BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1562–1568
doi:10.1093/bioinformatics/btl132

Sequence analysis

Computational recognition of potassium channel sequences

Burkhard Heil[1,*], Jost Ludwig[1], Hella Lichtenberg-Fraté[1] and Thomas Lengauer[2]

[1]Universität Bonn, IZMB, Kirschallee 1, 53115 Bonn, Germany and [2]Max Planck Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany

Received on July 11, 2005; revised on March 30, 2006; accepted on March 31, 2006
Advance Access publication April 4, 2006
Associate Editor: Alfonso Valencia

Use a binary way of encoding sequer

ABSTRACT

Motivation: Potassium channels are mainly known for their role in regulating and maintaining the membrane potential. Since this is one of the key mechanisms of signal transduction, malfunction of these potassium channels leads to a wide variety of severe diseases. Thus potassium channels are priority targets of research for new drugs, despite the fact that this protein family is highly variable and closely related to other channels, which makes it very difficult to identify new types of potassium channel sequences.

Results: Here we present a new method for identifying potassium channel sequences (PSM, Property Signature Method), which—in contrast to the known methods for protein classification—is directly based on physicochemical properties of amino acids rather than on the amino acids themselves. A signature for the pore region including the selectivity filter has been created, representing the most common physicochemical properties of known potassium channels. This string enables genome-wide screening for sequences with similar features despite a very low degree of amino acid similarity within a protein family.

Availability: The PSM software will be made available on request from the corresponding author.

Contact: Burkhard.Heil@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

INTRODUCTION

Ion channels are responsible for maintaining different concentrations of ions on either side of the membrane, resulting in a positively charged extracellular side and a negatively charged inside. This difference in ion concentration results in what is known as the resting potential, based on which signals can be created and transduced. A drop of the potential difference below a certain threshold creates the so-called action potential, which is the basis for sending stimuli along the cell membrane.

This potential can also be considered as a form of energy storage which is used in many other important cellular functions. Opening or closing of ion channels changes the potential and therefore can be used to activate different metabolic pathways, e. g. with calcium as a second messenger (Chay *et al*., 1990).

Potassium channels are key elements in maintaining and regulating the membrane potential (Yi *et al*., 2001). Owing to the role which potassium channels play in a great variety of important cellular processes many severe diseases are caused by malfunctions of potassium channels such as numerous heart disorders, e.g. Long-QT-Syndrome, different forms of epilepsy, deafness, cognitive disorders, ataxia and many more (Ashcroft, 2000). The importance of potassium channels becomes evident when observing that ~1% of all OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/Omim; Hamosh *et al*., 2000) entries are related to potassium channels. Both their crucial importance and the severity of diseases caused by possible malfunctions render potassium channels as priority targets for new drugs (Junker *et al*., 2002).

Rather than actively transporting ions, potassium channels provide a pore through the membrane and enable a passive flow of potassium ions through the membrane that is controlled by opening and closing the pore. The pore is formed by four discrete domains that are localized on the potassium channel α-subunits. On the extracellular side of the pore, a selectivity filter consisting of four amino acids is located. This filter allows only potassium ions to pass through. The side chains of these amino acids form hydrogen bonds with other parts of the protein and stretch the backbone in such a way that carboxy oxygens of the backbone can replace water oxygens of the hydrated K^+ ion. As a result, only dehydrated K^+ ions can pass through this filter and the arrangement of the backbone oxygens ensures that only K^+ ions can be dehydrated. This mechanism implements the selectivity (Doyle *et al*., 1998). The actual pore is the same as in other cation channels. It forms a water-filled cavity inside the membrane in order to decrease the energetic barrier that charged molecules have to overcome to pass the hydrophobic part of the membrane.

Owing to the ancient origin of potassium channels they emerge in many different topologies, sharing only the existence of a single pore and the selectivity filter. The known α-subunits comprise one or two pore domains, resulting either in tetrameric or dimeric channels. Up to now potassium channels with two, four, six, seven and eight transmembrane domains have been identified (Miller, 2000). The functional potassium channel might associate with other subunits, e.g. providing activation domains. For more details on potassium channels see Choe (2003) and MacKinnon (2002). Prior to the development of Property Signature Method (PSM), identification of potassium channels was based on signatures for potassium channel families. The InterPro database (Mulder *et al*., 2003) provides multiple entries for potassium channels which describe different potassium channel families with more or less specific signatures. The same accounts for the emotif system Huang and Brutlag, 2001, (http://motif.stanford.edu/emotif/). Our method is the only method specializing on the identification of new potassium channel genes. By using a special signature and an

*To whom correspondence should be addressed.

