Physiol Genomics 28: 223–231, 2007.
First published October 10, 2006; doi:10.1152/physiolgenomics.00173.2006.

Genome-wide analysis of SPAK/OSR1 binding motifs

Eric Delpire and Kenneth B. E. Gagnon

Department of Anesthesiology, Vanderbilt University Medical Center, Nashville, Tennessee

Submitted 5 August 2006; accepted in final form 4 October 2006

Delpire E, Gagnon KB. Genome-wide analysis of SPAK/OSR1 binding motifs. *Physiol Genomics* 28: 223–231, 2007. First published October 10, 2006; doi:10.1152/physiolgenomics.00173.2006.—Based on the alignment of 12 sequences of protein motifs that interact with the kinases SPAK (Ste20-related proline alanine-rich kinase) and OSR1 (oxidative stress response 1), we performed genome-wide searches of the sequence **[S/G/V]RFx[V/I]xx[V/I/T/S]xx**, where **x** represents any amino acid. The "Mus musculus" search resulted in the identification of 131 mouse proteins containing 137 SPAK/OSR1 putative binding motifs. Similar numbers were found for human, zebrafish, fruit fly, and worm. A little more than half of the mouse proteins containing SPAK/OSR1 binding domains (53%) were also identified in the human search, whereas ~17–18% of these common hits were identified in the zebrafish search. The mouse proteins could be divided into two broad categories: 2/3 had an identified function, whereas 1/3 were either predicted or of unknown function. The known proteins were grouped as transport proteins, other membrane proteins, kinases, phosphatases, cytoskeletal, ribosomal, nuclear, enzymes, and others. Analysis of the location of the SPAK/OSR1 binding motif within the protein sequence revealed distribution throughout the entire length, but with preference to the extreme amino- or carboxyl termini for a large number of proteins. Analysis of the amino acid composition of the motifs revealed a preponderance of serine residues at *positions 5, 6, 7,* and *8*. In summary, our new search found and thus confirms the 12 proteins previously shown to interact with the kinases and identifies 119 potential new targets for SPAK and OSR1 in the mouse proteome.

mouse genome; protein-protein interaction; docking site; Ste20 kinases; NCBI search; Ste20-related proline alanine-rich kinase; oxidative stress response-1

USING A LARGE-SCALE YEAST-2 hybrid screen and the amino terminus of KCC3 (K-Cl cotransporter-3), we uncovered a novel protein-protein interaction motif, recognized by two closely related Ste20 kinases: Ste20-related proline alanine-rich kinase (SPAK/PASK) and oxidative stress response-1 (OSR1) (5). Using deletion mutants and single residue mutagenesis, we determined that at minimum, nine residues in the target protein are required for kinase binding. The SPAK binding motif, published in 2002, has the sequence **[R/K]Fx[V/I]xxxxx**, where **x** represents any amino acid. Because the motif is based on yeast-2 hybrid data and this methodology cannot assess binding to residues located upstream of the positively charged arginine residue (that we will define as *position 1*), we cannot preclude the possibility that more than nine residues are involved in the interaction. Indeed, the smallest peptide (9 amino acids) that gives a positive interaction in a yeast-2 hybrid assay is fused to the binding domain of GAL4, and thus, amino acids linking this domain to the peptide might in fact participate to the binding. Methodologies different from those using fusion proteins (different than yeast-2 hybrid and glutathione-*S*-transferase pull-down) are needed to resolve the minimum length required for SPAK/OSR1 binding. Subsequent studies from our laboratory and others have demonstrated binding of SPAK to several additional proteins (3–6). Sequence alignment of the motifs found in these proteins is presented in Fig. 1.

Several conclusions can be made from this alignment: First, serine or glycine residues are found preferentially at position −1, with the only exception being the original KCC3 target, which contains a valine residue at that position. Second, despite our previous demonstration of a conserved protein-protein interaction between SPAK and NKCC2 when we substituted a lysine for the arginine residue (2), of the 12 proteins listed in Fig. 1, only arginine residues are found at *position 1*. Third, amino acid substitution at *position 2* implies the necessity for a phenylalanine residue (2), a fact supported by the alignment represented in Fig. 1. Fourth, only valine or isoleucine residues are found at *position 4*. Finally, valine, isoleucine, threonine, and serine residues are found at *position 7*. Thus, based on this analysis, we propose the following refined motif: **[S/G/V]RFx[V/I/]xx[V/I/T/S]xx**. Since proteins are made of 20 amino acids, at first approximation, the probability of finding one specific residue at any position is 1 divided by 20 or 0.05. Thus, the probability of finding the **[S/G/V]RFx[V/I/]xx[V/I/T/S]xx** motif can be calculated to be 7.5×10^{-6}. Because some amino acids are more abundant than others, the actual probability of finding the motif increases slightly to 8.46×10^{-6}, indicating that if the SPAK binding motif was distributed within the genome by chance only, this motif should be found once every ~120,000 (118,151) residues.

To determine the number of SPAK/OSR1 binding motifs existing within a genome and identify the proteins containing this motif, we analyzed the National Center for Biotechnology Information (NCBI) protein database. Because the search had to accommodate multiple residues at any specific position, we created a small program written in Visual Basic that allowed us to identify the binding motifs within entire genomes. We provide here a complete list of mouse proteins containing **[S/G/V]RFx[V/I/]xx[V/I/T/S]xx** motifs and provide information regarding conservation between different vertebrate genomes.

METHODS

The NCBI protein database(http://www.ncbi.nlm.nih.gov/) was searched for *Mus musculus,* and hits were saved in FASTA format within a single file. After opening the file and removing all "Hard Returns" using the find and replace function of WordPerfect, we saved the file in TEXT format. Next, using a small routine written in Visual Basic (Microsoft), we searched the entire text file for specific sequences allowing multiple residues per position. Basically, the first 60 characters are placed into 60 consecutive string variables, and the routine analyzes the file one variable at a time. Based on the identity

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: E. Delpire, Dept. of Anesthesiology, Vanderbilt Univ. Medical Ctr., T-4202 Medical Center No., 1161 21st Ave. S., Nashville, TN 37232 (e-mail: eric.delpire@vanderbilt.edu).

