CorrelationCheck: A Program Correlating Protein Function to the Distribution of Amino Acid Residue Types Within Multiple Related Proteins.

Taihao Jin and Diomedes Logothetis

The function of all inwardly rectifying potassium channels (Kir) requires PIP_2. The strength and specificity of channel-PIP_2 interaction display remarkable diversity among Kir channels. Identifying residues responsible for this diversity is important for understanding the detailed design of channel-PIP_2 interactions to carry out specific channel functions. In order to select the best candidates responsible for the diversified channel-PIP_2 interactions and carry out detailed experimental studies, we have developed a computer program to identify positions in a multiple sequence alignment, where physical-chemical properties of residues display strong correlation with functional aspects of proteins, such as the strength or specificity of channel-PIP_2 interactions.

In order to evaluate the correlation between protein function and distribution of amino acid types at a given position in a multiple sequence alignment, protein function and physical-chemical properties of amino acids need to be parameterized. Functional aspects of proteins are parameterized by assigning numerical scores to each protein based on the experimental data. In our attempt to select candidate amino acid residues responsible for the specificity of channel-PIP_2 interaction, ten members of the Kir family that were experimentally characterized (Rohács et al., under revision *PNAS*) were grouped into 4 groups based on the specificity of their activation by different phosphoinositides (for

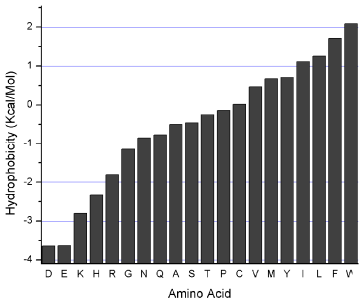

Figure1a. Hydrophobicity indexes the 20 amino acids. Assignment of the parameter, "Hydrophobicity, in our program is based on these values (see text). The figure is made using numbers from Wimley & White 1996.

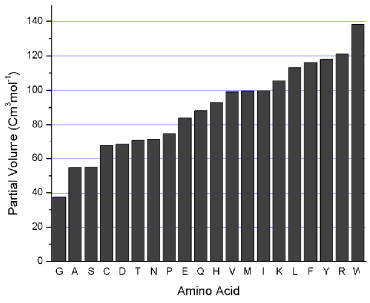

Figure1b. Partial volume of the 20 amino acids. Assignment of the parameter, "Volume", in our program is based on these values (see text).. The figure is made using numbers from Kharakoz 1997.

details, see the example presented below). Each of the four groups is assigned a numeric score, 1, 2, 3, or 4, with the lower value signifying greater specificity (see fig. 3). Then the group score is assigned to each of the group members, and we refer to it as the functional index of the sequence (protein) *i*, denoted as *F(i)* hereafter. Physical-chemical properties of each residue type are described with six parameters. **1.) Charge:** ASP and GLU are assigned as –1, ARG and LYS are assigned as +1 and all of the rest amino acids

are assigned as 0. **2.) Existence of a beta carbon:** ILE and VAL are assigned as 1 and all of the rest are assigned as 0. **3.) Hydrophobicity:** this parameter is assigned based on the hydrophobicity index determined by White and colleagues (Wimley and White, 1996) (see fig. 1a). Amino acids with hydrophobicity lower than –2 are assigned as –1, between –2 and 0 are assigned as 0, between 0 and 1 are assigned as 1, and larger than 1 are assigned as 2. **4.) Volume:** this parameter is assigned based on the partial volume determined by (Kharakoz, 1997) (see fig. 1b). Amino acids with partial volume less than 60 (unit Cm^3mol^{-1}) are assigned as –1, between 60 and 80 are assigned as 0, between 80 and 100 are assigned as 1, and those with hydrophobicity larger than 100 are assigned as 2. **5.) Aromaticity:** Aromatic amino acids PHE, TRP and TYR are assigned a value of 1 and all of the other amino acids are assigned a value of 0. **6.) Hydrogen bonding:** Amino acids that are capable of donating or accepting a hydrogen bond, ASP, GLU, HIS, LYS, ASN, GLN, ARG, SER, THR, TRP and TYR are assigned a value of 1, and the rest amino acids are assigned a value of 0.