Potassium channels	Voltage-gated	208 (80)
	Inward-rectifier	87 (29)
	Double-pore (2+2)	59 (30)
	Double-pore (6+2)	1 (1)
	Calcium-dependent (SK/IK)	16 (6)
	Calcium-dependent (BK)	39 (5)
	Kcsa + MthK	2 (2)
	Kch	14 (11)
	Hyperpolarization-activated	32 (20)
	unclassified	3 (3)
	Σ	461 (187)
other	Potassium-channel associated	188
	Calcium channel	169
	other Channels	9
	unspecified	591
	Σ	957

Fig. 1. Composition of the dataset. All sequences were ex-tracted from Swiss-Prot (Bairoch and Apweiler, 2000). The potassium channels represent the different families and topologies of known potassium channels. For the cross validation the non-potassium channels were used as false positives and all sequences with >80% sequence similarity were removed from the potassium channels (number of remaining channels in brackets). The (2 + 2) double-pore channels consist of two α-subunits with four transmembrane-domains each, the α-subunits of (6 + 2) double-pore channels possess eight transmembrane domains. The three unclassified potassium-channels cannot be unambiguously classified. The unspecified proteins were randomly chosen from Swiss-Prot.

enhanced searching algorithm PSM outscores conventional methods.

METHODS

Datasets

The dataset used for development of the PSM comprises 461 potassium channel α-subunits representing different families (Fig. 1). For the purpose of cross validation only potassium channel sequences with a pairwise sequence similarity <80% were used (187 sequences). In addition, the set contains 957 non-α-subunits, thus providing false positives (Fig. 1). In detail, these sequences include closely related ion-channels, proteins binding to potassium channels and randomly chosen proteins. The latter were included to ensure that the signature discriminates between potassium channels and all other proteins, and not only between potassium channels and related sequences. All sequences were extracted from Swiss-Prot (Bairoch and Apweiler, 2000); duplicate sequences were eliminated (the sequences are available in fasta format as a Supplement).

Potassium channel profile

A profile of the potassium channel pore region was created using the dataset described above. This profile is not used to describe the conserved amino acid positions in this region. Rather it describes all variations found in the different potassium channel families.

Algorithm

This profile is then translated into a descriptor, describing the different properties found in this sequence region. The amino acids at each position of the profile are analyzed and the properties whose absence or presence are conserved are used to describe this position. To locate regions in target sequences matching this descriptor, the algorithm searches for regions exhibiting the same conservation of properties independently from the amino acid composition. Hits are ranked according to the number ofproperties that are found in both, the property descriptor and the target sequence.

The screening algorithm was implemented in C++. A screen of yeast genomes takes roughly a minute on an Athlon 1800XP machine with 512 MB RAM.

Validation

The PSM was validated using 10-fold cross validation. The potassium channels of the dataset were randomly split into 10 equally sized sets. Each of the set was used as a test set while the remaining were used for training. The non-potassium channels were used as false positives. The performance was characterized using sensitivity and specificity.

THE PROPERTY SIGNATURE METHOD

The PSM uses an amino acid representation via a binary signature derived from physicochemical properties. Specifically, 23 properties are used and combined into five groups: side chain type, functional properties, secondary and tertiary structure preference and size (Fig. 2). For each amino acid a binary string is created in which a bit is set to 1, if the corresponding property applies to the amino acid. Altogether five bits are set, one for each property group. The remaining bits are set to 0. This property encryption results in 20 unique bit strings, one for each amino acid, which are used in the algorithm.

Owing to the small number of known potassium channel structures and the difficulty of modeling the structure of membrane proteins, the algorithm is strictly based on the amino acid sequence. As outlined in the introduction, using domain composition has limited effect since the known potassium channels differ highly in this regard. The only region providing a sufficient level of conservation is the pore domain including the selectivity filter.

The actual method is divided into two steps. First, a profile of the aligned pore domains is created which includes all amino acids present in at least 3% of the 461 potassium channels. Second, this profile is translated into a string representing the physicochemical properties of the sequences. Both steps are now described in detail.