	-1	1	2	3	4	5	6	7	8	9
KCC3a:	V	R	F	M	V	T	P	T	K	I
NKCC1:	S	R	F	Q	V	D	P	V	S	E
NKCC1:	G	R	F	R	V	N	F	V	D	P
NKCC2:	S	R	F	Q	V	H	V	I	N	E
WNK4 :	G	R	F	Q	V	T	S	S	K	E
WNK2 :	G	R	F	S	V	V	S	T	Q	D
AATYK:	S	R	F	T	V	S	P	T	P	A
AATYK:	S	R	F	S	I	T	H	I	S	D
HSP105:	G	R	F	V	V	Q	N	V	S	A
gelsolin:	G	R	F	V	I	E	E	V	P	G
CLH3 :	S	R	F	L	I	V	P	V	A	K
RELT :	G	R	F	R	V	A	R	I	P	E

Fig. 1. Alignment of the mouse KCC3 Ste20-related proline alanine-rich kinase/oxidative stress response-1 (SPAK/OSR1) binding motif with sequences of 11 other proteins experimentally shown to interact with the 2 kinases. Solid boxes delineate residues identified as essential for protein-protein interaction: R at *position 1*, F at *position 2*, and V/I at *position 4*. Dashed boxes identified residues exhibiting conservation at specific positions: V/S/G at *position −1*, and V/I/T/S at *position 7*.

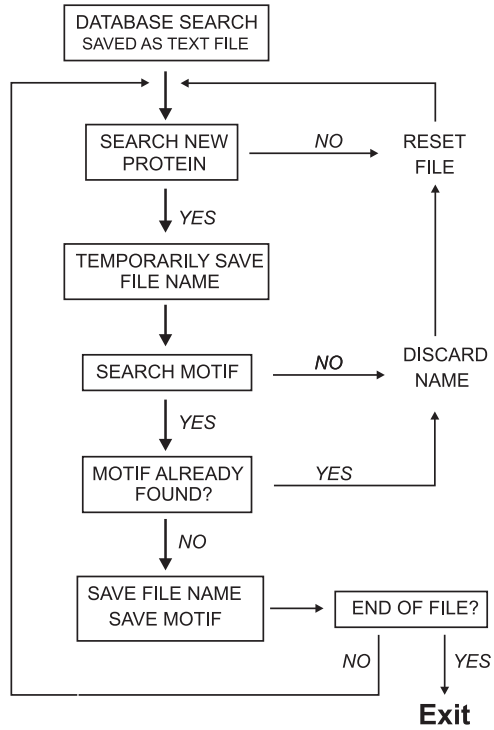

Fig. 2. Logic tree describing the general process used by the Visual Basic routine to identify SPAK/OSR1 binding motifs. See METHODS for details.

of the first variable, a specific action is taken (Fig. 2), and the routine moves one character over, resetting all 60 variables until the end of the file is reached. All NCBI FASTA entries begin with a line of text describing the sequence (i.e., >gi 51592076 ref NP_032089.2 solute-carrierfamily37). Our program identifies new protein entries when the first variable = ">", the second = "g", and the third = "i". The routine then increments its protein counter by one and seeks the fourth "|" character, which signals the beginning of the name or description of the protein. The 60 characters following the fourth "|" character are saved temporarily in a string variable. Next, the routine seeks 30 consecutive capitalized letters, which signals the beginning of the sequence (this method was chosen since all NCBI entries do not start with a methionine), and starts incrementing its residue counter. When the routine identifies a serine, valine, or glycine at the first position (first variable), followed by arginine and phenylalanine residues at *positions 2* and *3*, respectively, followed by a valine or isoleucine at *position 5*, and finally followed by valine, isoleucine, serine, or threonine residues at *position 8*, the routine captures the eight characters plus an additional seven characters following the motif into a temporary variable. This variable is then compared with 15 letter motifs previously saved, and if the string of 15 characters is unique, the motif belongs to a new protein and is saved along with the variable

Table 1. *Summary of the SPAK/OSR1 binding motif search in the NCBI protein database*

Search NCBI	Hits	Proteins with Motifs	Motifs Proteins	Duplicate	Residues	Time
All mouse	189,148	170	175	781	75,235,204	4h
Receptor	21,389	32	33	67	8,474,634	28′
Kinase	8,228	25	30	36	4,819,562	21′
Phosphatase	2,412	2	2	14	1,238,240	4′
Enzyme	3,018	11	13	12	1,509,775	5′
Cytoskeleton	1,389	9	9	0	1,079,907	3′
Transport	3,331	15	18	22	1,970,996	6′
Channels	2,601	9	9	41	1,551,394	5′
K^+ channels	551	1	1	3	279,708	57″
Na^+ channels	272	3	3	18	212,985	37″
Cl^- channels	184	0	0	0	113,045	20″
Ca^{2+} channels	307	0	0	0	267,320	45″
Pumps	147	1	1	0	82,138	16″
Homo sapiens	330,694	215	219	845	115,230,700	6 h 46′
Danio rerio	51,687	136	138	156	23,297,462	1 h 19′
Drosophila	81,930	103	104	275	39,321,787	1 h 56′
C. elegans	55,750	167	168	145	27,179,094	1 h 22′

The *top* portion of *column 1* lists the results of several Ste20-related proline alanine-rich kinase/oxidative stress response-1 (SPAK/OSR1) motif searches using different criteria attached to "*Mus musculus*." The *bottom* portion of *column 1* lists searches made for different species. The 2nd column lists the number of NCBI hits attached to each search. The 3rd column reports the number of proteins identified by the VisualBasic routine containing SPAK/OSR1 binding motifs. The 4th column reports the total number of motifs, as some proteins contain more than one motif. The 5th column lists the number of proteins that were eliminated because they were repeat hits. The 6th column reports the total number of amino acid residues analyzed by the routine, and the last column reports the times taken by the VisualBasic routine to return the data.

