SUPPLEMENTAL DATA

EXPERIMENTAL PROCEDURES

Determination of the 5'- and 3' ends of the SiMat1 mRNA by RACE-PCR. Isolation of total mRNA using oligo-dT-functionalized magnetic beads, and synthesis of cDNA was performed as described previously (1). For amplification of the 3'-end, nested PCRs were performed using the gene-specific sense primer 5'-ATTCTCCATTGCCACCAC and the antisense primer 5'-GGCCACGCGTCGACTAGTAC$(T)_{17}$ for the first PCR, and the gene-specific sense primer 5'- ACCGGTGGTTTCTCCACC and the antisense primer 5'-GGCCACGCGTCGACTAGTAC for the second PCR. To amplify the 5'-end, two nested PCRs were performed using sense primer 5'-GGCCACGCGTCGACTAGTACGGGIIGGGIIGGGIIG and the gene-specific antisense primer 5'- ATCATCTTACGGAGTCCACC and for the first PCR, and sense primer 5'-GGCCACGCGTCGACTAGTAC and gene specific antisense primer 5'-ACCAACGGTGCTGGAAGAGG for the second PCR. All PCR products were ligated into the pJet1.2 vector (Thermo Scientific) and sequenced (sequence information can be found in Fig. S6).

FIGURES

rCinY2

```
  1 AGGAGGTAAA ACAT**ATG**GGT ACTAACAAAA CGCTGCCACC GACCCCGTTC CCAGGCCGTC
 61 CGACCCCGAA TCCGACCATG GTTAACACTA TCGGTACCCC GGGTCCGAGC TTCATTGTTA
121 CCGAACAAAC GCCGGCACCT ACGCCGGGTG ACGTGTTGAC CCCGCAGCCG ACCCCGCTGC
181 CTACCCTGGG TGGTGTGCCG ACGACGAAGA TGCCGACGGA AATGAGCTAT GGCTATGGTT
241 ACGGTGACTA CGGCATCGTC GACTGCTTTG GTAAGTCTGG TAAGAGCGGC AGCGGCTGTG
301 GTAAATCCGG TAAAGGCTCC AAGAGCTCTG GTAAGTCGGG TAAGTCCGGT GGCGGCGGTG
361 GCGGTGGTTA CGGCTATGGT GACAACTACG CGGATGATTA CACCCCGAGC ACCGATGACT
421 ACGAGTACGG TTACGGTCAT GGTGGTAGCT CCGGCAAGTC CGGCAAAGGC AGCAGCGGCA
481 AGAGCGGTAA GAGCTCTAGC AAAAGCTCTA AAGGTTCGGG TAAAAGCAGC AAAAGCAGCG
541 GTAAATCGTC TAAAAGCTCC GGTAAAAGCG GCAAGGGTGG TAGCCGTGAC GATGGCCACG
601 GTTATGGCGG CTACGGTGGC TATGAGGGCT ACGGCGGTTA TGAGGGTTAT CAGTACGGCG
661 GTGATGAATA TGTCCGTCGC AATCGCCGCC TGGGCGCCAG CCACAATAAC CGTATTCACC
721 ACCACCATCA TCAC**TAA**GGA TCC
```