In a preparatory first step, an alignment of the pore region including the selectivity filter of the 461 potassium channels is generated which is used to create an unweighted sequence profile. This alignment contains no gaps since the structural demand on the pore and the filter prevents insertions or deletions. The profile covers only 25 sequence positions and includes the pore helix and the selectivity filter. Figure 3 shows a sequence logo representation (Schneider and Stephens, 1990) of these 25 positions. It becomes evident that in most parts of this region amino acids are not conserved but properties like hydrophobicity and polarity, respectively, are. For the exact position of the profile within the pore domain see Figure 4. The number of 25 amino acids seems to be small for representing a characteristic motif for such a diverse family. However, we did not consider a longer profile useful since the already low conservation level decreases significantly beyond the N-terminal end of the pore helix and C-terminal of the selectivity filter (Fig. 4). The profile contains all amino acids that occur at the respective position in >3% of all 461 potassium channels of the dataset. This threshold was chosen to prevent highly untypical amino acids to influence the construction of the profile. On the other hand 3% is still small enough considering that smaller potassium channel families like KCH or Ca^{2+} dependent potassium channels are taken into account, as well. Incorporating diverse sequence positions into the motif contrasts with common profiles, in which only highly conserved

		bit string	side chain type: aromatic	aliphatic	sulfur	imino	amide	hydroxyl	tertiary structure preference: acidic	basic	exposed	functional properties: intermediate	buried	secondary structure preference: acidic	basic	hydrophobic	polar	size: α-helix	β-strand	Coil	tiny	small	medium	large	very large
A	Alanine	10000000001001010010000	1									1				1			1		1			
C	Cysteine	00100000001000101001000			1							1				1			1			1			
D	Aspartate	00000010010000000100100								1			1		1					1			1		
E	Glutamate	00000010100100010000010								1			1		1				1					1	
F	Phenylalanine	01000000010001001000001		1								1				1				1					1
G	Glycine	10000000001000100110000	1										1					1			1				
H	Histidine	00000001100000110000010									1	1					1		1					1	
I	Isoleucine	10000000010001001000100	1									1				1			1				1		
K	Lysine	00000001100010010000010									1	1			1				1					1	
L	Leucine	10000000010001010000100	1									1				1			1				1		
M	Methionine	00100000010001010000010			1							1				1			1					1	
N	Asparagine	00001000100000100100100						1				1				1				1			1		
P	Proline	00010000001001000101000					1						1			1				1		1			
Q	Glutamine	00001000100000110000010						1				1				1			1					1	
R	Arginine	00000001100010010000001									1	1			1				1						1
S	Serine	00000100001000100101000							1				1			1				1		1			
T	Threonine	00000100001000101001000							1				1			1			1			1			
V	Valine	10000000010001001001000	1									1				1			1			1			
W	Tryptophan	01000000001001001000001		1									1			1			1						1
Y	Tyrosine	01000000001000101000001		1									1				1		1						1

Fig. 2. Bit string representation of the amino acids composed of 23 properties. Secondary and tertiary structure preferences were taken from Stryer (1995) The relative frequency of occurance in such a state was converted into binary values by majority vote. In 'size' the amino acids were categorized according to their molecular weight: tiny when ≤71 Da, small when ≤103 Da, medium when ≤115 Da, large when ≤137 Da and very large when >137 Da.

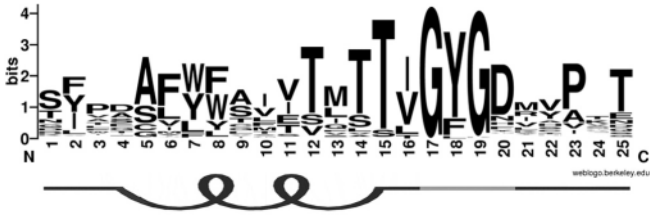

Fig. 3. Sequence logo representation (Schneider and Stephens, 1990) of the 25 sequence positions used for the signature. Even though there is high sequence variability within this region, certain properties like hydrophobicity or polarity are conserved. This suggests a property-based approach.

