	0.896	0.676
PROFTMB	1921	2736	590	276	0.843	0.874	0.822	0.685
PRED-TMBB	1483	3078	248	714	0.825	0.675	0.925	0.633
TMBpro	1689	2595	731	508	0.775	0.768	0.780	0.541
TMBETAPRED-RBF	1379	2995	331	818	0.792	0.628	0.901	0.559

[a] All residues of the OMP14 dataset were used to count true positive (TP), true negative (TN), false positive (FP) and false negative (FN) measures.

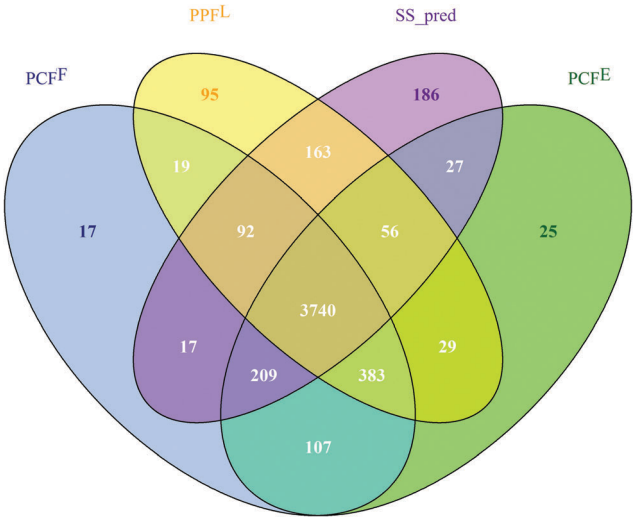

Fig. 3 Analysis of the complementarity of encoding-based predictors.

Table 3 Performance of transmembrane region prediction of various methods on the ADD14 dataset

Method[a]	TP	TN	FP	FN	Ac	Sn	Sp	Mcc
TransOMP	1512	2306	334	335	0.850	0.818	0.873	0.692
PROFTMB	1746	1916	724	101	0.816	0.945	0.725	0.663
PRED-TMBB	1321	2325	315	526	0.812	0.715	0.880	0.609
TMBpro	1627	2133	507	220	0.837	0.880	0.807	0.678
TMBETAPRED-RBF	1372	2286	354	475	0.815	0.742	0.865	0.615

[a] All residues of the ADD14 dataset were used to count true positive (TP), true negative (TN), false positive (FP) and false negative (FN) measures.

The TransOMP method is a weighted average model of the four encodings using eqn (6). As shown in Table 2, Ac, Sn and Sp scores of TransOMP predictions are 0.846, 0.769 and 0.896, respectively, which result in an overall Mcc value of 0.676. This Mcc value is approximately 5% higher than the best individual encoding-based prediction. Additionally, we also tested the TransOMP method using the ADD14 dataset (Table 3). The TransOMP method was trained using the OMP14 dataset and all proteins in the ADD14 dataset are not homologous with any protein in the OMP14 dataset (BLAST *e*-value > 0.01). TransOMP got the similar performance in both ADD14 and OMP14 datasets (Mcc = 0.692 *versus* 0.676).

3.2 Comparison with well-established methods

We relied on the OMP14 and ADD14 datasets to benchmark TransOMP against state-of-the-art transmembrane region prediction methods. The proteins of OMP14 and ADD14 datasets were directly submitted to PRED-TMBB, TMBpro and TMBETAPRED-RBF web servers. PROFTMB program was installed in our local computer. These programs represent typical and publicly available OMP transmembrane region prediction methods. As shown in Table 2, the TransOMP method can generate a higher Mcc value than the other methods. But this does not mean that TransOMP will replace others. In fact, the methods tested here are very diverse. For example, PRED-TMBB was trained based on HMM at the sequence level and its prediction results were significantly different from those of our method (*t*-test *p*-value $< 2.2 \times 10^{-16}$). At least, all methods benchmarked here can generate reasonable prediction performance. Therefore, they should be useful in the real application. In fact, the methods developed by some groups are very different. Regarding the HMM-based method PRED-TMBB, which was developed by Bagos *et al.*, for instance, its design of topology of the HMM, number of states and their connection need *a priori* fixed by taking insightful knowledge of known OMPs. Moreover, the development of consensus methods based on these methods can improve the prediction. There are slight differences in the predictions in the datasets of OMP14 and ADD14. The quality of these predictions should be further assessed through computing their confidence intervals. Confidence intervals are computed using the common assumption of a normal distribution by the following equation:

