Classifying Substrate Specificities of Membrane Transporters from *Arabidopsis thaliana*

Nadine S. Schaadt, Jan Christoph, and Volkhard Helms*

Center for Bioinformatics, Saarland University, D-66123 Saarbrücken, Germany

Received June 25, 2010

Membrane transporters catalyze the active transport of molecules across biological barriers such as lipid bilayer membranes. Currently, the experimental annotation of which proteins transport which substrates is far from complete and will likely remain so for much longer. Therefore, it is highly desirable to develop computational methods that may aid in the substrate annotation of putative membrane transport proteins. Here, we measured the similarity of membrane transporters from *Arabidopsis thaliana* by their amino acid composition, higher sequence order information, amino acid characteristics, or sequence conservation. We considered the substrate classes amino acids, oligopeptides, phosphates, and hexoses. Substrate classification based on the amino acid frequency yielded an accuracy of 75% or higher. Integrating additional information improved the prediction performance to 90% and higher.

INTRODUCTION

Most water-soluble organic and inorganic molecules require the activity of membrane transport proteins to cross phospholipid bilayer membranes. These integral transmembrane proteins act either as passive channels or as active carriers, being selective for a particular substrate or substrate group.[1−3] Although the amino acid sequences of many putative membrane transport proteins are known, their specific functions remain often unknown due to the large experimental effort involved in determining substrate specificities. Therefore, it is very desirable to complement these experiments by computational methods to predict putative transported substrates. Moreover, transporters are often able to carry out the transport of various substrates,[4] and often multiple transporters exist for one particular substrate. It is therefore quite likely that only computational methods may be able to provide an integrated picture of the 'transportome'. Previously, membrane transporters have been assigned to specialized transporter families based on multiple sequence alignments[5−7] or phylogeny.[8] However, it is often not possible to detect the transported substrate based on this classification. Furthermore, the common approaches may fail because of low levels of sequence similarity between transporters exhibiting the same function. Hence, we present in this study an alternative way of substrate prediction for membrane transporters using different characteristic features of their amino acid composition.

The amino acid composition (AAC) is a feature that has been frequently used to characterize proteins. In the context of protein classification, it was introduced in 1983 by Nishikawa et al. to distinguish between the inside and outside of biological cells, as well as between enzyme and nonenzyme.[9] Since then the amino acid composition has been applied to a large variety of properties such as secondary or tertiary structure prediction.[10−13] Subsequently, Chou et al. introduced the so-called pseudo amino acid composition (PseAAC)[14] to enhance the prediction quality of protein classification. The PseAAC includes physicochemical properties and correlations of residue pairs and thus partially accounts for long-range sequence-order information. Later, Park et al. utilized the pair amino acid composition (PAAC) to predict subcellular localizations.[15] Although subcellular localizations are closely related to the function of the corresponding protein, they only indicate whether a protein is a transporter or not and do not provide information about the corresponding substrate. Gromiha and Yakubi[16] used the AAC for functional classification of helical transmembrane proteins and grouped them into three categories of channels/pores, electrochemical potential-driven transporters, and primary active transporters. Furthermore, Ou, Chen, and Gromiha showed that it is possible to discriminate between transporters of six different families using AAC.[17] Davies et al.[18] predicted the function of G-protein-coupled receptors (GPCRs) based on the physicochemical properties of amino acids. Furthermore, a ranking algorithm to sort members of an unknown sample set according to their relationship to a true sample was used to predict blood-secretory proteins.[19]

Our aim here is to classify the sequences of membrane transporters according to the transported substrates. We did not only consider transporters belonging to the same family, but we also combined proteins of different families that transport the same substrate. The model organism *Arabidopsis thaliana* was selected for this exploratory investigation because the substrate annotations for this plant known so far from experiment are available through the manually curated database Aramemnon.[20] We show that the amino acid composition, together with some variants, can be used to very accurately predict the substrate specificities when given the protein sequence and a test set of proteins with known substrate annotations. The predictions are based on a large number of different ranking lists that are merged to a final ranking by a cross entropy Monte Carlo method.[21]

* Corresponding author e-mail: volkhard.helms@bioinformatik.uni-saarland.de.