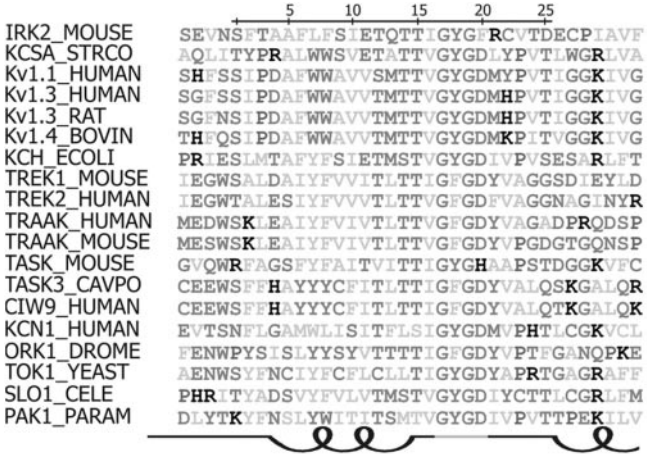

Fig. 4. This figure shows representative pore sequences from potassium channel α-subunits of different families. The secondary structure is schematically shown below with the selectivity filter in light gray. The names left of the sequences refer to entries in the SwissProt database. The numbers on top of the alignment show the region which was used to create the signature (from positions 1 to 25). Hydrophobic amino acids are shown in light gray, polar amino acids in gray and charged amino acids in black. Only the four amino acids of the selectivity filter exhibit a high level of conservation. Within the pore region properties like 'hydrophobic' or 'acidic' are conserved rather than individual amino acids.

residues are included. But the resulting profile is not used for the actual screening, rather it represents a preliminary selection of amino acids.

In the second step the profile is translated into a signature string which represents a consensus of the physicochemical properties of the amino acids within the sequence profile. This can be accomplished in a straightforward fashion using the bit strings defined in Figure 2.

For each sequence position and property-column, the numbers of '1's and '0's is determined. If this number exceeds a certain significance threshold (set >50%), a '1' or '0' is included in the signature, respectively. If neither the number of ones nor the number of zeros exceeds the significance threshold, a '.' (dot) is added to the signature (Fig. 5). Such positions show no clear tendency towards a certain property or towards the lack of that property. The significance threshold can be set adaptively, usually ranging >60% of the number of amino acids at this profile position. If the significance threshold is set low many positions are filled with dots '.'.

In the resulting signature a '1' indicates a conserved property which is regarded as crucial for the function of a potassium channel. A '0' represents a property which seems to interfere with the function. Dots in the signature indicate properties with no clear relation to the function of potassium channels. This signature now represents the physicochemical properties of the pore region of potassium channel α-subunits and can be used to screen genomes for unidentified potassium channels.

When a sequence is compared with this signature, the amino acids of the sequence must be first translated into the bit strings according to Figure 2. Thereafter, the bit string resulting from the query

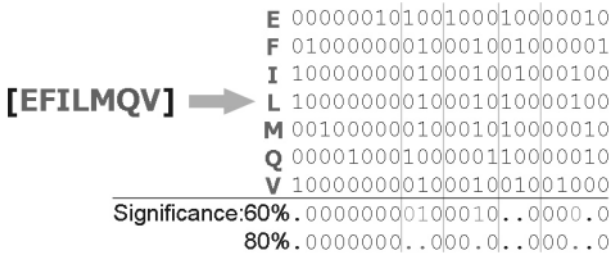

Fig. 5. A consensus signature is created similar to a consensus for amino acid sequences. The bit strings of the amino acids of one profile position are aligned and zeros or ones, respectively, are added to the signature if their frequency exceeds a certain threshold. If neither one passes the threshold, a dot is added. Obviously, the number of dots increases when the significance threshold is raised.

sequence and the signature are compared character by character. Two types of mismatches can arise when comparing a translated amino acid sequence to the signature. In the first case, a '1' in the signature is matched by a '0' in the query sequence. Such a property is called missing property, since the query sequence is missing a property which is conserved throughout the majority of the known potassium channel α-subunits. In the other case, a '0' in the signature is matched by a '1' in the query sequence. Such a property is called an unusual property since the query sequence shows a feature which is very uncommon to known potassium channel α-subunits. The score is bivariate and consists of the numbers of the missing properties and the unusual properties (see Fig. 5). Since the dots represent properties that show no clear tendency, the respective positions are not used for scoring.

The classification of tested sequences is accomplished using a k-nearest neighbor method. The distance of two test sequences is calculated using the Euclidean metric (mp = number of missing properties, up = number of unusual properties):

$$D_{i,j} = \sqrt{(\mathrm{mp}_i - \mathrm{mp}_j)^2 + (\mathrm{up}_i - \mathrm{up}_j)^2}. \quad (1)$$

Tested sequences were classified as potassium channels, when at least 5 of their 10 nearest neighbors were potassium channel α-subunits sequences (Hastie *et al*., 2001).