Table 2. *Mouse proteins with SPAK binding domains*

Protein	Motif	Accession	Position, %	Conserv.
	Transport Proteins			
Fatty acid (Slc27a4)	S**RF**HIPQVAE	NP_036119	58	H
Na-K-2C1 cotransporter 1. NKCC1 (Slc12a2)	S**RF**QVDPVSE	NP_033220	6.3	H
	G**RF**RVNFVD		10.1	
Na-K-2C1 cotransporter 2, NKCC2 (Slc12a1)	S**RF**QVHVINE	P55014	1.4	H
K-C1 cotransporter 3, KCC3 (Slc12a6)	V**RF**MVTPTKI	NP_598410	1.1	H
Organic cation transporter	G**RF**QVRLTIL	AAH21449	2.7	
ATP-binding cassette subfamily C member 9 isoform b	V**RF**AVKAIIS	NP_066378	38.1	H
Nicotinamide nucleotide transhydrogenase	V**RF**GIHPVAG	NP_032736	88.2	H
Gap junction membrane channel protein alpha-1	V**RF**WVLQIIF	NP_034418	19.6	
Gap junction membrane channel protein alpha-7	V**RF**WVFQIIL	NP_032148	18.7	HZ
Gap junction membrane channel protein alpha-9	S**RF**YIIQVVF	NP_034420	61.1	HZ
Gap junction alpha-12 protein (connexin-47)	V**RF**WVFQIVV	Q8BQU6	17.5	HZ
	Other Membrane Proteins			
SH3/ankyrin domain gene 3	V**RF**V**V**RSVSA	NP_067398	93.8	H
Transmembrane protein 24	V**RF**P**I**SVTQQ	NP_082185	22.8	H
Polyductin (PKHD1 gene)	V**RF**YVEGSNI	NP_694819	32.2	
Taste receptor type 2 member 136	V**RFCI**CVVWI	NP_851793	30.1	
Olfactory receptor MOR218-3	G**RFKI**FSTCI	AAL61334	74.6	
Down syndrome cell adhesion molecule	S**RFLI**TSTGA	NP_112451	8.5	H
Interleukin 12 receptor, beta 2	S**RFIV**RVTAI	NP_032380	24	
Tumor necrosis factor receptor superfamily, member 19-like	G**RFRV**ARIPE	AAI00311	81.2	HZ
	Kinases			
Aortic preferentially expressed gene 1 (S/T kinase)	V**RFRV**ACSNR	NP_031489	72.9	
alternatively spliced exon	G**RFGV**VRSC	BAC65770	78.9	
Serine/threonine kinase SADB	V**RFQV**DISSS	NP_001003920	61.3	H
Brain apoptosis-associated tyrosine kinase, AATYK	S**RFTV**SPTPA	NP_031403	97	HZ
	S**RFSI**THIS		98	
WNK1	G**RFQVSV**TMD	NP_941992	78	H
	G**RFQV**TTTA		82	
	G**RF**SVSRT		82	
WNK2	G**RF**SVVSTQD	AAH46464	73.7	H
WNK3	G**RFQVITV**PQ	XP_986322	74.7	HZ
WNK4	G**RFQV**TSSKE	NP_783569	81.4	H
	Phosphatases			
Protein phosphatase 2 regulatory subunit B (59 kDa)	V**RFTV**NPSDP	XP_923875	2.6	
Protein phosphatase 2 regulatory subunit A	V**RFNV**AKSLQ	NP_058587	91	HZ
	Cytoskeleton			
Gelsolin	G**RFV**IEE**V**PG	NP_666232	86.5	HZ
Titin isoform N2-B	V**RFGV**TITVH	NP_082280	93.6	HZ
	G**RFHI**ENTD		99.8	
Dynein axonemal heavy chain 11	V**RFSI**NDSTT	NP_034190	31.2	
Nexilin (actin filament-binding protein)	V**RFTV**KVTGE	NP_955759	87.3	HZ
Scin protein	G**RFII**EEVPG	AAH63328	87.3	
Advillin	G**RFLV**TEVTD	NP_033765	76.4	
	Ribosomal			
Ribosomal protein S11	V**RFNV**LKVTK	NP_038753	87.3	HZ
Ribosomal protein S23	V**RFKV**VKVAN	AAI00604	82.5	HZ
Mrps25 protein (mitochondrial)	G**RF**PIRRTLQ	AAH22953	2.9	HZ
	Nuclear			
Structure-specific recognition protein	V**RFYV**PPTQE	NP_892035	22.9	HZ
RecQ protein-like (helicase)	V**RFVI**HHSMS	NP_075529	59.1	HZ
Bloom syndrome protein homolog (helicase)	V**RFVI**HASLP	NP_031576	68.2	H
DNA-dependent ATPase SNF2L	V**RFEV**SPSYV	AAK52453	16.4	
Cardiac-specific RNA helicase	S**RFRI**IITTC	AAK77049	40	H
Nesprin (synaptic nuclear envelope 1 isoform 2)	S**RFQI**QQTAN	NP_071310	51.6	H
RanBP2 protein	S**RFKV**SRIKG	CAA60778	0.3	
Bromodomain containing 1	S**RFGI**STSLE	NP_001028446	98.8	
Minichromosome maintenance-deficient 2 mitotin	S**RFDV**LCVVR	NP_032590	72.5	H
Aquarius (helicase)	S**RFAI**RKIML	NP_033832	3.8	H
TAR DNA binding protein isoform 3	G**RFGV**HLISN	NP_001003899	92.9	H