10.1021/ci100243m
Published on Web 10/06/2010

Table 1. Data Subsets

substrate class	set size	protein families	sequence length	
			mean	standard deviation
amino acid	16	3	478.3	2.8
oligopeptide	17	2	662.3	23.1
phosphate	15	5	507.7	34.3
hexose	13	2	547.6	9.3

METHODS

Data Sets. The full data set used in this study contains 793 *Arabidopsis thaliana* membrane transport proteins annotated in the Aramemnon database[20] irrespective of their substrate class. For this, we retrieved all those protein sequences from the Aramemnon database for which the entries included the keywords *transport* or *carrier*. We removed redundant sequences from our data sets when their identity exceeded 90% using Blast.[22] We use this relatively high level of sequence identity in order not to remove too many transporters. Below, we will show that sequence identity turned out not to be useful for substrate prediction anyhow. We checked all members of this set using the secondary structure prediction tools Split 4.0[23] or MEMSAT[24] whether they contained at least one transmembrane helix. Otherwise, the corresponding transporter was discarded. This gave rise to 793 membrane transporters as mentioned above.

We then constructed subsets of the full data set whereby each subset contains those transporter sequences that are annotated in Aramemnon to transport a particular substrate. Proteins labeled as *putative* were not taken into account. We considered the following substrate classes: amino acid, oligopeptide, phosphate, and hexose, see Table 1, which all contain more than 10 members. These transporter groups are called positive sets hereafter. All positive sets contain members of multiple protein families. For example, the amino acid set consists of 14 TC 2.A.18 Amino Acid/Auxin Permease transporters and the two exceptions At2g02040 and At4g21120. The oligopeptide transporter set contains 17 transporters that belong to two different protein families, namely TC 2.A.67 Oligopeptide Transporter and TC 2.A.17 Proton-dependent Oligopeptide Transporter. The phosphate transporter positive set includes members from several different protein families (TC 2.A.1.9, TC 2.A.1.14, TC 2.A.20, TC 2.A.29). Further, the phosphate transporters show the largest variability with respect to the sequence length.

To evaluate the accuracy of the prediction results, we constructed additional negative sets that should be complementary to the positive sets. Because it is often unknown whether transporters exist that definitely do not transport a certain substrate, the negative sets were constructed by randomly selecting from the other nonrelated positive sets the same number of transporters as in the respective positive set and assuming that most of these actually do not transport the corresponding substrate. The transporter groups from which these negative sets were built were taken from Table 1. For example, we assumed that amino acid or oligopeptide transporters do not transport sugar molecules, such as hexoses.

Analysis of the Data Sets. We first estimated the similarities within each positive set and the pairwise similarities and dissimilarities between all positive sets. For the similarity in a set with q members, we summed the squared Euclidean distances between every pair of transporters i and j in this positive set using their individual amino acid compositions v_i and v_j:

$$d = \frac{2}{q(q-1)} \sum_{i,j} \sum_{k=1}^{20} (v_{i_k} - v_{j_k})^2 \quad (1)$$

Here, the normalization factor is the number of transporter pairs. The lower the value of d, the more homogeneous is the corresponding transporter group. The similarity between two sets was measured by the squared Euclidean distance between the mean amino acid compositions (μ) of both sets. Two sets were considered as similar to each other if this distance was below a certain threshold value. For this, we calculated the standard deviation (σ) of the amino acid composition in each transporter group. Because one can assume that most transporters of a group are located in the interval $[\mu - 2\sigma,\mu + 2\sigma]$ (95.4% in the case of a normal distribution), we tested the following condition for every amino acid type k:

$$\mu_{l_k} + 2\sigma_{l_k} \leq \mu_{m_k} - 2\sigma_{m_k} \vee \mu_{m_k} + 2\sigma_{m_k} \leq \mu_{l_k} - 2\sigma_{l_k} \quad (2)$$

If this was a true statement then the two positive sets l and m were considered dissimilar in the content of amino acid k.

Moreover, we aimed at deriving characteristic physicochemical features of the positive sets for which we grouped the 20 amino acid types according to 10 nondisjunct properties. For every group we checked if the values for the transporter sequences in the positive set assume a normal distribution. Afterward, we calculated p values to test the hypothesis that the mean values of two positive sets are the same using the analysis of variance (ANOVA). The hypothesis was rejected if the p value was less than 0.001.