Together with the score, a Kyte–Doolittle-Plot (Kyte and Doolittle, 1982) of the corresponding region is created for each classified sequence. This plot shows the level of hydrophobicity as a sliding average over 17–22 amino acids. Transmembrane regions show up in the plot as peaks. More importantly, pore regions appear as shoulders of a peak representing a transmembrane region (see Supplementary Figure 1S for an example plot). In addition, the distribution of the missing and unusual properties is shown in order to indicate at which sequence position these mismatches occurred (mismatches at the highly conserved selectivity filter should be taken more seriously than in the less conserved pore helix).

RESULTS

For the validation of PSM a 10-fold cross validation (Hastie *et al*., 2001) using the dataset described above was carried out. Only sequences with a pairwise similarity <80% were used. The 187

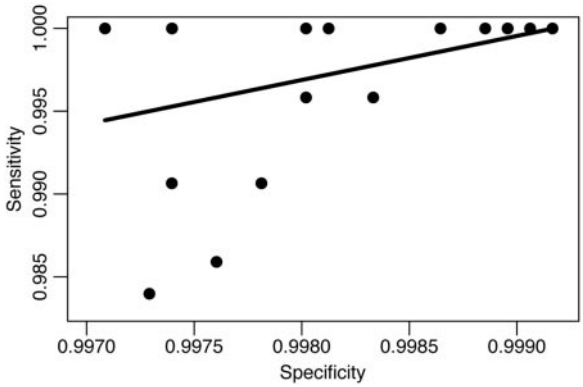

Fig. 6. ROC curve for the classification of the test sequences (regression line is shown in black). Sensitivity [Equation (2)] and specificity [Equation (3)] remain high throughout all significance levels. A change of the significance level has multiple effects on the consensus string and resulting in a non-monotonic curve.

potassium channels were randomly split into 10 equally sized sets. Each of these sets was used as a test set while the remaining sets were used for training. Pairwise sequence similarity within the sets and between the sets were ~30% (SD < 1%). The non-potassium channels were added as false positives to the test sets.

The performance of a method can be characterized by the two terms sensitivity and specificity, which are defined as follows (Hastie *et al*., 2001):

$$\text{sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}} \quad (2)$$

$$\text{specificity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}} \quad (3)$$

Sensitivity is the fraction of positives in the test data that are predicted as positive. Specificity is the fraction of negatives in the test data that are predicted as negative. In the receiver operating characteristic (ROC) in Figure 6, sensitivity is plotted against specificity to summarize the results of the cross validation. The data points in this figure represent the average of all 10 cross-validation runs. The variance (data not shown) is for all significance thresholds very low (maximum for sensitivity: 8.78E−04, maximum for specificity: 3.2E−06). An influence on the results due to the composition of the trainings sets can therefore be excluded.

As is evident from Figure 6, the separation of potassium channel α-subunits and other sequences is always guaranteed, almost independently of the significance threshold, to result in very high values for sensitivity and specificity. The reason for the non-monotonic behavior is the bivariate score. While the number of set bits increases monotonically with descending significance level, the number of unset bits can even drop. There is at maximum one bit set to 1 per block, but all positions of one property block could be set to 0 in some scenarios (e.g. eight amino acids, all

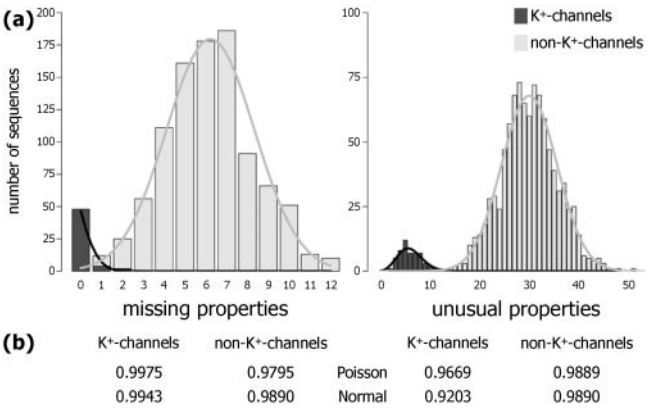

Fig. 7. Distribution of missing and unusual properties within the test sets. Bars indicate the number of properties, the continuous lines depict the fit to a suitable distribution (normal and Poisson, respectively). (**a**) shows a nearly perfect separation of the potassium channel α-subunits and the other sequences. The correlation coefficients for the fit to normal and Poisson distribution are presented in (**b**). In both cases, the potassium channels represent rather a Poisson distribution than a normal distribution; the non-potassium channels show a contrary behavior.