rCinW2

```
   1 AGGAGGTAAA ACAT**ATG**CAA CAAAGCTCTG TCCGCGGTGT CGCAACTACT AGCTCTCGCC
  61 AACTGGATGA ATGGGGTGAC GACGCGTGGG GTTCGAGCGA TAGCGGCTCT AGCGGTAAAT
 121 CCGGCAAATC GGGTGGTAGC GCCAGCTCTG GCGATGGCTG GGAAACCGAC GGTTGGGGCG
 181 GTGATTACTC GAGCTCTAAA AGCGGTAAAA GCGGCTCCGG CAAGTCCGGC AAAGGTAGCT
 241 CGGGTCCGCA CGGTCATTGG GTGTATATCG AGGACGATAG CAGCGACGGC AGCGGTAAGT
 301 CGGGTAAAGG CTCGAGCTCC AAAGGTTCTA AAGGTTCCAG CAAGTCGAGC AAGGGCAGCT
 361 CTAGCGATGA CAGCACGGAT GACTCCTGGG ACGGTGGCTG GGGCGGTCAC GGCGGTTGGA
 421 ATGGTGATAA CAGCGGCAAG TCTGGTAAAG GTTCTTATGG TAGCGGTAAG AGCGGTAAGG
 481 GTAGCAGCTA CCCGAGCAGC CACTGGGGTC CGAGCCATTG GGGTAGCGAC GACGACGATT
 541 CGTCCTCTAG CAAAAGCTCC AAGGGTTCTA GCGAGAGCAG CTCTAAGAGC AGCAAGGGCT
 601 CCAGCGACAG CAGCAGCAAA TCCAGCAAAG GTTCGTCCAG CAGCGAGGAC GAAGGCCATT
 661 GGGAGTGGGA AGGCGGCTAT GGTTCCGGCA AGAGCGGCAA AGGCAGCTAC AGCGGTAGCT
 721 CCGGTAAGTC CGGTAAGTCT GGCAGCGGTG ACAGCTGGGT TGGTGATTAC GGCAGCTCTG
 781 GTAAATCTGG CAAGGGTAGC TATGGCGGTG ATAGCTGGGG TGGTAACTAC AATGGTTGGG
 841 GTGGCCACTA CGACGTGGAC GTTGATGACG ATGATAGCAG CTCCAGCAAG AGCTCTAAGG
 901 GTTCGTCTAA GAGCTCGAAA GGCAGCAGCG AAGATAGCAG CAAGTCCTCT AAGGGCTCTA
 961 GCAGCAAAAG CAGCAAAGGC TCCAGCAGCT CCGAGGACGA GGGTCACTGG GTTTGGGAAG
1021 GTAGCTACGG TAGCGGCAAA AGCGGCAAGG GTTCCTACTC CGGTAGCAGC GGCAAATCCG
1081 GTAAATCGGG TAGCGGCGAT GAGGGTTGGT ATAGCGGTTG GCATCATCAC CACCACCAC**T**
1141 **AA**GGATCC
```

FIGURE S1. DNA sequences of codon-optimized recombinant cingulins. Translation start and stop codons are highlighted in bold letters.

rCinY2 (240 amino acids, pI = 9.4)

```
  1 MGTNKTLPPTPFPGRPTPNPTMVNTIGTPGPSFIVTEQTPAPTPGDVLTPQPTPLPTLGG
 61 VPTTKMPTEMSYGYGYGDYGIVDCFGKSGKSGSGCGKSGKGSKSSGKSGKSGGGGGGGYG
121 YGDNYADDYTPSTDDYEYGYGHGGSSGKSGKGSSGKSGKSSSKSSKGSGKSSKSSGKSSK
181 SSGKSGKGGSRDDGHGYGGYGGYEGYGGYEGYQYGGDEYVRRNRRLGASHNNRIHHHHHH
```

rCinW2 (375 amino acids, pI = 6.3)

```
  1 MQQSSVRGVATTSSRQLDEWGDDAWGSSDSGSSGKSGKSGGSASSGDGWETDGWGGDYSS
 61 SKSGKSGSGKSGKGSSGPHGHWVYIEDDSSDGSGKSGKGSSSKGSKGSSKSSKGSSSDDS
121 TDDSWDGGWGGHGGWNGDNSGKSGKGSYGSGKSGKGSSYPSSHWGPSHWGSDDDDSSSSK
181 SSKGSSESSSKSSKGSSDSSSKSSKGSSSSEDEGHWEWEGGYGSGKSGKGSYSGSSGKSG
241 KSGSGDSWVGDYGSSGKSGKGSYGGDSWGGNYNGWGGHYDVDVDDDDSSSSKSSKGSSKS
301 SKGSSEDSSKSSKGSSSKSSKGSSSSEDEGHWVWEGSYGSGKSGKGSYSGSSGKSGKSGS
361 GDEGWYSGWHHHHHH
```