Let $R_j(i,k)$ represent parameter j of a residue of sequence i that is located at position k in a multiple sequence alignment. The Pearson coefficient (Rosner, 1995) of correlation between the functional index and parameter j at position k, $r_j(k)$, is calculated with equation 1. The statistical variable, $t_j(k) = r_j(k)(N-2)^{1/2}/(1-r^2)^{1/2}$, follows a t distribution

$$
\begin{aligned}
r_j(k) &= \frac{\delta^j_{R,F}(k)}{\delta^j_R(k)\delta_F(k)} \qquad (1)\\
\delta^j_R(k) &= \sqrt{\sum_{i=1}^{N}(R_j(i,k)-\bar{R}_j(k))^2}\\
\delta_F(k) &= \sqrt{\sum_{i=1}^{N}(F(i)-\bar{F})^2}\\
\delta^j_{R,F}(k) &= \sum_{i=1}^{N}(R_j(i,k)-\bar{R}_j(k))(F(i)-\bar{F})\\
\bar{F} &= \sum_{i=1}^{N}\frac{F(i)}{N}, \quad \bar{R}_j(k) = \sum_{i=1}^{N}\frac{R_j(i,k)}{N}
\end{aligned}
$$

with degrees of freedom $df = N-2$, where N is the number of sequences. Thus, statistical significance, p value, of the correlation at position k can be evaluated based on $t_j(k)$ by looking up the table "Percentage points of the t distribution" that is built in our program. The program scans the multiple sequence alignment, and calculates the Pearson correlation coefficient between the functional index and each of the six physical-chemical properties at each position. Positions at which the calculated correlation coefficient is larger than a predetermined threshold will be selected as potential candidates.

We also tried to consider the possibility that at some positions, the functional index may not correlate well with any single parameter, but may correlate with a certain combination of the parameters. In other words, the functional index may correlate with the "overall properties of amino acids". Therefore, our program also evaluates correlation

$$
\begin{aligned}
r_{\Delta}(l) &= \frac{\delta_{\Delta^R,\Delta^F}}{\delta_{\Delta^R}\delta_{\Delta^F}} \qquad (2)\\
\delta_{\Delta^R} &= \sum_{m=1,n>m}^{N} (\Delta^R_{aa(m,l),aa(n,l)} - \bar{\Delta}^R)^2\\
\delta_{\Delta^F} &= \sum_{m=1,n>m}^{N} (\Delta^F_{m,n} - \bar{\Delta}^F)^2\\
\bar{\Delta}^R &= \frac{2}{N(N-1)} \sum_{m=1,n>m}^{N} \Delta^R_{aa(m,l),aa(n,l)}\\
\bar{\Delta}^F &= \frac{2}{N(N-1)} \sum_{m=1,n>m}^{N} (F(m) - F(n))
\end{aligned}
$$

between the difference in residues properties and difference in functional index. We tried to quantify the "difference" of amino acid based on substitution matrix for sequence alignment algorithms. In the current version, we used the substitution matrix PAM250. The "difference" between amino acid type *i* and type *j* is defined as $\Delta^R_{i,j} = 0.5*(M_{i,i} + M_{j,j}) - M_{i,j}$, where $M_{i,j}$ stands for the substitution matrix element at row *i* and column *j*. At each position of a multiple sequence alignment of *N* sequences, we compare residues from every possible pair of sequences. There are total $C_N^2 = N(N-1)/2$ number of comparisons. Let's use *aa(m,l)* to represent amino acid type of sequence

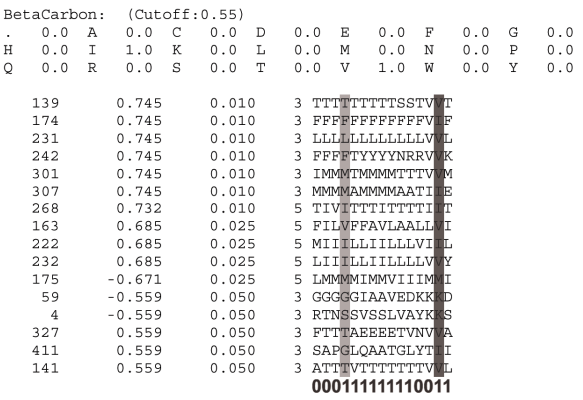

Figure 2. Positions where Pearson correlation coefficients between functional index and existence of beta carbon are larger than cutoff, 0.55 (corresponding p value < 0.05). The upper part of the figure lists the parameter: existence of beta carbon for each of the 20 amino acids. The lower part of the figure lists the positions. In each row: the first column is the position number (Kir2.1 numbering), the second column is the Pearson correlation coefficient, the third row is the p value, the fourth column is the number of hits (see text for definition) and the fifth column lists residues of all of the sequences at the same position. Sequences that are experimentally characterized are indicated in the last row with 1, while the rest of the sequences are marked with 0. Residues of the sequences with the largest (Kir6.2, dark gray) and the second largest (Kir 3.4, light gray) functional index are highlighted.