Substrate Prediction. Figure 1 shows a general overview of the calculation steps. From the amino acid composition v_j of every transporter j, an average amino acid composition μ of the positive set is built as a search profile. For this, we used different ways of characterizing the amino acid composition, namely the standard amino acid composition (AAC), the pseudo amino acid composition (PseAAC)

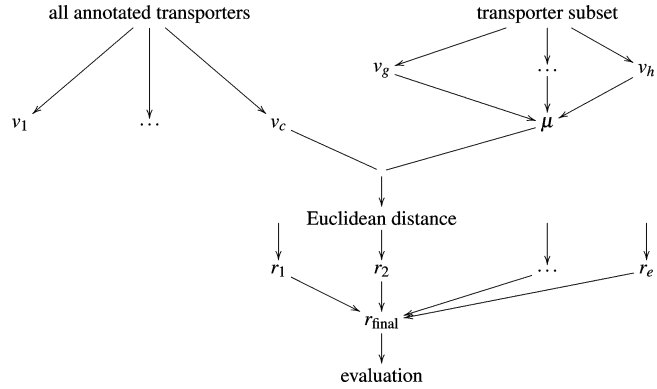

Figure 1. The computational procedure comprises the calculation of the individual amino acid compositions (v_j) for a particular transporter, the average amino acid composition for a positive set (μ), the distance scoring, the construction of rankings (r), and a final evaluation.

introduced by Chou,[14] the pair amino acid composition (PAAC) developed by Park et al.,[15] a combination of PAAC with the λ last entries of PseAAC, termed PsePAAC, and a profile-based version that we called MSA-AAC. The AAC is a vector with 20 entries for the frequencies of the 20 amino acid types. The PseAAC is an extended version of AAC with $20 + \lambda$ entries where the λ additional entries incorporate neighborhood correlations describing amino acid characteristics such as mass, hydrophobicity, or isoelectric point.[14] The PAAC vector contains 400 frequencies for the 20×20 possible pairs of amino acids. PsePAAC consists of $400 + \lambda$ entries, where the first 400 correspond to the frequencies of all amino acid pairs and the other to the neighborhood correlations. Its entries are given in eq 3.

$$v_y = \begin{cases} \dfrac{u_y}{n} & \text{if } y \leq 399 \\ \dfrac{\omega\theta_{y-399}}{n} & \text{else} \end{cases} \tag{3}$$

The two factors ω and n were used for weighting and for normalization of the pair frequencies and the λ additional entries. The correlation factors (θ) describe neighborhood correlations of hydrophobicity, hydrophilicity, side-chain mass, pK values, and isoelectric point as presented by Shen et al.[25] The MSA-AAC method is a profile method that uses a full multiple sequence alignment (MSA) of the corresponding transporter built by ClustalW[26] in the same way as in ref 27. For this, a maximum of 1000 homologous sequences was searched in the nonredundant database using Blast, which were used to build an initial MSA. Sequences with an identity below 25% were removed and a final MSA that we used for our calculations was generated from the remaining sequences. The occurrence of every amino acid in all sequences of the alignment was normalized by the numbers of included amino acids.

The similarity between any transporter sequence (including those of the positive set) and a positive set was measured by a simple Euclidean distance between the amino acid composition v_j of the considered transporter and the mean composition μ of the positive set, see Figure 1. Based on this measurement, a ranking was set up, where the transporter sequence showing the highest similarity to the positive set was put on the top followed by the sequence with second-highest similarity and so on. In this way, multiple rankings were constructed for multiple combinations of representing the amino acid composition and by using subsets of sequences in the positive set. The latter procedure was implemented to avoid systematic shiftings caused by a few outliers. For a positive set with q members, we constructed all possible subsets with $(q - 1)$ members each and all subsets with $(q - 2)$ members each. Additionally, the whole positive set was considered. Hence, we generated

$$1 + \sum_{z=1}^{q} z$$

different methods. The quality of each method was characterized by leave one out cross validation (LOOCV). For this, a single element was deleted from the positive set. The predictions were done for this reduced positive set of size $(q - 1)$, and it was checked whether the removed element was found by the method within the 10 top positions of the ranking. This was repeated for all elements of the positive set. A method was called successful if the LOOCV found at least one element. In this way, a large number of different rankings was obtained from the application of the successful methods. Ideally, those rankings, which are ordered lists, should be combined to one single final ranking. In order to do this in an optimal way, we used a cross entropy Monte Carlo (CEMC) method[21] for merging the rankings. This method integrates sorted lists by an iterative procedure so that the sum of distances between the positions of the entries in the final and the original lists becomes minimal. This procedure was followed by a validation step in which the numbers of correctly and wrongly predicted transporters were detected. Here, we counted the number of positive (true positives (TP)) and negative (false positives (FP)) set transporters among the $2q$ top positions of the final ranking. The numbers of false negatives (FN) and true negatives (TN) were derived from TP, FP, and the number of set members. Then the performance measures sensitivity (sens), specificity (spec), and accuracy (acc) were defined as usual:

$$\text{sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{6}$$

The sensitivity characterizes the likelihood of finding positive set elements and thus is the most important measurement here. The specificity depends on the negative set, and the accuracy combines both. Because the choice of the negative sets is not unique, we validated our results using comparisons with randomly generated "positive sets". For this, we set up a random set of the same size as the corresponding positive set, that contains randomly selected transporters from the full data set of 793 sequences, performed the whole computational procedure, and repeated this 20 times. Then, we averaged the evaluation measurements (e.g., $\overline{\text{sens}}_{\text{ra}}$) and compared these values with those from the actual prediction. We also computed the standard deviation of the random sensitivities ($s_{\text{sens}_{\text{ra}}}$) and calculated the integer value f with the following properties:

$$\overline{\text{sens}}_{\text{ra}} + f s_{\text{sens}_{\text{ra}}} < \text{sens}_{\text{p}} < \overline{\text{sens}}_{\text{ra}} + (f + 1) s_{\text{sens}_{\text{ra}}} \tag{7}$$

where sens_{p} is the sensitivity of the real positive set. Additionally, we performed a one-sample t test with two significance levels $\alpha_1 = 0.001$ and $\alpha_2 = 0.1$ and the null hypothesis

$$\overline{\text{sens}}_{\text{ra}} \leq \text{sens}_{\text{p}}$$

whereby $\overline{\text{sens}}_{\text{ra}}$ is again the mean sensitivity of the random predictions and sens_{p} the corresponding sensitivity of the real positive set. We rejected our hypothesis if the calculated T value was higher than $t_1 = 3.883$ or $t_2 = 1.328$, respectively.

RESULTS AND DISCUSSION

The aim of this work was to derive a sequence-based prediction of the putative substrates transported by an

Table 2. Similarity Scores of the Positive Sets

positive set	score
oligopeptide	0.130
amino acid	0.139
hexose	0.223
phosphate	0.231
full data set	0.374

arbitrary membrane transporter. For this, we first had to derive a mathematical representation of the members of the individual substrate classes.

Analysis of the Data Sets. The substrate predictions need to be based on a mathematical representation of the transporters according to which functionally related proteins are more similar to each other than to the remaining. By comparing the considered transporter groups, we now show that the amino acid composition is a suitable representation for this. The similarity within each set and the similarity between the four positive sets was computed based on Euclidean distances of the amino acid compositions. Table 2 lists the similarities in each positive set. Accordingly, the set of oligopeptide transporters is the most homogeneous set, followed closely by the amino acid transporter set. The set of phosphate transporters is the most heterogeneous set. This relatively large dissimilarity between the phosphate transporters is as expected, because this set contains members of five protein families. For comparison, the averaged similarity score over the full data set is 0.374. The scores of all positive sets are much lower than this score; see Table 2. Hence, functionally related proteins are similar with respect to their amino acid composition.

To quantify the similarity between the positive sets, the squared Euclidean distance between the mean amino acid compositions of the two different groups was computed. On the basis of this score, the oligopeptide transporters were found to be closely related to the amino acid transporters. This is as expected, because oligopeptides consist of a small number of amino acids so that oligopeptides have analogous characteristics to single amino acids. Additionally, the oligopeptide and hexose transporter sets are similar to phosphate transporters. Figure 2 visualizes these similarity characteristics in a graph. Sets connected by dashed lines are similar according to a distance less than 0.0014. On the basis of the considered similarity score, there is no connection between amino acid and oligopeptide transporters to sugar transporters.