$$\left(\mu - Z\frac{\mathrm{SD}}{\sqrt{n}},\ \mu + Z\frac{\mathrm{SD}}{\sqrt{n}}\right) \quad (15)$$

where μ and SD are mean and standard deviation of the samples, n is the sample size, and Z is the critical value and Z equals to 1.96 at a 95% confidence level. The mean (μ) and standard deviation (SD) of sensitivity and specificity scores were calculated by bootstrap resampling for 1000 repeats. The confidence interval values estimated in two datasets are listed in Tables 4 and 5. The sensitivity and specificity confidence intervals of the TransOMP method at the 95% level are [0.806, 0.829] and [0.863, 0.881] in the OMP14 dataset. Similarly, the sensitivity and specificity confidence intervals of the TransOMP method at the 95% level are [0.809, 0.828] and [0.866, 0.886] in the ADD14 dataset. Generally speaking, the confidence interval is judged to be better than another if it leads to intervals whose lengths are typically shorter. The shortest lengths of sensitivity confidence intervals in the two datasets are PROFTMB.

Table 4 Confidence intervals at 95% level for sensitivity and specificity estimated in the OMP14 dataset

Method	Sensitivity	Specificity
TransOMP	[0.806, 0.829]	[0.863, 0.881]
PROFTMB	[0.867, 0.881]	[0.815, 0.828]
PRED-TMBB	[0.838, 0.908]	[0.964, 0.987]
TMBpro	[0.789, 0.996]	[0.650, 0.844]
TMBETAPRED-RBF	[0.731, 0.763]	[0.841, 0.885]

Table 5 Confidence intervals at 95% level for sensitivity and specificity estimated in the ADD14 dataset

Method	Sensitivity	Specificity
TransOMP	[0.809, 0.828]	[0.866, 0.880]
PROFTMB	[0.939, 0.950]	[0.716, 0.734]
PRED-TMBB	[0.704, 0.725]	[0.874, 0.887]
TMBpro	[0.873, 0.888]	[0.800, 0.816]
TMBETAPRED-RBF	[0.732, 0.753]	[0.858, 0.872]

3.3 Large-scale benchmark of OMP identification

It is also valuable to investigate the performance of TransOMP in OMPs identification. In this work, we use the R-dataset, which is compiled by the Söding group, and the dataset consists of 2164 OMPs and 5000 non-OMPs. To remove homologous sequences, we searched the 2164 OMPs against the 14 structurally known proteins in the OMP14 dataset. We removed 128 sequences that have better significance than the BLAST *e*-value of 0.01 with the sequences in the OMP14 dataset. Finally, 2036 OMPs and 5000 non-OMPs were retained. The sequences in the R-dataset are not homologous with the 23 structurally known OMPs, which were used to derive 486 cluster consensus sequences,[17] at the sequence level. When tested on the R-dataset, the PSI-BLAST *e*-value of 0.01 was used as the criterion to remove homologous sequences with the 941 non-OMPs. The performance of SSEA-OMP and TransOMP in OMP identification was compared *via* ROC analysis. Because the performance at low false positive rates is more important in real-world application, we paid more attention to compare the performance of different methods at $<1\%$ false positive rates (*i.e.* 50 false positive instances). SSEA-OMP correctly recognized 1215 OMPs before including 5 false positives, whereas TransOMP can detect 1099 OMPs (Table 6 and Fig. 4). Although the performance of TransOMP is not as good as that of SSEA-OMP, the two methods are well complementary and combining them can result in a higher accuracy prediction. As shown in Table 6 and Fig. 4, combining two methods (*i.e.* I_CScore) can identify more OMPs at the same false positive instances. For example, the I_CScore method can detect 1343 and 1450 OMPs before including 5 and 50 false positive instances, which are higher than the numbers identified by SSEA-OMP and TransOMP. Here, we do not list other OMP identification methods in Table 6, because we have proved that SSEA-OMP is comparable to existing methods in our previous work.[18]

Table 6 Comparison of receiver operator characteristics table ($\leq$50 false positives) for different methods

	Receiver operator characteristics ($\leq$50 false positives[a])									
	5	10	15	20	25	30	35	40	45	50
I_CScore	1343	1385	1401	1405	1413	1418	1421	1424	1437	1450
SSEA-OMP	1215	1260	1291	1314	1316	1329	1350	1368	1385	1405
TransOMP	1099	1203	1213	1222	1225	1227	1231	1231	1233	1234

[a] False positives correspond to those non-OMPs predicted as OMPs.