Continued

Table 2.—*Continued*

Protein	Motif	Accession	Position, %	Conserv.
	Nuclear (Continued)			
Myst4 protein	G**RF**L**I**DFSYL	AAH95974	80	HZ
Zmym4 protein	S**RF**M**I**ELTKL	AAH50924	82.2	H
Pseudouridine synthase (Rpusd3)	S**RF**HVMATGR	AA**H**3808	60.7	
	Enzymes			
Glycosylphosphatidylinositol specific phospholipase D1	G**RF**Q**I**GRVYI	NP_032182	48.5	
Similar to O-acyltransferase (membrane bound)	V**RF**Y**I**TLSSL	XP_134120	36.7	
Acetyl-coenzyme A carboxylase alpha	S**RF**I**I**GSVSE	NP_579938	1	H
Endothelin converting enzyme 1	S**RF**RVIGSLS	NP_955011	95.3	HZ
Phosphodiestarase 1b (Pde1b)	S**RF**K**I**PTVFL	AAH58531	42.4	H
Phosphodiesterase 1c isoform a	S**RF**K**I**PISAL	NP_035184	30.6	H
Neurolysin (metallopeptidase M3 family)	S**RF**DIEMSMR	NP_083723	18.5	H
	Extracellular Space			
Cathelicidin (antimicrobial peptide)	V**RF**R**V**KETVC	NP_034051	42.8	
Roundabout homolog 1	S**RF**S**V**SQTGD	NP_062286	22.6	HZ
Pentraxin (Crp, membrane-bound calcium-binding protein)	V**RF**M**V**SEIPE	NP_031794	44.9	
Fibroblast growth factor 4	S**RF**F**V**AMSSR	NP_034332	63.9	H
	Other Proteins			
Dennd2c protein (suppression of tumorigenicity)	V**RF**FVELVGH	AAH46440	79.1	H
Dennd2a gene	V**RF**F**V**EIVGH	NP_766065	89.7	
Selenoprotein X 1	S**RF**U**I**FSSSL	NP_038787	79.3	HZ
Heat shock protein 105	G**RF**V**V**QNVSA	NP_038587	53.6	HZ
Eef2 protein (elongation factor)	V**RF**D**V**HDVTL	AAH60707	81.3	H
High glucose-regulated protein 8	G**RF**D**V**RWIFV	NP_663368	85.1	HZ
Procollagen type XVIII alpha 1	G**RF**G**I**NGSYA	NP_034059	36.7	
IgH antibody heavy chain VDJ region	G**RF**I**V**SRSRD	AAB95229	54.8	
Follistatin-like 4	G**RF**I**V**SVSNK	NP_796033	82.9	
Flywch1 protein	G**RF**L**V**YESFL	AAH25645	60.8	H
FERM domain containing 4A	V**RF**Y**I**ESISY	NP_766063	8.1	HZ
Golgi-associated band 4.1-like protein	V**RF**Y**I**ESISF	NP_660130	8.6	H
Carcinoembryonic antigen	V**RF**H**V**HPILL	CAA33409	29.7	
Low-density lipoprotein receptor-related protein 6	S**RF**V**I**INTEI	NP_032540	10.7	
Proline-rich 14 protein	S**RF**R**I**RRTPV	BC006909	58.5	H
Poly(ADP-ribose) polymerase family, member 14	S**RF**P**V**DVVVN	NP_001034619	45.3	
Ccdc60 protein	S**RF**L**I**QCVKI	AAH75662	27.5	H
HEF-like protein	S**RF**L**I**PRVEQ	NP_001028710	47.4	
Sfrs14 protein	S**RF**G**I**EIIKW	AAH57305	44.9	
Ig heavy chain var. region MANEX1B against dystrophin	G**RF**T**I**SRVRN	S56008	60.9	H
Ig heavy chain variable region scfv fragment	G**RF**T**I**SKTSS	AAL15163	53.9	
IgH antibody heavy chain VDJ region	G**RF**F**V**SRSRD	AAB94782	56.7	
Polymeric immunoglobulin receptor	G**RF**S**V**LITGL	NP_035212	39.6	
Pdzx protein	G**RF**S**V**ELTRG	AAH99459	38.6	
Gtf3c1 protein	V**RF**R**I**SNSST	AAH32208	5.7	H
Glutaredoxin cysteine-rich 1 protein	V**RF**R**I**ASSHS	NP_001018019	5.5	H
D19Ertd652e protein	V**RF**Q**V**NLTVQ	AA107403	61.3	
Vacuolar protein sorting 37D	G**RF**G**I**LSTGQ	NP_808242	9.7	
Msid2	G**RF**K**I**DVSDT	AAS87379	59.8	
Left right determination factor 1	G**RF**L**V**SETST	NP_034224	23.8	
	Unknown Proteins			
Unnamed protein product	V**RF**L**I**CFIGM	BAC33608	94.5	
Unnamed protein product	V**RF**L**V**SWSWL	BAE26787	95.3	
Unnamed protein product	V**RF**C**V**FPTGG	BAE24015	56.3	
Unnamed protein product	S**RF**T**V**AAIRR	BAB25824	74.2	
Unnamed protein product	S**RF**S**II**SSSS	BAE38028	3.9	H
Unnamed protein product	S**RF**R**V**KLTLE	BAC27063	22.9	H
Unnamed protein product	S**RF**L**V**ILSFF	BAE33644	32.4	
Unnamed protein product	G**RF**S**V**CFSFP	BAE24254	90	
Unnamed protein product	G**RF**F**I**PSVTQ	BAE23369	41.9	
Unnamed protein product	G**RF**K**V**FRISP	BAE42406	78.9	
Unknown protein	V**RF**S**V**PGTTW	BAE24216	36.1	
Unknown protein	V**RF**R**I**AKIVA	BAE22912	38.1	
RIKEN cDNA 1110057K04	S**RF**P**V**WIISH	AAH46986	22.4	
RIKEN cDNA 4631403P03	G**RF**RVTHIEK	AK044097	55.1	H

Continued

Table 2.—*Continued*

Protein	Motif	Accession	Position, %	Conserv.
	Unknown Proteins (Continued)			
C730048C13Rik protein	G**RF**Q**I**LQISF	AAH50265	3.1	
mKIAA0539 protein	G**RF**T**V**NKVAL	BAD90144	72.4	
Amphoterin-induced gene	S**R**FPVELIKD	AAH10598	34.8	
	Predicted Proteins			
Similar to 40S ribosomal protein S4, X isoform	G**RF**AVHHITL	XP_001004419	40.1	H
Similar to TBC1 domain family member 4 (Akt substrate)	G**RFEI**NLISP	XP_619279	45.5	
Sodium channel voltage-gated type IX alpha 8	SRFSVSQVAN	XP_904778	68.1	
Importin 9 isoform 5 (1, 6)	S**RF**TVAMSPE	XP_903902	72.4	H
Similar to hyaluronoglucosaminidase 4	S**RFR**VRESLR	XP_995969	57.8	
Similar to proline-rich 6 isoform 4 (or 6)	V**RFEV**WASAD	XP_921840	53.2	H
Similar to Bex4 protein	G**RFVV**QGTEV	XP_982566	67	
Similar to T-cell receptor alpha chain V region HPB-MLT precursor isoform 2	G**RFSV**KHSKA	XP_925983	51.3	H
Hypothetical protein LOC332309	V**RFKI**SSSYS	NP_001028598	7.5	H
Hypothetical protein LOC382137	V**RF**GVDCTQL	NP_941077	9.8	H
Hypothetical protein LOC107373 (hydrolase, peptidase)	V**RFCI**HAVGS	NP_080916	29.2	
Hypothetical protein LOC231279	S**RFPV**EDIRN	NP_766299	6.6	H
Hypothetical protein LOC71874	S**RFIV**DYSRD	NP_765999	13.3	
Hypothetical protein LOC212943 isoform 1	S**RFFI**DFSDI	XP_135029	71.5	H
Hypothetical protein LOC73453	G**RFSI**SPSHD	NP_898920	42.3	
Hypothetical protein LOC100532	G**RF**RVTKVEH	NP_666035	77.7	H
Hypothetical protein LOC75619	G**RFSI**LNSQH	NP_766010	47	
Hypothetical protein LOC226089 isoform 3	G**RFLI**IFIAI	XP_905557	92.8	
Hypothetical protein LOC75905	G**RFLV**STSTR	NP_780437	56	
Hypothetical protein LOC210463	G**RFQI**CMIIL	NP_759010	2.5	
Hypothetical protein LOC271981 (kinase)	G**RFQI**LKTIT	NP_766620	5.2	H
Similar to Kelch repeat and BTB domain-containing protein	SRFQIPSVFT	XP_915191	18.3	H