m at position *l*. Then difference of sequence *m* and sequence *n* at position *l* is $\Delta^{R}_{aa(m,l),aa(n,l)} = 0.5*(M_{aa(m,l),aa(m,l)} + M_{aa(n,l),aa(n,l)}) - M_{aa(m,l),aa(n,l)}$. The Pearson correlation coefficient between differences in residue types and functional indexes can be calculated according to equation 2. If the coefficient is larger than threshold, then the position will be suggested as a candidate.

When the Pearson correlation coefficient is used as the only criterion, the program favors highly conserved (regarding to the given parameter) positions. Figure 2 lists positions where Pearson correlation coefficients between the functional index and the existence of beta carbon are larger than output threshold. The positions are listed in descending order of the Pearson correlation coefficient. As one can see, in all of the six positions that have highest coefficients, Kir6.2 (highest functional index, dark gray) is the only experimentally characterized sequence that has VAL or ILE at those positions. These positions have large coefficients because of smaller value of $\delta_{Rj(k)}$ (see equation 1). Intuitively, positions where the sequence Kir3.4 (second highest functional index, light gray) also possesses the residue VAL or LUE, such as position 222 and 232 etc., correlated better. In order to enable the program to reflect this judgment, we introduced

$$H_j(k) = \left| \sum_{i=1}^{N} h_i(i,k) \right| \qquad (3)$$

$$h_j(i,k) = \begin{cases} 1, & when \ (R_j(i,k) - \bar{R}_j)(F(i) - \bar{F})\delta_{R,F}(k) > 0 \\ -1, & when \ (R_j(i,k) - \bar{R}_j)(F(i) - \bar{F})\delta_{R,F}(k) < 0 \\ 0, & when \ (R_j(i,k) - \bar{R}_j)(F(i) - \bar{F})\delta_{R,F}(k) = 0 \end{cases}$$

an additional criterion, *H(k)*, number of hits as defined in equation 3. As one can see in figure 2, the six positions that have the highest coefficient values have lower hit, *H(k)*, value than the immediately following five positions. Alternatively, we also introduced an optional rescaling to the Pearson correlation function as equation 4.

$$r_j^s(k) = r_j(k) \frac{\delta_R^j(k)}{\delta_R^j(k) + S(\delta_{R_{max}}^j - \delta_R^j(k))} \qquad (4)$$

$$\delta_{R_{max}}^j = Max\{\delta_R^j(k) | k = 1 \cdots M\}$$

As discussed below in the example, rescaling the coefficients with rescale factor *S = 0.3* ranked all positions with a H(k) value of 5 in figure 2, higher than those six positions that have the highest correlation coefficients in figure 2.

The program needs three input files, a multiple sequence alignment, a parameter file assigning functional indexes for sequences in the alignment file, and a parameter

initialization file. The multiple sequence alignment file should be in GCG format. An example of a parameter input file is shown in figure 3. It contains three columns, the sequence names, functional indexes and an integer "key" assigned to each sequence. Assigning a "key" value of 1 indicates the corresponding sequence has been experimentally characterized and thus will be taken into account when evaluating the correlation. A "key" value of 0, on the other hand, indicates the corresponding sequence has not been experimentally characterized and thus will be ignored in the process of evaluating the correlation. The purpose of this "key" variable is to enable the user to use a multiple alignment file containing additional sequences from those that have been characterized. The parameter initialization file is shown in figure 4. The entire file has eight fields. Each field contains a parameter name followed by an assignment of the parameter value to each of the 20 types of amino acids. The deletion mark in the GCG format of multiple sequence alignments, denoted by a dot "." , is also a parameter value. In the current version of the program, however, correlation is evaluated only at positions that do not contain deletions. Users could "comment out" a field by inserting # at the beginning of the parameter name, if they do not wish to check the correlation of a certain parameter. Users can also include additional parameters by adding additional fields. Each additional field should follow the same structure, i.e., a parameter name followed by assignment of the parameter values to the 20 types of amino acids and the deletion mark (i.e. a dot ".").

```
kir3.1      0     0
hKir3.2     0     0
hkir3.3     0     0
hkir3.4     3     1
hKir1.1     2     1
kir2.1      1     1
hkir2.2     2     1
hkir2.3     2     1
hkir2.4     1     1
hkir4.1     1     1
hkir4.2     2     1
hkir5.1     0     0
hKir6.1     0     0
hkir6.2     4     1
kir7.1      2     1
```