3.4 Proteome-wide OMP identification in *Escherichia coli*

To confirm the performance of the I_CScore measure in genome-scale application, it was used to identify the 'putative' and 'probable' OMPs in the *E. coli* proteome. The whole proteome of *E. coli* K-12,[45] which contains 4126 protein sequences, was downloaded from the NCBI database. In-depth manually annotating the OMPs in the *E. coli* proteome is critically essential for analyses. In our previous work,[18] we collected a known OMP dataset consisting of 120 proteins from the *E. coli* proteome by retrieving the annotations from NCBI, PSORTdb[46] and OMPdb[47] databases. In this work, we further scanned the remaining sequences against the UniProtKB database and found 6 multi-location proteins (gi numbers are 170080821, 170080837, 170082168, 170082515, 170081952 and 170082684) whose subcellular locations are annotated as 'cell outer membrane' plus 'other places' (*e.g.* lipid-anchor or peripheral membrane). To validate whether they are OMPs, we submitted the 6 sequences to SPARKS-X,[48] which represents a typical approach of the fold recognition method for protein structure prediction. The models of the two proteins, 170081952, known as 'outer membrane-bounded periplasmic space', and 170080837, known as 'outer-membrane lipoprotein LolB', have the characteristic shape of antiparallel β-strand barrels. These two proteins are very likely to be OMPs and we added them to the dataset. Finally, a database of known OMPs, consisting of 122 proteins, was constructed. The 4126 sequences were directly fed into SSEA-OMP and TransOMP algorithms. The final prediction result was determined by the I_CScore measure. There are 118 proteins predicted to be potential OMPs with a false positive rate control of 1% (ESI,† S2).

Among the 118 detected OMPs, 77 proteins have been included in the known *E. coli* OMP dataset. Therefore, these

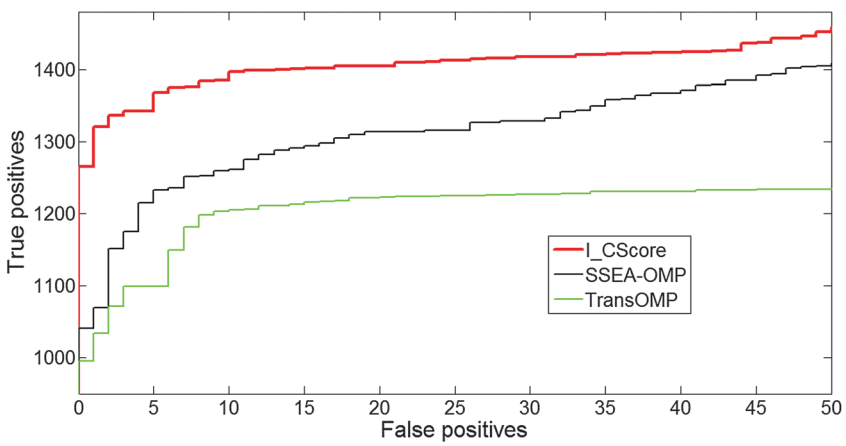

Fig. 4 Comparison of receiver operator characteristic curves ($\leq$50 false positives) for different methods.