Column 1 lists all mouse proteins that were identified as including **[S/G/V]RFx[V/I]xx[V/I/T/S]xx** binding sites. The exact sequence of the motif(s) is reported in the 2nd column with residues at *positions 1, 2,* and *4* highlighted (bold). The 3rd column contains the accession number of the protein, the 4th column indicates the location of the motif within the protein (as a percentage). In the last column, we indicate whether the motif is also found in the human (H) or zebrafish (Z) orthologs.

containing the name of the protein, and the motif counter is incremented by one. In contrast, if the string of 15 character is already saved, this indicates a redundant protein and the "duplicate protein" counter is incremented by one. The likelihood that two different proteins contain the exact same 15 characters is infinitesimal, as the probability to find a specific string of 15 residues can be estimated at 3×10^{-20}. The routine then continues its search without saving the motif. At the end of the search, all variables (protein counter, motif counter, protein names, and motifs) are copied into a new single text file.

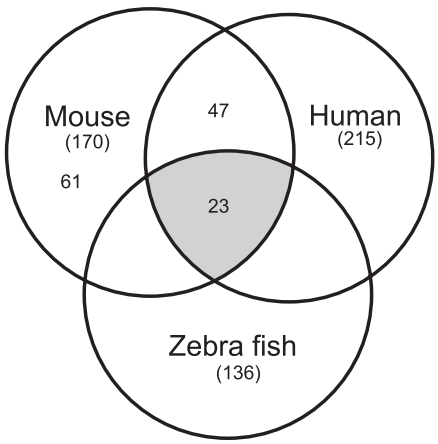

Fig. 3. Conservation of SPAK/OSR1 binding motifs. The value in parentheses indicates the number of protein hits in 3 vertebrate genomes: mouse (*Mus musculus*), human (*Homo sapiens*), and zebrafish (*Danio rerio*). We identified a total of 131 mouse proteins from the 170 hits. There were 23 out of 131 proteins conserved in mouse, human, and zebrafish genomes. Sixty-one proteins were specific to the mouse genome, and 70 (47 + 23) proteins were conserved between the mouse and human genomes. The list of these proteins is reported in Table 2.

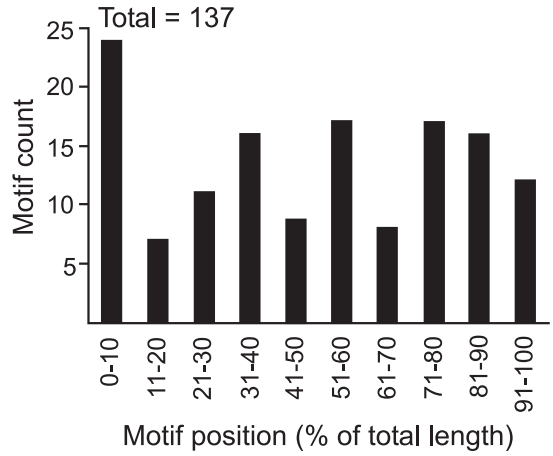

Fig. 4. Distribution of the SPAK/OSR1 motifs within the length of proteins. For each motif identified, its location (*position −1*) was divided by the total length of the protein. There were 137 motifs identified in the mouse search. Note that the motif is distributed throughout the length of the protein with a larger number of proteins with SPAK/OSR1 binding motif located within the 1st 10% of the protein length. The motif location for each individual protein is reported in Table 2.