Furthermore, we also identified those amino acids that are mainly responsible for the dissimilarities between transporter

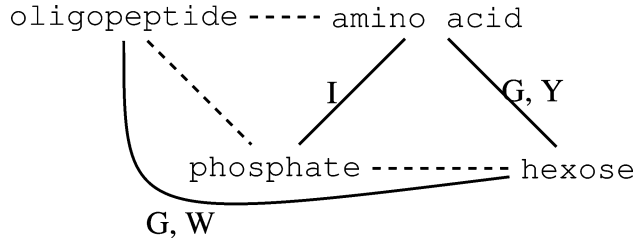

Figure 2. Similarity graph: Two substrate groups are connected by a dashed line if a constitutive similarity exists between them. The straight lines indicate significant dissimilarities between sets whereby the edge labels name those amino acids in one letter code with the largest differences between the two connected transporter groups.

Table 3. Average Characteristic Frequencies in the Positive Sets Where the 20 Amino Acids Are Grouped According to Their Physicochemical Properties[a]

characteristic	amino acid	oligopeptide	phosphate	hexose
tiny	0.264	0.241	0.285	0.266
small	0.519	0.501	0.522	0.505
large	0.205	0.216	0.201	0.184
aromatic	0.147	0.151	0.137	0.123
aliphatic	0.265	0.256	0.239	0.276
positive	0.090	0.094	0.098	0.085
negative	0.059	0.067	0.064	0.071
charged	0.150	0.161	0.162	0.156
polar	0.426	0.446	0.435	0.406
hydrophobic	0.637	0.628	0.606	0.607

[a] Groups are tiny (A, C, G, S), small (A, C, D, G, N, P, S, T, V), large (F, K, R, W, Y), aromatic (F, H, W, Y), aliphatic (I, L, V), positive (H, K, R), negative (D, E), charged (D, E, H, K, R), polar (C, D, E, H, K, N, R, Q, S, T, W, Y), and hydrophobic (A, C, F, I, L, M, P, T, V, W, Y).

Table 4. ANOVA p Values Corresponding to Table 3[a]

pair of positive sets	tiny	small	large	aromatic	aliphatic
amino acid−oligopeptide	0.0013	**0.0008**	0.0203	0.4651	0.1235
amino acid−phosphate	−	0.7023	0.4607	0.0335	0.0046
amino acid−hexose	0.8164	0.0762	0.0077	**0.0008**	0.1368
oligopeptide−phosphate	−	**0.0003**	0.0085	0.0288	0.0225
oligopeptide−hexose	0.0069	0.5839	**0.0001**	0.0012	**0.0007**
phosphate−hexose	−	0.0477	0.0478	0.0388	**0.0002**

pair of positive sets	positive	negative	charged	polar	hydrophobic
amino acid−oligopeptide	0.2890	−	0.0520	0.0049	−
amino acid−phosphate	−	−	−	−	−
amino acid−hexose	0.1196	−	0.2537	0.0211	**0.0006**
oligopeptide−phosphate	−	−	−	−	−
oligopeptide−hexose	0.0206	0.3862	0.4866	≤**0.0001**	−
phosphate−hexose	−	−	−	−	−

[a] If one of the considered sets is not normal distributed for that characteristic, no p value is given. Values in bold are considered of high significance with $p < 0.001$.

groups using eq 2; see Figure 2. For example, the hexose transporters differ from the amino acid transporters mainly in the frequencies of the amino acids glycine and tyrosine. Further, the hexose transporters are different from oligopeptide transporters in the content of glycine and tryptophan.

Table 3 shows which amino acid properties are over- and underrepresented in the individual transporter sets. For this, groups of amino acids with the same characteristics were combined into 10 nondisjoint groups. In general, the differences between the positive sets are relatively small. The hexose transporters contain fewer aromatic amino acids than the other sets and the smallest amount of positively charged, polar, and large amino acids. Further, the amino acid transporter set contains the smallest fraction of negatively charged amino acids. To detect significant differences in these characteristics between positive sets, ANOVA p values are given in Table 4. The hypothesis, that the mean of the two considered positive sets is the same, is rejected if the p value is less than 0.001. Accordingly, the amino acid and the oligopeptide transporter sets differ in the content of small amino acids. Further, the amino acid transporters are different from the hexose transporters in the frequency of aromatic and hydrophobic amino acids. There exists also a significant difference between oligopeptide and phosphate transporters in the content of small amino acids and between oligopeptide and hexose transporters in the frequency of large, aliphatic,

and polar amino acids. Finally, the phosphate and the hexose transporters differ with respect to aliphatic amino acids.