77 predicted OMPs should be regarded as true positives with high confidence. In fact, 74 out of the 77 proteins are ranked among top 100 by the I_CScore. In the remaining 41 proteins, there exist 19 proteins whose subcellular localizations are annotated as 'unknown' or 'this protein may have multiple localization sites' in the PSORTdb database. These may be potential OMPs that have not been previously discovered. For example, protein 170080707, annotated as 'unknown' by the PSORTdb database and 'the displayed sequence is further processed into a mature form' in the UniProtKB database, was predicted to be an OMP like β-barrel structure by SPARK-X. Thus, protein 170080707 is very likely to be an OMP. When the PSORTdb database was further searched, the remaining 22 hits are clearly annotated as non-OMPs in terms of subcellular localization information, suggesting that they are very likely to be false positives. Among the 22 proteins, 7 are annotated as 'Cytoplasmic Membrane' and this may show that these proteins have very similar characteristics to OMPs. In fact, it is estimated that 97–98% protein sequences in the *E. coli* genome are non-OMPs, therefore, at 1% false positive rate, it is reasonable to have 30–40 false positives.

In summary, there are 77 true positives with high confidence, 19 possible true positives and 22 false positives identified by TransOMP at 1% false positive rate. In order to reduce the false positives, we may resort to other bioinformatics tools. For example, false positives could be further reduced by employing a signal peptide predictor according to the fact that most OMPs have a signal peptide.[46] Alternatively, we may choose the threshold value at a higher confidence level, but the identified true positives will be reduced accordingly. Additionally, to maximize the performance of SSEA-OMP and TransOMP, a regularly updated library which covers all sequence/structure spaces of known OMPs is highly desired.

3.5 Top ranked amino acid pairs in the evolutionary profile

It is useful to know the 'important' amino acid pairs of the composition-based encodings, in which some conserved and functional motifs may be found in OMP sequence families. The information gain algorithm that was implemented in the Weka package[49] was employed to quantify the relative importance of features in the most effective encoding (*i.e.* PCF^F). Based on the OMP14 dataset, the top 20 *k*-spaced amino acid pairs are listed in Table 7. These residue pairs are the most informative in the prediction that was identified by the Weka program. We also list the occurrence probability of each amino acid in the top 100 ranked *k*-spaced pairs (Table 8). As could be seen from the data, the top 5 amino acids are P, S, E, Q and D. Interestingly, Gromiha *et al.*[31] found that S and E distribute most significantly differently between OMPs and non-OMPs. Here, S and E ranked 2nd and 3rd. This may suggest that the top residues, which are listed in Table 4, distribute significantly differently in the transmembrane and non-transmembrane regions. For example, proline (P) is the most informative feature according to Table 7. The compositions of proline are 1.5% and 4.5% in transmembrane and non-transmembrane regions of the OMP14 dataset with *t*-test *p*-value < 0.01. This suggests significantly higher frequency in non-transmembrane regions than transmembrane regions. To probe the reasons for their distribution may need literature investigation.

Table 7 The top features in PCF^F encoding

#	Top residue pairs
1	PxS[a]
2	PA
3	SxP
4	PS
5	VP
6	SxxP
7	PxT
8	TxP
9	PP
10	PT
11	NxxP
12	PxxP
13	TxxP
14	AxP
15	PxxxQ
16	PxxxS
17	PxxxP
18	PxE
19	PxQ
20	AP

[a] The feature 'PxS' represents a 1-spaced residue pair of 'PS', where x stands for any amino acid. The same representation is applied to other *k*-spaced residue pairs.

Table 8 The amino acid composition in the top features of PCF^F encoding e

#	Occurrence probability in the top 100 features
1	P(0.47)[a]
2	S(0.085)
3	E(0.055)
4	Q(0.05)
5	D(0.05)
6	T(0.045)
7	N(0.045)
8	L(0.04)
9	A(0.04)
10	V(0.025)
11	R(0.025)
12	K(0.02)
13	G(0.02)
14	Y(0.015)
15	I(0.015)
16	F(0)
17	M(0)
18	C(0)
19	W(0)
20	H(0)

[a] The value inside the parentheses denotes the occurrence probability of the corresponding residue in the top 100 residue pairs.

3.6 The web server for OMP prediction

To aid the research community, a web server implementing the TransOMP and SSEA-OMP methods was constructed. The server was designed using Java, Perl and HTML. The web server is freely accessible at http://genomics.fzu.edu.cn/OMP/index.html. For each running, the web server calculates the secondary structure alignment score by SSEA-OMP and transmembrane

region score by TransOMP. Depending on the I_CScore measure, a query sequence determines whether it is an OMP. The 3D structures will also be provided by an in-house Bayes probability-based profile–profile alignment (paper in preparation). The multi-thread technique was employed and the computational time for processing a query sequence depends on the length of the query sequence. It is estimated that the job will be finished in 10 minutes if the query sequence is less than 500 amino acids. The prediction results including OMP identification, transmembrane region prediction, and 3D structural models will be mailed to users when the jobs are finished.