Table 3. *Mouse proteins with putative [S/G/V]KFx[V/I]xx[V/I/T/S]xx motifs*

Protein	Motif	Accession	Position (%)
Transport Proteins			
Aquaporin 6	G**KF**AVHWIFW	NP_780296	70
Plasma membrane calcium ATPase 1	V**KFFI**HGVTV	NP_080758	34.2
Potassium inwardly rectifying channel J15 isoform a	V**KFH**VDSSSE	NP_001034146	65.2
Equilibrative nucleoside transporter 3	S**KF**FVPLTVF	Q99P65	72.7
ATP-binding cassette, subfamily A (ABC1), member 17	S**KFGI**GYSLQ	NP_001026792	91.6
Sperm-associated cation channel 2	S**KFII**SLIFL	NP_694715	18.5
Solute carrier family 12, member 2	SKFRIDFSDI	NP_033220	88.8
Other Membrane Proteins			
Purinergic receptor P2X1	G**KFDII**PTMT	NP_032797	81.2
Purinergic receptor P2X3	G**KFNII**PTII	NP_663501	79.1
Purinergic receptor P2X4	G**KFDII**PTMI	NP_035156	83.4
Purinergic receptor P2X5	G**KFSII**PTVI	NP_201578	72.3
Cadherin 23	G**KFEI**DESTG	NM_023370	37.4
Olfactory receptor 1496	V**KFVV**GLSVI	NP_667200	63.1
Olfactory receptor 898	V**KFAV**KRTMK	NP_667082	95.9
Olfactory receptor 116	S**KFTV**RIICP	NP_666843	96
Olfactory receptor 452	G**KFSV**FTSKL	NP_001011869	96.5
Huntingtin	V**KFVV**MTVEA	NP_034544	71.5
Protocadherin alpha 5	S**KFTI**DSSSG	NP_034089	31.1
Cadherin-related neural receptor 6	S**KFTI**DSSSG	BAA29051	28.4
Killer activatory receptor-like protein p91D	G**KF**FIPSVTQ	AAC40073	41.9
PTX1 protein isoform 1	G**KFIV**EIICC	NP_080444	89.1
Macrophage scavenger receptor 2	S**KFTI**SAISK	NP_109632	31.8
Anti-ricin antibody light chain variable region	V**KFLI**YSTSN	ABD47464	40
Kinases			
Protein kinase cGMP-dependent type II	S**KF**CVACVTE	AA113206	72.6
Brain-selective kinase 2 isoform (alpha, beta, gamma isoforms)	V**KF**QVDITYT	NP_083702	90
Bruton agammaglobulinemia tyrosine kinase	S**KF**PVRWSPP	NP_038510	84.5
Thymidylate kinase family LPS-inducible member	G**KF**QVIAIEG	NP_065582	50.4
Hexokinase domain containing 1	S**KF**RVLKVQV	NP_663394	9.6
Phosphatases			
Protein tyrosine phosphatase nonreceptor type 9	G**KF**TILNVRD	NP_062625	15.9
Nuclear			
Topoisomerase (DNA) II alpha	V**KFVI**KMTEE	NP_035753	63.2
Topoisomerase (DNA) II beta	V**KFVV**KMTEE	NP_033435	60.5
DNA methyltransferase 3A isoforms 1 & 2	G**KFSV**VCVEK	NP_031898 (1)	36.3
Polymerase (RNA) II (DNA directed) polypeptide C	V**KFII**ENTDL	NP_033116	6.9
Methyl-CpG binding domain protein 4	G**KFDV**YFISP	NP_034904	16.4
MYST histone acetyltransferase 1	G**KFLI**AFSYE	NP_080646	71.8
SWI/SNF-related matrix associated actin dependent Regulator of chromatin, subfamily a, member 4	G**KF**NV**L**LTTY	NP_035547	52.7
Chromodomain helicase DNA binding protein 3	V**KFH**VLLTSY	CAI35246	44
DEAH (Asp-Glu-Ala-His) box polypeptide 35	V**KFQV**PKVSS	NP_665685	86.9
Enzymes			
RAB5A-B-C, members RAS oncogene family	V**K**FEIWDTAG	NP_080163 (5A)	32.1
Dipeptidylpeptidase 4	V**KFFI**VNIDS	NP_034204	34.2
2′-5′ Oligoadenylate synthetase 1G	V**KFEV**QSSWW	NP_035982	32.4
2′-5′ Oligoadenylate synthetase 1C & 1F	V**KFEV**QSSEE	NP_291019 (IC)	32.9
Carnosine dipeptidase 1 (metallopeptidase M20 family)	G**KFSI**RLVPT	NP_803233	70.3
Carnosine dipeptidase 2 (metallopeptidase M20 family)	G**KFSI**RLVPD	NP_075638	71.2
Arhgef10 (guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases)	V**KF**IVSATAF	AAH66074	85.8
RAS p21 protein activator 2	S**KFTI**EDSVA	NP_444498	93.7
ADP-ribosylation factor 6	V**KFNV**WDVGG	NP_031507	32.6
Phosphoglucomutase 1	G**KFEI**SAIRD	NP_079976	83.7
Transglutaminase 3 E polypeptide	G**KFK**VTGILA	NP_033400	70
4-Hydroxyphenylpyruvic acid dioxygenase	V**KF**AVLQTYG	NP_032303	33.3
Peptidase 4	S**KF**NVNNTIL	NP_032846	33.9
Rab geranylgeranyl transferase a subunit	S**KFLV**ENSVL	NP_062392	75.3
UTP-glucose-1-phosphate uridylyltransferase	S**KFKIF**NTNN	AAH25585	62.6
UDP-glucuronosyltransferase 2 family, member 5	S**KF**DVLLSDP	BAE22997	28.7
UDP glucuronosyltransferase 2 family, polypeptide B36	S**KFDV**ILSDA	NP_001025038	27.2
Serine (or cysteine) peptidase inhibitor, clade A, member 3B	G**KF**IVDRSR**H**	NP_766612	54

Continued

Table 3.—*Continued*

Protein	Motif	Accession	Position (%)
Extracellular Space			
Tubulin tyrosine ligase-like family, member 12	V**KF**D**I**RYIVL	NP_898838	72.3
Pregnancy zone protein	V**KF**R**V**VSVDI	NP_031402	9.5
Other Proteins			
Pad-I-like isoform 1	V**KFTI**STTSK	NP_081569	68
YTH domain family 1	G**KF**D**V**KWIFV	NP_776122	84.4
Ythdf3 protein	G**KFEV**KWIFV	AAH67040	85.3
Sulfatase modifying factor 1	G**KFPV**SNTGE	NP_666049	60.9
Ly-6C variant	G**KFPV**TYIKK	BAA13049	61.8
Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin)	G**KFRI**GFIDE	NP_082247	58.2
Neuron specific gene family member 2	G**KFRV**PKIAE	NP_032767	35.1
Phf13 protein	S**KFDI**RRSNR	AAH59282	93.2
Nup214 protein	V**KFAV**QDVND	AAH39282	16.5
Sperm acrosome associated 1	**VKFTV**YTTNE	NP_080569	67.2
Translocon-associated protein alpha	V**KFLV**GFTNK	AAK82421	35.3
Tumor-associated antigen	V**KFLV**AKVYE	AAB17502	23.9
Hydin (hydrocephalus-inducing protein)	S**KFTI**SPSSI	AAO44953	61.7
Eukaryotic elongation factor, selenocysteine-tRNA-specific	G**KFKI**HITEQ	NP_075547	87.3
DEP domain containing 1B	S**KFII**HNVYS	NP_848798	40.2
Apolipoprotein B	S**KFKV**SHVEK	AAH38263	97
Immunoglobulin H-chain V-region	S**KFDV**WGTGT	BAD24599	87.5
Melanoma-derived leucine zipper, extra-nuclear factor	S**KFAI**KGIIN	NP_113555	46.5
Unknown Proteins			
Unnamed protein product	V**KFCI**DASQP	BAC28336	2.8
Predicted Proteins			
Hypothetical protein LOC67105	V**KFYI**EGSEP	NP_080245	78.3
Similar to Serum amyloid P-component precursor	G**KFDV**KQSFV	XP_136329	70.6
Similar to Extracellular calcium-sensing receptor	G**KFLV**GIIGA	XP_981998	17.3
Similar to Zinc finger protein 208	S**KFFV**YPSRL	XP_995677	66.8
Similar to FLJ44048 protein	V**KFTV**PVTLP	XP_918762	22.2
Similar to 60S ribosomal protein L38	V**KFRV**RCSRY	XP_983985	45.7
Similar to trinucleotide repeat containing 6A	S**KFVV**GSSSN	XP_987987	17.9
Similar to enolase 1, alpha	S**KFGV**NIILG	XP_357379	22.1
Similar to mucin 17	S**KFGI**ATSKN	XP_992907	71
Similar to Zinc finger protein 208	S**KFFV**PHSQL	XP_995677	66.8
Similar to nucleolar protein NOP56	V**KFNV**NRVDN	BAB27647	45.4
Similar to CG31901-PA	G**KFVI**STSDG	XP_913556	80.4
Similar to GTPase activating protein testicular GAPI	V**KFLI**ENSLK	XP_995294	44.7
Similar to putative binding protein 7a5	G**KFTV**SDIRK	XP_995079	47.7
Hypothetical protein LOC210711	G**KFTV**NLSGC	NP_666067	60.3
Hypothetical protein LOC241134	G**KFTV**PASHS	NP_766437	23.7
Hypothetical protein LOC329562	G**KFYV**AIVCS	NP_808519	76.2
Hypothetical protein LOC67484	V**KFEI**CVSSK	NP_080465	16.5
Hypothetical protein LOC66367	V**KFAI**KRTLV	NP_780316	90.1
Hypothetical protein LOC268706	S**KFNI**LGTVS	NP_848861	55.5
Hypothetical protein LOC626359	V**KFEV**MCVVL	NP_001033016	70.8
Hypothetical protein from cochlea	V**KFQV**VKIKG	CAA67982	92.2