It is reasonable to assume that the characteristics of membrane transporters are related to the physicochemical properties of their substrates. Consequently, transporters with the same function can be expected to be more similar to each other than to other transporters, i.e., one expects a higher similarity within a positive set and a lower similarity between different positive sets. This is exactly what the analysis of our data set indicates. Hence, we showed that the amino acid composition is a suitable measure to distinguish between transporter groups. Along the same line, the small differences found between amino acid transporters and oligopeptide transporters may be related to their similar substrate types (single or multiple amino acids). In our classifications, oligopeptide transporters were sometimes classified as amino acid transporters. This is also observed in reality. For example, At2g02040 is annotated both as an oligopeptide and histidine transporter in the Aramemnon database.

For comparison, we tested whether a similar distinction between substrate-based transporter groups can also be made based on pairwise sequence alignments using BLAST.[22] The average sequence identity between members of the two protein families that transport oligopeptides is 3.01%. In contrast, the average sequence identity of the oligopeptide transporters of both families to members of the other positive sets is 2.78%. The identity between the different families in the amino acid transporter set is 5.92%, whereas the whole set has an identity to the other positive sets of 4.35%. Members of different phosphate transporter families have an identity of 12.24% to each other and 6.46% to members of other positive sets. The different families of the hexose transporters have no detectable sequence identity. The hexose transporter set has an identity of 7.95% to the other positive sets. Generally, all these values are much lower (<15%) than what is usually considered a reliable measure of sequence homology. Given that both values appear indistinguishable, we conclude that sequence alignment-based measures cannot be used to classify substrate specificity of membrane transporters beyond the established protein families.

Substrate Prediction. After having discussed the properties of the positive sets in the previous section, we now discuss the predictivity of the substrate classifications. The results presented in Table 5 were obtained based on the simple Euclidean distance as score measurement. Other ways of scoring led to similar results (data not shown). All methods performed very well with accuracies over 80%. The PseAAC gave similar results as the simple AAC. The profile based method MSA-AAC that uses information from multiple sequence alignments yields a higher specificity and higher or equal sensitivity and accuracy in comparison to the simple AAC method, except for the oligopeptide transporters. The PAAC method that also considers the content of amino acid pairs enhances the prediction conspicuously. Its specificities are similar to those obtained by MSA-AAC, but the sensitivities are higher. Hence, it resulted in very high accuracies of at least 87.5%. Again the use of PsePAAC gave results of similar quality as with PAAC.

For oligopeptide transporters, the predictions often achieved a very high specificity. This is easily explained by the fact that the subset of oligopeptide transporters annotated so far is one of the two most homogeneous sets. All oligopeptide transporters belong to one of only two different protein families. Interestingly, the predictions worked equally well for very heterogeneous sets, namely that of the phosphate transporters. In this case, we found clear differences between the different methods, however. Whereas the simple AAC method did not yield very accurate results, the more elaborate MSA-AAC method and, in particular, the PAAC method provided results of similarly high accuracies (over 90%) as for the homogeneous sets.

For comparison, Table 6 shows the performance of different methods obtained for randomly compiled positive sets. Because the size of the full data set is quite large in comparison to the size of the positive and negative sets used here, the probability of finding negatives is also small for random sets. Hence, we expect a relatively high specificity. In contrast, the sensitivity based on the simple AAC is only 33.7% for a random set corresponding to the size of the amino acid transporter set, whereas the prediction with actual amino acid transporters resulted in a clearly higher sensitivity of 87.5%. A similar remarkable difference was found for the other methods as well. The PAAC or MSA-AAC methods gave a sensitivity of about 26% for the random sets and of about 90% for the actual positive set, respectively.

Table 5. Behavior of Different AAC-Based Measures for Different Transporter Groups: Sensitivity, Specificity, and Accuracy Are Shown

positive set	method	sens	spec	acc
amino acid	AAC	0.875	0.875	0.875
	PseAAC	0.875	0.875	0.875
	PAAC	0.938	0.867	0.903
	PsePAAC	0.938	0.875	0.906
	MSA-AAC	0.875	1.000	0.938
oligopeptide	AAC	0.941	1.000	0.970
	PseAAC	0.882	1.000	0.939
	PAAC	0.933	1.000	0.968
	PsePAAC	1.000	1.000	1.000
	MSA-AAC	0.882	1.000	0.939
phosphate	AAC	0.800	0.667	0.733
	PseAAC	0.800	0.667	0.733
	PAAC	0.933	1.000	0.968
	PsePAAC	0.933	1.000	0.968
	MSA-AAC	0.933	1.000	0.968
hexose	AAC	0.769	0.909	0.833
	PseAAC	0.769	0.909	0.833
	PAAC	0.769	1.000	0.875
	PsePAAC	0.769	1.000	0.875
	MSA-AAC	0.769	1.000	0.875