FIGURE S2. Amino acid sequences of recombinant cingulins. Lysine-rich regions are shown in red with KXXK motifs highlighted in yellow. Tyrosine-rich regions are depicted in green, tryptophan rich regions in blue, and regions rich in both tryptophan and tyrosine shown in purple. RXL motifs are highlighted in black with white letters.

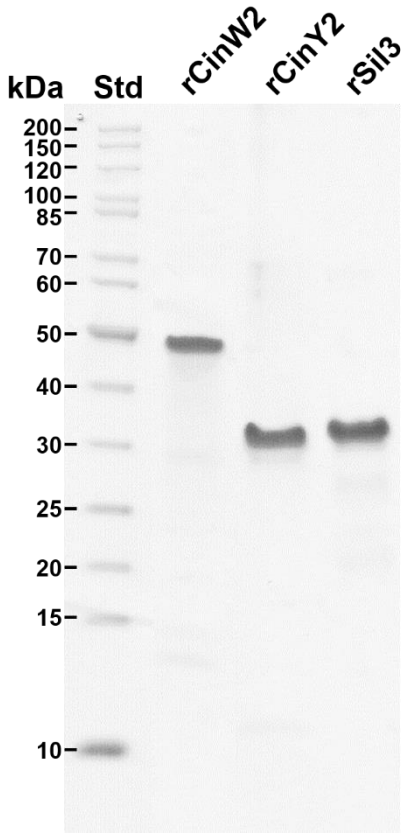

Figure S3. SDS-PAGE analysis or purified rCinY2, rCinW2, and rSil3. Two µg purified protein were loaded on each lane. The gel was stained with Coomassie Blue.