4 Conclusions

Taken together, we have clearly shown that the position- and composition-based encodings are effective features to predict the transmembrane regions of OMPs as well as to identify OMPs from genome-scale sequences. The success of our method should be ascribed to the encodings we used which can effectively represent the characteristics of the sequence environment surrounding the target residues in OMPs. Furthermore, a simple secondary structure probability-based prediction model was developed. The TransOMP method was constructed by combining these effective and complementary encodings.

Although Chen *et al.*[41] used *k*-spaced amino acid pairs of the PSSM profile for the classification of integral membrane proteins, it should be emphasized that our methods have made significant improvement with respect to PSFM and the standard logistic function introduced, which are more biologically meaningful for amino acid composition calculation. More importantly, the application aspects of Chen *et al.* and ours are different. Interestingly, both the PSFM profile and the standard logistic function-based transformation can result in higher Mcc values in the transmembrane region prediction. To the best of our knowledge, this is the first application of composition-based encoding of a sequence profile to the transmembrane region prediction of OMPs.

The TransOMP method has been benchmarked against state-of-the-art methods and results showed that the TransOMP method can generate higher Mcc values than other methods tested in this work. But this does not mean that TransOMP will replace other methods. The purpose of developing TransOMP is to provide the community with a practical tool. Moreover, we hope that the development of such novel methods will be helpful to accelerate the exploration of the sequence–structure protein landscape in OMPs. Last but not the least, as TransOMP relies on the evolutionary information of sequence profiles, it would raise the issue that the method may hamper its performance when the sequence profiles contain some false homologous sequences.

Acknowledgements

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 M. Punta, L. R. Forrest, H. Bigelow, A. Kernytsky, J. Liu and B. Rost, *Methods*, 2007, **41**, 460–474.
2 C. Dong, K. Beis, J. Nesper, A. L. Brunkan-Lamontagne, B. R. Clarke, C. Whitfield and J. H. Naismith, *Nature*, 2006, **444**, 226–229.
3 W. C. Wimley, *Protein Sci.*, 2002, **11**, 301–312.
4 M. M. Gromiha and M. Suwa, *Biochim. Biophys. Acta*, 2006, **1764**, 1493–1497.
5 M. M. Gromiha, S. Ahmad and M. Suwa, *Comput. Biol. Chem.*, 2005, **29**, 135–142.
6 M. M. Gromiha, S. Ahmad and M. Suwa, *Nucleic Acids Res.*, 2005, **33**, W164–W167.
7 A. G. Garrow, A. Agnew and D. R. Westhead, *BMC Bioinf.*, 2005, **6**, 56.
8 F. S. Berven, K. Flikka, H. B. Jensen and I. Eidhammer, *Nucleic Acids Res.*, 2004, **32**, W394–W399.
9 Y. Zhai and M. H. Saier, Jr., *Protein Sci.*, 2002, **11**, 2196–2207.
10 M. M. Gromiha and M. Suwa, *Proteins*, 2006, **63**, 1031–1037.
11 K. J. Park, M. M. Gromiha, P. Horton and M. Suwa, *Bioinformatics*, 2005, **21**, 4223–4229.
12 I. Jacoboni, P. L. Martelli, P. Fariselli, V. De Pinto and R. Casadio, *Protein Sci.*, 2001, **10**, 779–787.
13 M. M. Gromiha, S. Ahmad and M. Suwa, *J. Comput. Chem.*, 2004, **25**, 762–767.
14 P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos and S. J. Hamodrakas, *Nucleic Acids Res.*, 2004, **32**, W400–W404.
15 P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos and S. J. Hamodrakas, *BMC Bioinf.*, 2004, **5**, 29.
16 H. Bigelow and B. Rost, *Nucleic Acids Res.*, 2006, **34**, W186–W188.
17 M. Remmert, D. Linke, A. N. Lupas and J. Soding, *Nucleic Acids Res.*, 2009, **37**, W446–W451.
18 R. X. Yan, Z. Chen and Z. Zhang, *BMC Bioinf.*, 2011, **12**, 76.
19 S. D. Emr, J. Hedgpeth, J. M. Clement, T. J. Silhavy and M. Hofnung, *Nature*, 1980, **285**, 82–85.
20 N. Wiedemann, V. Kozjak, A. Chacinska, B. Schonfisch, S. Rospert, M. T. Ryan, N. Pfanner and C. Meisinger, *Nature*, 2003, **424**, 565–571.
21 S. Kim, J. C. Malinverni, P. Sliz, T. J. Silhavy, S. C. Harrison and D. Kahne, *Science*, 2007, **317**, 961–964.
22 R. Voulhoux, M. P. Bos, J. Geurtsen, M. Mols and J. Tommassen, *Science*, 2003, **299**, 262–265.
23 A. Belaaouaj, K. S. Kim and S. D. Shapiro, *Science*, 2000, **289**, 1185–1188.
24 K. R. Wong and J. T. Buckley, *Science*, 1989, **246**, 654–656.
25 R. James, *J. Bacteriol.*, 1975, **124**, 918–929.
26 A. V. Rodionov, *Bioorg. Khim.*, 1990, **16**, 1687–1689.
27 W. C. Wimley, *Curr. Opin. Struct. Biol.*, 2003, **13**, 404–411.
28 J. W. Fairman, N. Noinaj and S. K. Buchanan, *Curr. Opin. Struct. Biol.*, 2011, **21**, 523–531.