Table 6. Behavior of Random Computations: Mean (μ) and Standard Deviation (σ) of the Sensitivity and the Mean Values of the Specificity and Accuracy Averaged over 20 Computations with Different Random Positive Sets

positive set	method	$\overline{\text{sens}}_{\text{ra}}$	$s_{\text{sens}_{\text{ra}}}$	$\overline{\text{spec}}_{\text{ra}}$	$\overline{\text{acc}}_{\text{ra}}$
amino acid	AAC	0.325	0.097	0.837	0.580
	PAAC	0.263	0.214	0.753	0.505
	MSA-AAC	0.257	0.064	0.894	0.576
oligopeptide	AAC	0.324	0.092	0.916	0.613
	PAAC	0.308	0.207	0.809	0.555
	MSA-AAC	0.294	0.072	0.916	0.595
phosphate	AAC	0.373	0.095	0.880	0.627
	PAAC	0.370	0.229	0.787	0.578
	MSA-AAC	0.273	0.094	0.930	0.602
hexose	AAC	0.358	0.122	0.918	0.611
	PAAC	0.319	0.297	0.586	0.442
	MSA-AAC	0.273	0.089	0.927	0.573

Table 7. Comparison between Random and Actual Positive Sets[a]

	AAC	PAAC	MSA-AAC
amino acid	5	3	9
oligopeptide	6	2	8
phosphate	4	2	7
hexose	3	1	5

[a] The factor f is defined in eq 7.

Because of the similar size of the amino acid, oligopeptide, phosphate, and hexose transporter sets, the probability of finding their members by chance should be similar. This was checked for the oligopeptide, the phosphate, and the hexose transporters based on the different AAC variations. In fact, the sensitivity for the simple AAC was 32.4% for random positive sets of the size of the oligopeptide transporter set, and 35.8% for the size of the phosphate and hexose transporter sets. Thus, these values are about the same as for the amino acid transporter group. The mean random value for PAAC was even lower than that for AAC, whereas that for MSA-AAC was the lowest. Additionally, the standard deviation was very low with MSA-AAC. The behavior of the predictions particularly for the phosphate transporters shows that our approach is applicable to this problem and performs better than sequence comparisons.

A further comparison between the sensitivity of the actual and the random computations is based on the values of f defined in eq 7, see Table 7. High values (≥ 3) indicate a very clear separation between real predictions and random results. This criterion is fulfilled particularly well for the MSA-AAC method, whereas the mediocre PAAC results may hint at overtraining. The t test showed for all methods independent of the choice of the positive set that the null hypothesis must not be rejected for both confidence intervals. This means that the actual sensitivities are significantly higher that the random ones.

Overall, one may wonder why the relatively simple AAC-type measures have such predictive power. AAC has been widely used to predict protein function,[9,14] subcellular localization,[15] or even protein−protein interaction.[28] With regard to membrane transporters, Gromiha and Yakubi[16] showed that the AAC can be used successfully to classify membrane proteins into different transporter families. Here, we employed the AAC for substrate prediction and also obtained high accuracies. Hence, it seems that the composition includes important information. In our view, such approaches would not work if the AAC was not related to functional properties of proteins and of membrane transporter families, in particular.

CONCLUSIONS

We have developed a novel computational approach to predict substrate specificities of membrane transporter proteins from their amino acid sequence. The method utilizes sophisticated variants of the amino acid composition and constructs a merged ranking list based on the Euclidean distance between the amino acid composition of a test sequence and that of positive sets of transporters for a particular substrate. Our approach has a high accuracy around 90% compared to around 60% for randomized data. The predictivity of our approach is best reflected by the similarly high sensitivity of 80−90% compared to 30−35% for randomized data. Based on this high sensitivity, the predictions should be useful to guide the experimental determination of substrate specificities of given transporters and to detect new transporters of a given substrate class.