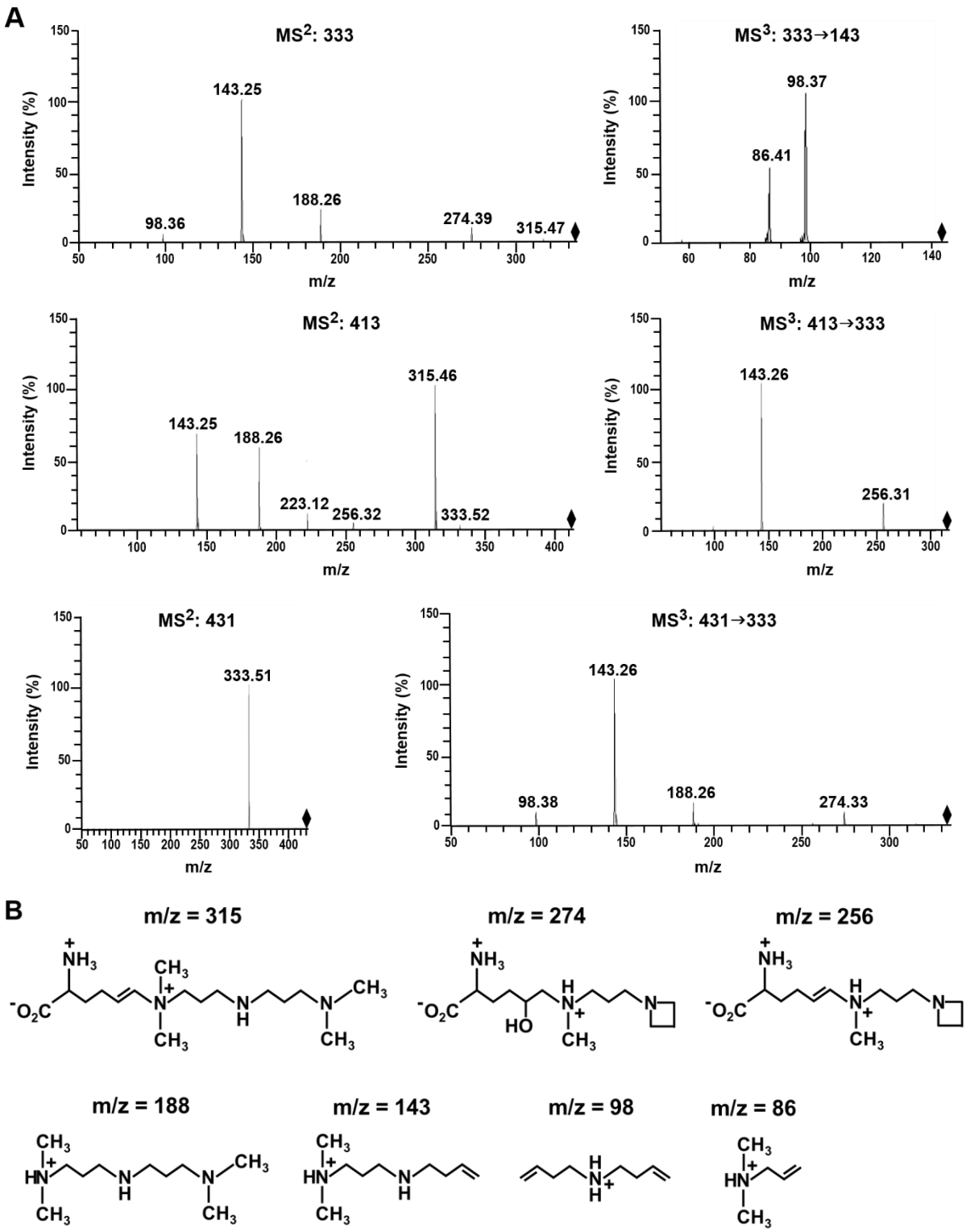

FIGURE S4. Mass spectrometric fragmentation analysis of lysine derivatives. *A*, MS^2 and MS^3 spectra were obtained by collision-induced fragmentation of singly charged ions with m/z ratios indicated at the top of each image. The black diamond indicates the position of the parent ion that was fragmented. *B*,predicted structures of the fragment ions from the spectra in *A*.

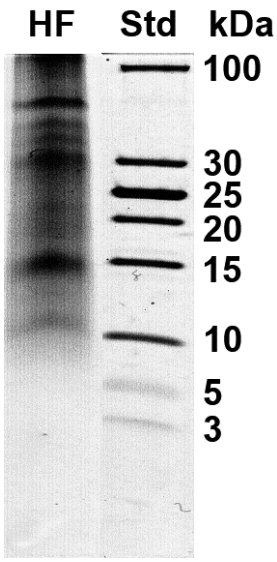

FIGURE S5. HF extract from the insoluble organic matrices. The HF extract was subjected to Tris-Tricine SDS-PAGE (16% acrylamide) (2) and stained with Coomassie Blue. Std = molecular mass standard.

SiMat1 (TP-ID21094, SFLP44)

```
  1 MKLSKFTTVATALAAASSSTVGVSALEGGLRKMMKDGYGYTDDYGYEHGLHSSSKSGKSG
 61 SSKGGKSESHGYAGYSMDYVETIVEEYGDDYGYGSGHGSSSKGGKSGSSKSGKGSSGDSS
121 KSGKSGSSKSGKGSKGGYHIMGKGSKSGGGGGYGSGYGYKYDDGYAADDNYVGADDVIGE
181 PSPTDDANFDDVVATTNPTGGFSIATTNPTGGFSTGATGTGSTGFGTGSTDFATGPTGTE
241 FFSTGPTV
```