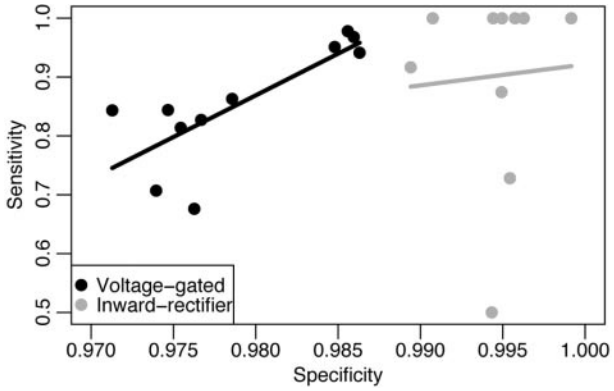

Fig. 8. ROC curves for the separation between certain groups of potassium channels and other potassium channels. True positives for the black points are voltage-gated potassium channels and for the gray points inward-rectifying potassium channels. The lines represent linear regression lines. Even though a decrease in sensitivity and specificity in comparison with the general detection of potassium channels is apparent, there is still clear discrimination between both groups and the other potassium channels.

with different chemical properties, will create a group consisting completely of zeroes at thresholds <87.5% significance). Since this block is unsatisfiable, all positions of this block will be set to '.'. This explains the missing monotony in the number of encoded properties and in the ROC curve as well. Also, increasing sensitivity does not imply descending specificity because the addition of certain properties can increase both, sensitivity and specificity. Therefore, the user can choose between a high significance threshold (only highly conserved properties are considered) and a lower significance threshold without losing specificity or sensitivity. In practice, a significance threshold of 80% has proved to provide a suitable compromise between too many lowly conserved properties and too few highly conserved properties.

Another criterion for the performance of PSM is the distribution of the scores of potassium channels and non-potassium channels, respectively (Fig. 7). Here the scores indicate the number of missing and unusual properties. The potassium channel scores follow a Poisson distribution, whereas the scores of non-potassium channels are distributed normally. This reflects the discriminatory power of the signature. The Poisson distribution is a clear sign of a biased comparison, while the normal distribution of the non-potassium channel α-subunit is an indication of a random comparison against the signature.

Furthermore, 10-fold cross validation also revealed that PSM is capable of discriminating between different groups of potassium channel sequences. Although there is a loss in sensitivity and specificity (Fig. 8), the resulting values arestill sufficiently high to recognize special groups of potassium channels. This was tested by creating property signatures using only members of the corresponding families. For cross validation, voltage-gated potassium channels and inward-rectifying potassium channels, respectively, were used as true positives and the remaining potassium channels as false positives. In this experiment, the test and trainings sets contained also sequences with a pairwise similarity >80%, otherwise the number of sequences would have been too low. The other groups of potassium channels are too small to obtain a reliable result.

PRACTICAL PERFORMANCE OF PSM

Comparison to conventional methods

In order to compare the performance of our method with existing methods, the test database was searched with a conventional pattern recognition method using the NPS@ Web Server (http://npsa-pbil.ibcp.fr). The seven potassium channel specific signatures of the InterPro entry IPR001622 (Mulder *et al*., 2003) were used as the search term. The test database was searched with each signature and the hits of all searches were added; duplicate hits were only counted once. Using the lowest stringent parameter configuration, this method was able to recover ~90% of the potassium channels in the test database—at the price of 30% false positives. PSM had a 10-fold lower false positive rate—even when recovering 99% of the potassium channels (no false positives when the parameters were adjusted to recover ~90%).

A direct comparison of the results of the cross validation to emotif (Huang and Brutlag, 2001) cannot be carried out. The emotif system does not allow for screening a target sequence set provided by the user. Emotif provides a wide range of potassium channel signatures. However, these are specific for certain families and return only very few sequences matching the signature–even if mismatches are allowed. These signatures represent only a fraction of the potassium channels and thus are not suited for a genome-wide screen for unknown potassium channels, which can differ strongly from known potassium channel sequences. Therefore, they might not be identified as potassium channels by emotif. Other signatures contain transmembrane regions and are too unspecific to beused for genome-wide screening. Using the 25 amino acid region to create a signature with emotif results in highly unspecific signatures, since the majority of the relevant 25 positions contain a wide variety of amino acids which do not match any of the emotif substitution