Abbreviations: AAC, amino acid composition; PAAC, pair amino acid composition; PseAAC, pseudo amino acid composition; PsePAAC, combination of PAAC with the λ last entries of PseAAC; MSA-AAC, profile-based version based on multiple sequence alignments (MSA).

ACKNOWLEDGMENT

We thank Sikander Hayat for automatic generation of multiple sequence alignments with ClustalW, and Michael Hutter for careful reading of the text.

REFERENCES AND NOTES

(1) Griffith, J.; Baker, M.; Rouch, D.; Page, M.; Skurray, R.; Paulsen, I.; Chater, K.; Baldwin, S.; Henderson, P. Membrane transport proteins: implications of sequence comparisons. *Curr. Opin. Cell Biol.* **1992**, *4*, 684–695.
(2) Bush, D. Proton-coupled sugar and amino acid transporters in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **1993**, *44*, 513–542.
(3) Saier, M., Jr. Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol. Rev.* **1994**, *58*, 71.
(4) Lee, T.; Paulsen, I.; Karp, P. Annotation-based inference of transporter function. *Bioinformatics* **2008**, *24*, i259.
(5) Saier Jr., M. Genome archeology leading to the characterization and classification of transport proteins. *Curr. Opin. Microbiol.* **1999**, *2*, 555–561.
(6) Li, H.; Benedito, V.; Udvardi, M.; Zhao, P. TransportTP: A two-phase classification approach for membrane transporter prediction and characterization. *BMC Bioinformatics* **2009**, *10*, 418.
(7) Ren, Q.; Kang, K.; Paulsen, I. TransportDB: a relational database of cellular membrane transport systems. *Nucl. Acid Res.* **2004**, *32*, D284.
(8) Saier, M., Jr. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Rev.* **2000**, *64*, 354.
(9) Nishikawa, K.; Kubota, Y.; Ooi, T. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *J. Biochem.* **1983**, *94*, 981–995.
(10) Rost, B.; Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **1994**, *19*, 55–72.
(11) Bahar, I.; Atilgan, A.; Jernigan, R.; Erman, B. Understanding the recognition of protein structural classes by amino acid composition. *Proteins* **1997**, *29*, 172–185.
(12) Genfa, Z.; Xinhua, X.; Chun-Ting, Z. A weighting method for predicting protein structural class from amino acid composition. *Eur. J. Biochem.* **1992**, *210*, 747–749.
(13) Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **2005**, *347*, 827–839.
(14) Chou, K. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.* **2000**, *278*, 477–483.
(15) Park, K.; Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **2003**, *19*, 1656–1663.
(16) Gromiha, M.; Yabuki, Y. Functional discrimination of membrane proteins using machine learning techniques. *BMC Bioinformatics* **2008**, *9*, 135.
(17) Ou, Y.; Chen, S.; Gromiha, M. Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. *Proteins* **2010**, *78*, 1789–1797.
(18) Davies, M.; Secker, A.; Freitas, A.; Mendao, M.; Timmis, J.; Flower, D. On the hierarchical classification of G protein-coupled receptors. *Bioinformatics* **2007**, *23*, 3113.
(19) Liu, Q.; Cui, J.; Yang, Q.; Xu, Y. In-silico prediction of blood-secretory human proteins using a ranking algorithm. *BMC Bioinformatics* **2010**, *11*, 250.
(20) Schwacke, R.; Schneider, A.; van der Graaff, E.; Fischer, K.; Catoni, E.; Desimone, M.; Frommer, W.; Flugge, U.; Kunze, R. ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiol.* **2003**, *131*, 16–26.

(21) Lin, S.; Ding, J. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics* **2009**, *65*, 9–18.
(22) Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
(23) Juretic, D.; Zoranic, L.; Zucic, D. Basic charge clusters and predictions of membrane protein topology. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 620–632.
(24) Jones, D. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **2007**, *23*, 538.
(25) Shen, H.; Chou, K. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2007**, *373*, 386–388.
(26) Thompson, J.; Higgins, D.; Gibson, T. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acid Res.* **1994**, *22*, 4673.
(27) Park, Y.; Hayat, S.; Helms, V. Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinformatics* **2007**, *8*, 302.
(28) Roy, S.; Martinez, D.; Platero, H.; Lane, T.; Werner-Washburne, M. Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions. *PLoS ONE* **2009**, *4*, e7813.

CI100243M