SiMat2 (TP-ID25898, SFLP84)

```
  1 MKSSILLTIVFLPLILARRDGTTISEINEDVAPAIITETFATESLVATTDAATTDAPELV
 61 ELQIDEDVLRGTNKTLAPTPFPGRPTIPGTTPFPTENTPAPSPGEFGTRPPSPYSAKPPT
121 TGCSKSGKGGKGSKSGSMDYIIDCIDLSTKSGKANGGYGSSSKGGKSGGGKGSKGAYGTA
181 GSGKSGKPSNGYGGYGDDNYNHDDNYSPKAQYRVFDGPFRNGAVEAAAPEDILDGQIDAE
241 IQATLVQAELDEELMRGTNKTLAPTPFPKRPTPPGTTPFPTENTPAPSPGFGTKPPTPYS
301 AKPPTTGCSKSGKGGKGSKTGSMDYVIDCIDLSTKSGKTDGYGSGGSSKGGKSSGGYGDS
361 NDDTYAGVDDYVSGGKGGKSDGSSKSGKSSTGSYDGSGYGYGARFLLEEAVMDTVMAEGT
421 GFEGCNKFAVSTDWRYLFTGDNLPNM
```

SiMat3 (TP-ID22349, SFLP52)

```
  1 MRIGTSLMAVAALPLAVAASVDGHRQLSSKSGGGGWGNGGGWYHDHHDSSWSSSSSKSSK
 61 SGSSKSSKSGSSKSSKSGSSKSSKSGSSKSSKSGSSISSGDWGHGGWEGYSGKGGKSGGG
121 NWGGNWHSSSSSSSNSSGKSGKESKSSSKSSSKSSKSGGGYYHYNPRPQWDDDGHSD
181 HNSPGICAPDFHSGGLDCGHDSDDVDFKYCSEPGSDGECGSGFSCYDKKLCPKYSYEGGY
241 FSPEGVCDDNGKGLDCADVTTYCTEPGGHGECGAGRQCHDAGLCGYGGKGLDDYDGVCGA
301 AYGGIHGGGLDCGEVSTSCTDVGSKAECAHGATCFSSGMCHESGKHAGVCGSDDGFFCDD
361 VKTYCSSPGESGQCGRGKSCYDAGLCHGSGSDYSGVCAPKEGKGGLECGDVDQYCTEPGS
421 QGECGEGHTCYASDICYGGDSHGASGAVCFHNGDKVSCSGNPEGDAVYVQFSYSAETSGV
481 DPDDVVGPLEDKILSAVADHVSSGYGGSVTMISSDPSDYIKDDEACSTKYSGDKCSVIKG
541 EMTVYANGHELDACEYADIVEYSMYDKDFSNVHGVEKAEYIRAESSCNTNAIIGSGNAAE
601 DESLGAGAMIGLAMLAAALAAIALLAARRLRRKPATDKDFDIVSLDSNEGRFLGMGNDPF
661 ASTVDVHKCTSMYCNCNKGLGGTTFIPAPKKADMNKTLEAQGIASPQGVGEAQGFFVGEP
721 EAVDDLEMDEEPQDNIIRVNPSATEAHRSLTPVHEIAHDSEIDTEFESEGEEDMDSIPPP
781 PPLPPGHMQRAGLYDREMRDDEMSI
```

SiMat4 (TP-ID25912)

```
  1 MKIHSFTLSAFLALVTAATTNAQVPLVGYTKFGDNVCADGQYKFSPQPFSYVVTSYNGLSA
 61 GACAATCDTYGDSNGVMKQINSFTLAVDINQCRCNFDNGSTIPESIKDDAVIRFGVGDVEA
121 MVNGDIGEAHCYTRDGYKEPVPTQTPTASPTIPDIKFYYPATPNLDFGPCADSQGRILPFG
181 LVPFLDEPSNDARDAEICARNCFDFGSIYNNWEEQVGMGLTVDLDTDPMAVYCKCYFESNE
241 NLPRTIDGEFIRFPKLGTGAVAGIGGQGDEEKETESTICYVREAYVASSAPSVSSRPSISP
301 KPTRSPTRSPVTPRPSRSPTKSPSVSPTLAPSTSPTLSPSLSPTLSPSASPTTPTTSPTLS
361 PSLSPTLSPTLSPTLSPTLSPTFYPTFSPTMSPSNNPTQAPSISPSRSPNNNVVDDFFPTM
421 SPTMDSVPNIGCSKSSKRGKGLRTKSGKAGRCKSGKGNRHEIDWVRPQKPWKGSGFNKKDM
481 RFVFDDDSNTDDNGLPLDIVVGPSGKPLKKSVVNATITTTVNNEDI
```