signature position	#	conserved properties 60%	80%
$[DGNRST]_1$	6	polar,loop	-
$[FILVWY]_2$	6	internal, hydrophobic, β-strand	hydrophobic, β-strand
$[AFIL\ STVW]_3$	9	hydrophobic	-
$[ADEGHIST]_4$	8	-	-
$[ACGS]_5$	4	no tertiary., polar	no tertiary.
$[FILMVY]_6$	6	internal, hydrophobic, β-strand	internal, hydrophobic
$[FLWY]_7$	4	aromatic, hydrophobic, β-strand, very large	-
$[FKLWY]_8$	5	aromatic, hydrophobic, β-strand, very large	-
$[ACGILSTV]_9$	8	aliphatic, no tertiary.	-
$[FILMSTV]_{10}$	7	internal, hydrophobic	-
$[EISTV]_{11}$	5	β-strand, small	-
$[HSTV]_{12}$	4	polar, small	-
$[EFILMQV]_{13}$	7	internal, hydrophobic	-
$[ALSTV]_{14}$	5	aliphatic, no tertiary., hydrophobic, small	-
$[CST]_{15}$	3	hydroxyl, no tertiary., polar, β-strand, small	no tertiary., polar, small
$[ILTV]_{16}$	4	aliphatic, internal, hydrophobic, β-strand	-
$[G]_{17}$	1	aliphatic, no tertiary., polar, loop, very small	aliphatic, no tertiary., polar, loop, very small
$[FLY]_{18}$	3	aromatic, internal, hydrophobic, β-strand, very large	-
$[G]_{19}$	1	aliphatic, no tertiary., polar, loop, very small	aliphatic, no tertiary., polar, loop, very small
$[DFNRSY]_{20}$	6	-	-
$[IKLMQRVY]_{21}$	8	α-helical	-
$[ACHRSTVY]_{22}$	8	no tertiary., polar	-
$[AI\ V]_{23}$	4	aliphatic, hydrophobic	hydrophobic
$[EGHIKLNQSTVY]_{24}$	12	-	-
$[DEGNQST]_{25}$	7	polar	-
Σ properties		63	19

Fig. 9. Conservation of properties at 60 and 80% significance level, respectively. A scheme of the secondary structure is drawn left of the signature positions. Despite the low amino acid conservation there are properties which are conserved in 80% of the sequences. As expected from the analysis of potassium channel pores (Doyle *et al*., 1998; Jiang *et al*., 2002), hydrophobic residues dominate the pore, with a few polar residues to decrease the energetic barrier for the charged K^+ ions.

groups (Wu and Brutlag, 1996). The 30 best signatures created with the emotif standard configuration contain about thirty percent variant positions.

Screening genomes

The genome of *Saccharomyces cerevisiae* was screened using PSM. Both of the only two hits found were the pore domains of the two-pore potassium channel TOK1, the only known potassium channel of *S.cerevisiae*. Despite the close relationship and quite high homology of the two potassium transporters TRK1 and TRK2 to the potassium selective pore domains of TOK1, those two were correctly classified as non-potassium channels.

Another test was performed with *Caenorhabditis elegans* whose complete genome sequence was published in 1998 (Hodgkin *et al*., 1998). Its genome is well understood in terms of potassium channel sequences: about 40 two-pore-domain potassium channels are annotated. All of them were recovered using PSM, additionally one new potential pore domain was identified.

Analysis of the potassium channel signature

Figure 9 depicts a summary of the conserved properties at 60 and 80% conservation threshold. Despite the high divergency in the set of sequences, 63 properties are conserved at the 60% significance level and 19 properties are conserved at the 80% significance level. Not shown are the unusual properties coded in the signature (about 350 properties at 60% significance level and 330 properties at 80% significance level). These properties contribute significantly to the specificity of the method.

Within the transmembrane part of the pore, hydrophobicity is conserved at several positions, but also a few, scattered polar residues can be found. These results are in accordance with the analyzed crystallized potassium channel structures of KcsA and MhtK (Doyle *et al*., 1998; Jiang *et al*., 2002). The pore-helix is dominated by hydrophobic residues. Only a few residues are polar in order to reduce the energetic barrier which charged ions like K^+ have to overcome when passing through the membrane. Another property which seems to play an important role, is amino acid size. At 60%, nine positions require a substantial size of the amino acid. The size is important for the amino acids of the selectivity filter and influences the diameter of the pore, as well.