29 E. k. Teerasak, R. Burchmore, P. Herzyk and R. Davies, *BMC Bioinf.*, 2012, **13**, 63.
30 M. M. Gromiha, *Biophys. Chem.*, 2005, **117**, 65–71.
31 M. M. Gromiha and M. Suwa, *Bioinformatics*, 2005, **21**, 961–968.
32 P. L. Martelli, P. Fariselli, A. Krogh and R. Casadio, *Bioinformatics*, 2002, **18**(suppl 1), S46–S53.
33 P. G. Bagos, T. D. Liakopoulos and S. J. Hamodrakas, *BMC Bioinf.*, 2005, **6**, 7.
34 C. F. Reboul, K. Mahmood, J. C. Whisstock and M. A. Dunstone, *Bioinformatics*, 2012, **28**, 1299–1302.
35 S. Hayat and A. Elofsson, *Bioinformatics*, 2012, **28**, 516–522.
36 M. M. Gromiha and Y. Y. Ou, *Briefings Bioinf.*, 2013, DOI: 10.1093/bib/bbt015.
37 J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter and E. E. Abola, *Acta Crystallogr.*, 1998, **54**, 1078–1084.
38 E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider and A. Bairoch, *Methods Mol. Biol.*, 2007, **406**, 89–112.
39 M. A. Lomize, A. L. Lomize, I. D. Pogozheva and H. I. Mosberg, *Bioinformatics*, 2006, **22**, 623–625.
40 S. Henikoff and J. G. Henikoff, *J. Mol. Biol.*, 1994, **243**, 574–578.
41 K. Chen, Y. Jiang, L. Du and L. Kurgan, *J. Comput. Chem.*, 2009, **30**, 163–172.
42 D. T. Jones, *J. Mol. Biol.*, 1999, **292**, 195–202.
43 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.
44 H. Chen and P. C. Boutros, *BMC Bioinf.*, 2011, **12**, 35.
45 F. R. Blattner, G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao, *Science*, 1997, **277**, 1453–1462.
46 S. Rey, M. Acab, J. L. Gardy, M. R. Laird, K. deFays, C. Lambert and F. S. Brinkman, *Nucleic Acids Res.*, 2005, **33**, D164–D168.
47 K. D. Tsirigos, P. G. Bagos and S. J. Hamodrakas, *Nucleic Acids Res.*, 2011, D324–D331.
48 Y. Yang, E. Faraggi, H. Zhao and Y. Zhou, *Bioinformatics*, 2011, **27**, 2076–2082.
49 E. Frank, M. Hall, L. Trigg, G. Holmes and I. H. Witten, *Bioinformatics*, 2004, **20**, 2479–2481.