FIGURE S6 (part 1).

SiMat5 (TP-ID21757)

```
  1 MLLIPLILRLFAITTASSSEWQHTNPSAASLRRQLQDDATSTTSNTTPDDGDLPFHKAT
 61 FLASHNAHANRDAASSFFETLGINQDSSIYDQLSNNDVRGLLLDIKLDPNFADEQLRLV
121 HGPLDFGGFSSVANENLIPFLEENPNAIVTLILETTGDSGEYEATIRANILKELQTIFS
181 ALSVNGQPLKEITFKYDDLLWQNHDNWPTLSEIRQSGQRLFIFSDRSELANSEYGFMHN
241 QQVMKENYWEGVVDCIAQFGWDLSTVSLPSNQSWSRLFMMNHFCCESGAESFGRVVGEA
301 LLGGGDNGWGILYPRIQNCMANNGGVTPNFIALDWVVNSEEARAVRDYLKFGGAVGRGQ
361 TCDDDSQCATSSCNTAAGICQCQECASNSLDICPGCASGQYCQSAGDSSANQCIVKERI
421 ENSYVCSTSFDSAVASCNTAVRCPNGNDDCPVDQVCFNAVDCLPAPTLQPTDQSSSSSS
481 SDATTSVQVQEITTTVPATAETSESPSAAVIEPTAAPITPASRYCGESYEDAKTTCSEI
541 TACPKGYECPSGLTCYDGVKCFTRRPTSSPTDTPTTASPISPPSASPVTDAPIAPSMSP
601 TDRPTRAPFDFFNEYFCGGNFTEAQSSCYTTTPCPTGSPASCLNGETCYGGIKCIAPPS
661 ISPTLQPTDKPPTKSPSAVVQEQTTSPPFNWLNTNGGMAAMSGGYAIVAMIVGVAGVML
721 W
```

SiMat6 (TP-ID24761)

```
  1 MPMKKVLLALLLIIVEAASVSANTLRGSEESKRELYGGTNRSGGYGYRHPYRFIGRRGGC
 61 KPDLYLVEDDSERSTKKCTDNLRDEFWREVNSVKTVHTAIGARYGDLIYYVNQATNCGDV
121 WIEIDEDIRLHRGRIRCYRHSYGYDGSSKSSSSSQDNDNDSTSSEDNDNDSTSSEDNDND
181 STSSEDNDD
```

SiMat7 (TP-ID24710)

```
  1 MKFVLPAILLATATANPFAPKQTRNTKKAAYAASLMRGATPLRRLEDAYDGQVDVDLSGY
 61 SVKFEKCQFVKQYEGGEGGNNNNKNGNGEQFLSTKRFVIFRLCPDSSCSSCNYNYGEYIV
121 DMDTYLESTLQYKQEEQETYCQSCQQCVEMQANANNGDANDDQNNDNAWMCNNIDTSTCY
181 DECQNIENMEANGYMDASELTGCVKMNYQDNYGNAYYAGAMCASSGTRIKIGVFSDEQCS
241 QVVEDADIETYLAYGNNNDNNNGVTMKLSYHLLKQTFPESGCVSSCLKQNENQNNNNNNN
301 GEQQAAEVNEICENLYEVAGKCESTHGFKTGYANYDNYENQIRNEELVCDFISSVSAGHY
361 DQTGEIVVSGGRTTLGGGVATTGGQKFALTFFILGSVGLAGYAAMLHQQLTKGAKADLSR
421 QGGAMA
```

FIGURE S6 (part 2). Amino acid sequences of SiMat proteins. Regions rich in both lysine and serine are shown in red. Tyrosine-rich regions are shown in green, tryptophan rich regions in blue, and regions rich in both tryptophan and tyrosine are shown in purple. Asparagine-rich regions are depicted in orange. KXXK motifs are highlighted in yellow, and RXL motifs are shown in white letters on a black background.