Figure 3. A sample input file for functional indexes. In each row, the first column is the sequence name, the second column is the functional index, and the third column is a "key" value indicating whether the sequence will be taken into account in the evaluation of correlations. A "key" value 1 indicates that the sequence will be taken into account and a "key" value 0 indicates that the sequence will not be taken into account.

```
8
Charge:
.   0.0  A   0.0  C   0.0  D  -1.0  E  -1.0  F   0.0  G   0.0  H   0.0  I   0.0  K   1.0  L   0.0
M   0.0  N   0.0  P   0.0  Q   0.0  R   1.0  S   0.0  T   0.0  V   0.0  W   0.0  Y   0.0

BetaCarbon:
.   0.0  A   0.0  C   0.0  D   0.0  E   0.0  F   0.0  G   0.0  H   0.0  I   1.0  K   0.0  L   0.0
M   0.0  N   0.0  P   0.0  Q   0.0  R   0.0  S   0.0  T   0.0  V   1.0  W   0.0  Y   0.0

Hydrophobicity:
.   0.0  A   0.0  C   0.0  D  -1.0  E  -1.0  F   2.0  G   0.0  H  -1.0  I   2.0  K  -1.0  L   2.0
M   1.0  N   0.0  P   0.0  Q   0.0  R  -1.0  S   0.0  T   0.0  V   1.0  W   2.0  Y   1.0

Volume:
.   0.0  A  -1.0  C   0.0  D   0.0  E   1.0  F   2.0  G  -1.0  H   1.0  I   1.0  K   2.0  L   2.0
M   1.0  N   0.0  P   0.0  Q   1.0  R   2.0  S  -1.0  T   0.0  V   1.0  W   2.0  Y   2.0

#Proline:
.   0.0  A   0.0  C   0.0  D   0.0  E   0.0  F   0.0  G   0.0  H   0.0  I   0.0  K   0.0  L   0.0
M   0.0  N   0.0  P   1.0  Q   0.0  R   0.0  S   0.0  T   0.0  V   0.0  W   0.0  Y   0.0

#Cystein:
.   0.0  A   0.0  C   1.0  D   0.0  E   0.0  F   0.0  G   0.0  H   0.0  I   0.0  K   0.0  L   0.0
M   0.0  N   0.0  P   0.0  Q   0.0  R   0.0  S   0.0  T   0.0  V   0.0  W   0.0  Y   0.0

Aromaticity:
.   0.0  A   0.0  C   0.0  D   0.0  E   0.0  F   1.0  G   0.0  H   0.0  I   0.0  K   0.0  L   0.0
M   0.0  N   0.0  P   0.0  Q   0.0  R   0.0  S   0.0  T   0.0  V   0.0  W   1.0  Y   1.0

H-bond:
.   0.0  A   0.0  C   0.0  D   1.0  E   1.0  F   0.0  G   0.0  H   1.0  I   0.0  K   1.0  L   0.0
M   0.0  N   1.0  P   0.0  Q   1.0  R   1.0  S   1.0  T   1.0  V   0.0  W   1.0  Y   1.0
```

Figure 5. Parameter initialization file. The first line contains the number of parameters. The following are eight fields, each of which contains parameter name and parameter value assignment of 20 types of amino acids and the deletion mark, designated by a dot ".".

Reference List

Kharakoz,D.P. (1997). Partial volumes and compressibilities of extended polypeptide chains in aqueous solution: additivity scheme and implication of protein unfolding at normal and high pressure. Biochemistry *36*, 10276-10285.

Rosner,B. (1995). In Fundamentals of Biostatistics, (Boston: Duxbury Press).