DISCUSSION

The PSM was developed in order to respond to the limits of conventional methods in classifying potassium channel α-subunits. Previous large-scale analysis of potassium channel sequences have shown the interest and need in identifying these proteins, as well as the problem one has to overcome when it comes to analysing potassium channel sequences (Harte and Ouzounis, 2002; Moulton *et al*., 2003). The potassium channel family is highly diverse, on the one hand, and closely related to other ion channels, on the other hand. Using amino acids to classify potassium channels has shown to be to imprecise.

Harte and Ouzounis (2002) use a combination of hidden Markov models and BLASTp (Altschul *et al*., 1997). Moulton *et al*. (2003) also employ a BLAST algorithm. In addition, they use potassium channel motifs from the PRINTS Database (Attwood *et al*., 1997). Both approaches use multiple methods because a single of the above methods is only able to recognize sequences of a certain subset of the potassium channel family. This again shows the preeminence of the PSM, which is able to detect properties which are representative for all subsets of the potassium channel family.

PSM analyzes the physicochemical properties of amino acids in order to enable a more sensitive extraction of information coded in the amino acid sequences. As the results of the validation and the comparison with conventional pattern recognition methods indicate, PSM is superior to conventional methods for the search for sequences with a very low conservation level. The main advantage of PSM is that the signature describes, for each amino acid position, which of the selected properties are frequent and which of the properties are uncommon in the potassium channel α-subunits. Therefore, a query searches for sequences matching a property profile rather than for sequences with a similar amino acid sequence. This abstraction has shown to be much more sensitive and specific than known methods using only amino acids.

Using position-bound properties in the signature has another advantage: The interpretation of the results is very simple. Next to the number of missing properties and the unusual properties, the method returns, for each sequence, a vector that displays which sequence positions contain the missing and the untypical residues, respectively. This facilitates fast analysis of the sequence, e.g. despite a low number of missing properties, sequences with such properties within the selectivity filter can be left aside.

ACKNOWLEDGEMENTS

This work was supported by EC grant QLK3-CT2001-00401.

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3398–3402.

Ashcroft,F.M. (2000) *Ion Channels and Diseases,* 2nd edn. Academic Press, New York.

Attwood,T. *et al*. (1997) The PRINTS protein fingerprint database: functional and evolutionary applications. *Nucleic Acids Res*., **25**, 3398–3402.

Bairoch,A. and Apweiler,R. (2000) The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*., **28**, 45–48.

Chay,T.R. *et al*. (1990) The effect of ATP-sensitive K^+ channels on the electrical burst activity and insulin secretion in pancreatic beta-cells. *Cell Biophys*., **17**, 11–36.

Choe,S. (2003) Potassium channels. *FEBS Lett*., **555**, 62–65.

Doyle,D.A. *et al*. (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science*, **280**, 69–77.

Hamosh,A. *et al*. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat*., **15**, 57–61.

Harte,R. and Ouzounis,C. (2002) Genome-wide detection and family clustering of ion channels. *FEBS Lett*., **514**, 129–134.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning*. 1st ed. Springer, New York.

Hodgkin,J. *et al*. (1998) *C.elegans*: Sequence to Biology. *Science*, **282**, 2011.

Huang,J.Y. and Brutlag,D.L. (2001) The EMOTIF database. *Nucleic Acids Res*., **29**, 202–204.

Jiang,Y. *et al*. (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature*, **417**, 515–522.

Junker,J. *et al*. (2002) Amiodarone and acetazolamide for the treatment of genetically confirmed severe Andersen syndrome. *Neurology*, **59**, 466.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **157**, 105–132.

MacKinnon,R. (2002) Potassium channel structures. *Nat. Rev. Neurosci*., **3**, 115–121.

Miller,C. (2000) An overview of the potassium channel family. *Genome Biol*., **1**, Reviews 0004.

Moulton,G. *et al*. (2003) Phylogenomic analysis and evolution of the potassium channel gene family. *Recept. Channels*, **9**, 363–377.

Mulder,N.J. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **18**, 315–318.

Schneider,T.D. and Stephens,R.M. (1990) Sequence Logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.

Stryer,L. (1995) *Biochemistry*. 4th ed. W.H. Freeman and Company, New York.

Wu,T.D. and Brutlag,D.L. (1996) Discovering empirically conserved amino acid substitution groups in databases of protein families. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **4**, 230–240.

Yi,B.A. *et al*. (2001) Controlling potassium channel activities: interplay between the membrane and intracellular factors. *Proc. Natl. Acad. Sci. USA*, **98**, 11016–11023.