Column 1 lists all mouse proteins that were identified substituting the arginine residue at *position 1* with a lysine residue. The sequence of the alternative SPAK/OSR1 motif is reported in the 2nd column. The 3rd column lists the accession number, and the 4th column indicates the location of the motif within the protein (as a percentage).

For each of the proteins returned by the Visual Basic routine, we queried the NCBI protein database to locate and confirm the motif within the amino acid sequence of the protein. Additionally, the curation process provided the opportunity to eliminate any remaining redundant or nonmurine proteins and allowed grouping of the proteins into categories based on their general description and features.

RESULTS AND DISCUSSION

The NCBI protein database(http://www.ncbi.nlm.nih.gov/) was first searched for *Mus musculus*. The search was performed on June 14, 2006, and returned some 189,153 hits. More directed searches such as *Mus musculus* AND "descriptor" returned: 21,389 hits for "receptors," 8,228 hits for "kinases," 3,018 hits for "enzymes," 1,389 hits for "cytoskeleton," 3,331 hits for "transport," 2,412 hits for "phosphatases," 2,601 hits for "channels." Further defining the type of ion channel reduced the number of hits to 551 for "K^+ channels," 307 for "Ca^{2+} channels," 272 hits for "Na^+ channels," and 184 hits for "Cl^- channels" (see Table 1). An obvious first observation that can be made from these searches is that NCBI returns many more hits than the number of proteins existing in any one specific genome. Indeed, the NCBI search of *Mus musculus*

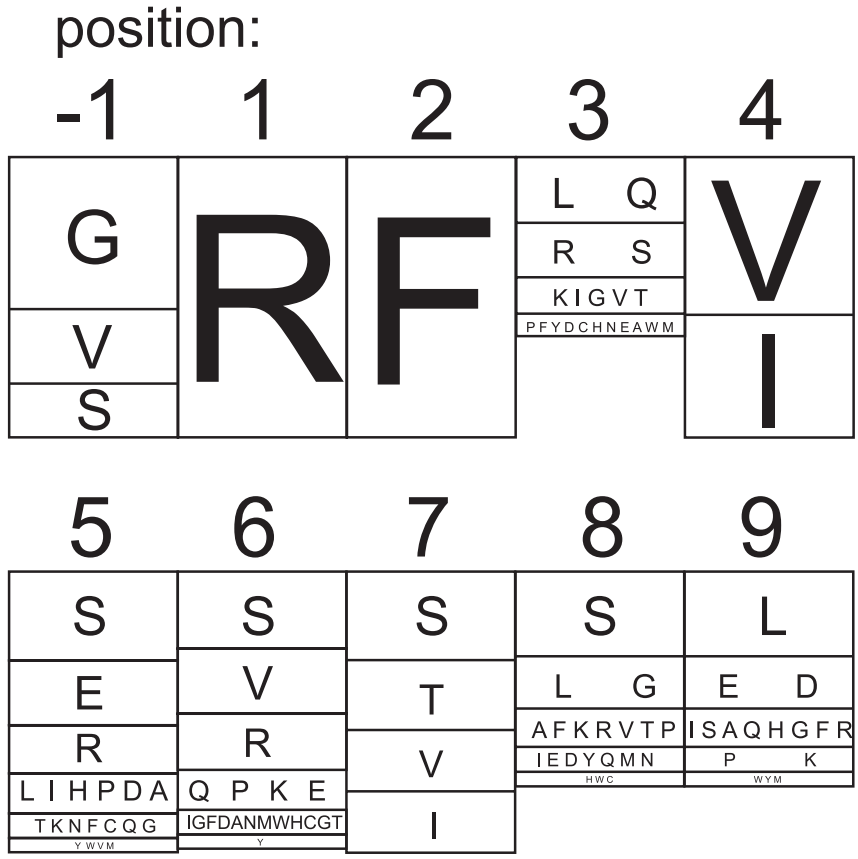

Fig. 5. Relative frequency of occurrence of amino acids at each position of the SPAK/OSR1 binding motif. Analysis was based on the 137 binding motifs listed in Table 2. Size of the letter is proportional to the occurrence of the amino acid. The arginine residue was arbitrarily placed at *position 1*.

returned close to 200,000 hits, whereas estimation of the number of proteins in the mouse genome is closer to 24,000–25,000 (7). A second observation is that, due to the extensive cross-referencing of NCBI entries, proteins are often identified using unrelated or indirect searches. As an example, only a fraction of the 8,228 entries found with *Mus musculus* AND "kinases" are actual kinases. The Na-K-2Cl cotransporter (NKCC1) is identified in the *Mus musculus* AND "kinases" search due to the multiple NCBI links existing between SPAK/PASK kinases and the cotransporter.