Wimley,W.C. and White,S.H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat. Struct. Biol. *3*, 842-848.

An example of running the program **CorrelationCheck**

Bellow is an example of running **CorrelationCheck** to identify positions in a multiple sequence alignment where protein function is correlated with distributions of amino acid residue type. We run the program to identify candidate residues that are possibly important for the diversity of the specificity of channel-PIP_2 interaction among members of the inwardly rectifying K^+ (Kir) channel family.

1. Preparing input files. The GCG format of multiple sequence alignment of 15 Kir family members (listed in fig. 2, and also see fig. 7) were made with CLUSTAWL (at web site: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_clustalw.html). The specificity profile of channel-PIP_2 interaction of most of Kir family members was recently determined (Rohács et al., under revision *PNAS*). The specificity profile of a Kir channel is assayed by determining the effects of different analogues of diC_8 PIP_2, $PI(3,4)P_2$ and $PI(3,4,5)P_3$, expressed as a percentage of the effect of diC_8 $PI(4,5)P_2$. Ten channels whose specificity was studied were divided into four groups. The first group includes channels that show marginal or no activation by $PI(3,4)P_2$ and their activation by $PI(3,4,5)P_3$ is less than 10%. The second group contains channels that show marginal or no activation by $PI(3,4)P_2$, but show significant activation (40-80%) by $PI(3,4,5)P_3$. The third group of channels is activated by $PI(3,4)P_2$, although less so than they are by $PI(3,4,5)P_3$. The fourth group contains channels that are activated by $PI(3,4)P_2$ and $PI(3,4,5)P_3$ even slightly higher than they are by $PI(4,5)P_2$. Channels in these four groups are assigned functional indexes 1, 2, 3, and 4 respectively, as shown in figure 2 (also see fig. 7). Each channel in the alignment file whose specificity profile has been studied is assigned a "key" value 1, while each of the rest of the channels in the alignment file is assigned a "key" value 0, so that they are not taken into account when evaluating the correlation. The amino acid parameter file initialization file was constructed as outlined in the description section (see fig. 5).

2. Running the program. When running the program, users will be prompted to read in the multiple sequence alignment file and the functional index file. The parameter initialization file will be read in without a prompt, thus the file should not be moved to another folder or have its name changed. The program will display a dialogue window where users can change running options. As shown in figure 6, there are five areas in the window where users can enter five groups of parameters. The first area contains two radio buttons, which allow users to decide to use the Pearson correlation coefficient "r value" (default) or a corresponding "p value" as an

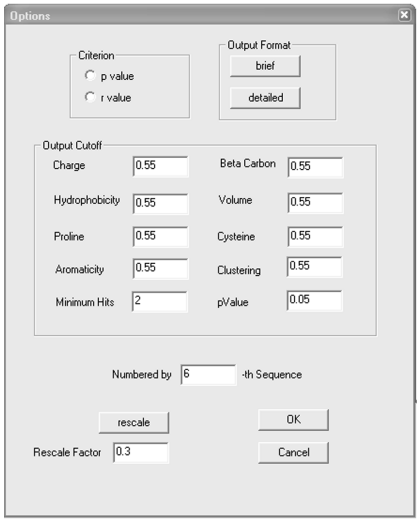

Figure 6. The parameter dialog window

output criterion. The second area contains two buttons to allow users to choose the output format. The brief format (default) is recommended for most situations. Choosing the detailed format will output some intermediate results for more detailed analysis. The third area is where users enter the output cutoff of the correlation coefficient for different parameters. The values shown in figure 6 are default values. In the fourth area, there is an edit box where users can decide numbering of which sequence in the alignment will appear in the output file. The fifth area contains a button and an edit box. If the "rescale" button is pressed, then the program will rescale the Pearson correlation coefficient, using the value in the edit box "Rescale Factor" (see equation 4). Finally, users will be prompted to choose the location and file name of the final output file.

3. Positions suggested based on **CorrelationCheck**. The output file contains a head section and field for each parameter. The head section contains the information in the functional index input file, degree of freedom, and a table that lists a series of Pearson correlation coefficients, the corresponding statistical value $t = r(N-2)^{1/2}/(1-r^2)^{1/2}$ and p value. The field for each parameter is similar to what is shown in figure 2. In figure 7, the positions shown are those where the functional index highly correlated with different physical chemical properties of residue types. Four positions are suggested based on the correlation between specificity and the charge of residues. At **position 51**, the most nonspecific channel Kir6.2 has a positively charged residue K, the secondly nonspecific channel Kir3.4 has a neutral polar residue S, while all three of the most specific channels, Kir2.1, Kir2.4 and Kir4.1 have the negatively charged residue D. At **position 64**, both of the most nonspecific and the second nonspecific channel, Kir 6.2 and Kir3.4 have negatively charged residue E, while two of three most specific channels, Kir2.1 and Kir4.1, have the positively residue K. At **position 185**, both of the most nonspecific and the second specific channel, Kir6.2 and Kir3.4, have a neutral polar residue Q, while all the rest of the channels have the positively charged residues K or R. At **position 216**, both the most nonspecific and second nonspecific channels, Kir6.2 and Kir3.4, have negatively charged residues, while all the rest of the channels have a neutral polar residue N. Three positions, 222, 232 and 268 are suggested as candidate positions based on correlation of functional index with the existence of a beta carbon. **Position 222 and 232** are most favored by visual examination, not only because all of the nonspecific channels have residue I or V, but the most specific channels have residue L in that position. Therefore it is convincing that the major difference between residues of nonspecific channels and specific channels is the existence of a beta carbon. Position 268 also shows high correlation with hydrophobicity and H-bonding parameters. Two positions, 58 and 268 are suggested based on the correlation between the specificity and hydrophobicity of residues. At **position 58**, both of the nonspecific channels Kir6.2 and Kir3.4 have the large hydrophilic residue H, while the rest of the characterized channels have the large or medium size hydrophobic residues F, I or M. At **position 268**, both of the nonspecific channels Kir6.2 and Kir3.4 have a big size hydrophobic residue, I, while all of the most

specific channels have the medium size more hydrophilic residue T. Position 181 is suggested for correlation between specificity and size of residue. At **position 181**, the nonspecific channels Kir6.2 and Kir3.4 have a medium size nonpolar residue M and V, while all of the specific channels have the small nonpolar residue, A. Position 180 is selected based on the correlation between specificity and aromaticity. At **position 180**, nonspecific channels Kir6.2 and Kir3.4 have aromatic residue F, while all the rest have the large or medium size hydrophobic residues L, M, or V. Two positions, 58 and 268 are selected based on the correlation between specificity and the capability of side chain of residues to form

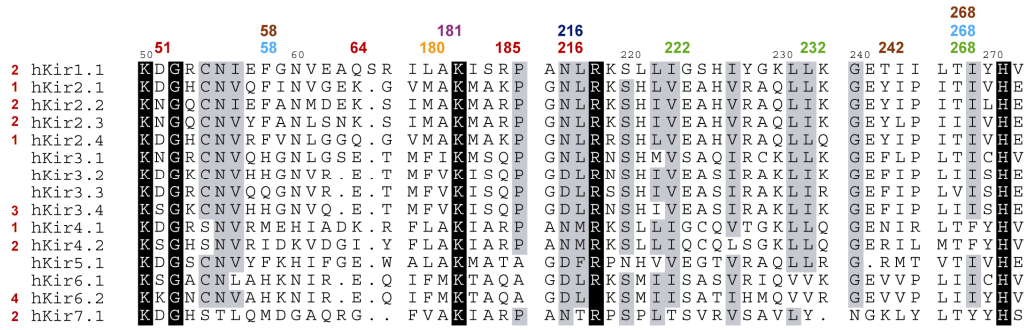

Figure 7. Suggested candidate positions based on the correlation between function and residue type distribution. Functional indexes of the sequences whose specificity profiles are determined are given at the first column. Colors of the labels indicating the suggested candidate positions match the types of parameters listed at the left bottom of the figure.

hydrogen bonds. Both positions are also selected based on the hydrophobicity. At **position 58**, the residue found in the nonspecific channels Kir6.2 and Kir3.4 is H, whose side chain is capable of forming a hydrogen bond, while the residues of specific channels are not. At position 268, residues of specific channels are capable of forming a hydrogen bond, while residues of the nonspecific channels are not. Position 216 is selected based on the correlation between differences in specificities of channels and the differences of residue types (see equation 2). This position is also selected based on the correlation between specificity and charge.