Using a routine written in Visual Basic, we searched for occurrence of the **[S/G/V]RFx[V/I]xx[V/I/T/S]xx** within the 189,153 downloaded mouse proteins. The routine identified 170 proteins containing 175 motifs. Analysis of the proteins showed that the routine was able to eliminate all redundant entries and alternatively spliced products, thus identifying one product per gene. However, a small number of proteins identified in the "*Mus musculus*" search were not mouse proteins but belonged to other species. After curation and elimination of these proteins the final list comprised 131 proteins containing 137 motifs. There were five instances where a single open reading frame encoded a protein containing more than one SPAK/OSR1 binding motifs. WNK1 contains three independent SPAK/OSR1 binding motifs, and four other proteins (NKCC1, AATYK, titin isoform N2B, and mKIAA1297) contain two SPAK/OSR1 motifs. The complete list is presented in Table 2. Two-thirds of these proteins are involved in transport, enzymatic, cytoskeletal, nuclear, ribosomal, and other functions, whereas the remaining one-third of the proteins listed in Table 2 either are only predicted or have undetermined function.

Comparison among genomes revealed a high number of proteins with conserved SPAK/OSR1 binding motifs. Indeed, 70 out of the 131 mouse proteins were also found in the human genome, indicating a 53% conservation. However, a significantly lower number of mouse and human proteins containing SPAK/OSR1 binding motifs (23 out of 131 mouse proteins) were present in the genetically more distant zebrafish genome, corresponding to 17.5% of mouse proteins. These data are summarized in Table 2 and Fig. 3.

Next, we examined the location of the SPAK/OSR1 binding motif within the entire set of 131 proteins. As seen in Fig. 4, a larger number of motifs are located at the extreme NH_2 terminus of the protein. The remaining motifs are distributed along the length of the target protein. However, proteins such as transporters and channels, which have a transmembrane core and cytosolic amino- and carboxyl termini, have the SPAK/OSR1 binding motifs preferentially located on the cytosolic tails (see Table 2). In other proteins such as kinases, which have well-defined catalytic and regulatory domains, the **[S/G/V]RFx[V/I]xx[V/I/T/S]xx** motifs are found exclusively in the regulatory domain. Indeed, the average location for sites found in kinases displaying SPAK/OSR1 binding motifs is 80 ± 3.2% ($n = 11$). With the exception of dynein, all SPAK/OSR1 binding motifs found in cytoskeletal proteins are also found at the extreme carboxyl-terminus (Table 2). Finally, note the opposite location of the SPAK/OSR1 binding motifs in the two regulatory subunits of protein phosphatase 2 (Table 2). These

motifs are again preferentially located at the very beginning or very end of the protein.

The identification of 137 **[S/G/V]RFx[V/I]xx[V/I/T/S]xx** motifs within 131 mouse proteins allow us to further examine the distribution of residues in the motif at *positions 3* and *5–9*. As shown in Fig. 5, residues at *position 3* can be very diverse, from a mostly hydrophobic leucine to a very hydrophilic arginine. One interesting feature of this analysis of the expanded motif, is the preferential use of serine residues at *positions 5, 6, 7,* and *8*.

The list of proteins that contain the SPAK/OSR1 binding motif is rather extensive despite the somewhat restrictive definition of the motif based on the 12 sequences reported in Fig. 1. However, if we were to allow additional residues at *positions −1, 1,* or *7,* the total number of proteins would increase significantly. For example, we expanded our search to include the possibility of a lysine residue at *position 1*. We identified an additional 100 proteins with **[S/G/V]KFx[V/I]xx[V/I/T/S]xx** motifs (Table 3). To date, however, no SPAK/OSR1 interacting proteins have been identified as containing a KFxV motif. As the motifs were identified through gene searching and not from peptide libraries, some of the identified motifs might not constitute actual SPAK/OSR1 binding sites. Therefore, whether proteins identified in this search do indeed interact with the kinases requires experimental testing. Anchoring of kinases to their substrates constitutes one important mechanism of substrate specificity (1). Thus, for any protein-protein interaction to reach biological significance, overlap of their temporal and spatial distribution is required. While SPAK/OSR1 interaction with cation-chloride cotransporters, chloride channels, and TNF receptor has been evidenced experimentally (3–6), this study provides a list of new proteins that potentially interact with SPAK and OSR1 and suggests a significantly greater physiological role for each of these kinases.

GRANTS

This work was supported by National Institutes of Health Research Grant NS-36758.

REFERENCES

1. **Bardwell AJ, Flatauer LJ, Matsukuma K, Thorner J, Bardwell L.** A conserved docking site in MEKs mediates high-affinity binding to MAP kinases and cooperates with a scaffold protein to enhance signal transmission. *J Biol Chem* 276: 10374–10386, 2001.
2. **Delpire E, Piechotta K.** Ste20 kinases and cation-chloride cotransporters. In: *Cell Volume and Signal Transduction*, edited by Lauf PK and Adragna NC (Advances in Experimental Medicine and Biology). New York: Springer Science and Business Media 559: 43–53, 2004.
3. **Denton J, Nehrke K, Yin X, Morrison R, Strange K.** GCK-3, a newly identified Ste20 kinase, binds to and regulates the activity of a cell cycle-dependent ClC anion channel. *J Gen Physiol* 125: 113–125, 2005.
4. **Piechotta K, Garbarini NJ, England R, Delpire E.** Characterization of the interaction of the stress kinase SPAK with the Na^+-K^+-$2Cl^-$ cotransporter in the nervous system: Evidence for a scaffolding role of the kinase. *J Biol Chem* 278: 52848–52856, 2003.
5. **Piechotta K, Lu J, Delpire E.** Cation-chloride cotransporters interact with the stress-related kinases SPAK and OSR1. *J Biol Chem* 277: 50812–50819, 2002.
6. **Polek TC, Talpaz M, Spivak-Kroizman T.** The TNF receptor, RELT, binds SPAK and uses it to mediate p38 and JNK activation. *Biochem Biophys Res Commun* 343: 125–134, 2006.
7. **Waterston RH, Mouse Genome Sequencing Consortium.** Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562, 2002.