5’-end of SiMat1

```
  1 GGGGGGGGGG GGTTATTTGC AAATCGCAGC ATTACACACA ATCACAATTG ATACACAATC
 61 ACACGCTACA CCTCCACCTC TTTGTTCTCT GTTCTCTGTC ATCGTTCAAC GTCAATCTGT
121 AATCATGAAG CTCTCAAAGT TTACTACCGT CGCCACAGCT CTCGCTGCTG CCTCTTCCAG
181 GCACCGTTGG TATCTTCTA
```

3’-end of SiMat1

```
  1 TACCGGTGGT TTCTCCACCA GTGCTACTGA GACGGGAAGT ACTGGCTTCGG AACGGGAAG
 61 CACCGACTTC GCCACTGGCC CTGCTGGAAC CGAGTTCTTT TCTACTGGTCC TACCGTGTA
121 GAGAGGATGT TGTACTACAG TCACAATCTT TAGAGATACC TGCCATGAGGT AGTTGTCAG
181 ATACTTTGAT CCTCGAGAGA TCAATTCCTT AGAGATGTTT GCAAGCTGAGT TCTATATGC
241 ACCGGACGGA AGCGCATCAT CTATTAGTTA CAGTTACTAC ATTGATATTGC TGTATTCTC
301 TCATGAGAAA AGAGAGAGTC TTAAACCTCG TCACGAAAAA TGCCTTTCAAA GAAGAAAGA
361 TTCAAGACAA TTATGAAGCA GCAACGATTA CATACCAACT AAAGTAGATTC TTAGATTAG
421 TTGAAAAAAA AAAAAAAAAA AA
```

Figure S7. DNA sequences obtained from 5’- and 3’-RACE PCR for the SiMat1 cDNA. Start and stop codons are highlighted in red, polyG- and polyA-tails are underlined.

TABLES

TABLE S1. Amino acid composition of the insoluble organic matrices. The analysis was performed three times from independent isolates of the insoluble organic matrices. Standard deviations are provided. 'Trace' indicates detectable levels of an amino acids with a relative abundance <1.0 mol-%. Hyp = hydroxyproline, Dihyp = dihydroxyproline.

	Amino Acid	mol-%
Non-polar	Ala	3.7 ±0.4
	Gly	27.7 ±3.4
	Ile	1.7 ±0.3
	Leu	1.9 ±0.3
	Met	trace
	Phe	2.4 ±1.9
	Pro	1.9 ±1.4
	Trp	trace
	Val	2.0 ±1.1
Polar, uncharged	Cys	trace
	Ser[+]	25.0 ±1.4
	Thr[+]	3.2 ±1.4
	Tyr	10.6 ±2.8
	Dihyp	2.3 ±0.6
	Hyp	trace
Cationic	Arg	trace
	His	1.1 ±0.1
	Lys[§]	trace
Anionic	Asx	10.3 ±0.4
	Glx	2.7 ±1.5

[+]includes O-phosphorylated derivatives
[§]does not include modified lysines

TABLE S2. Monosaccharide composition of the insoluble organic matrices. 'Trace' indicates detected amino acid with a relative abundance <1.0 mol-%.

	Monosaccharide	mol-%
Hexoses	Fructose	trace
	Galactose	9
	Mannose	23
Pentoses	Ribose	trace
	Xylose	20
Deoxyhexoses	Fucose	3
	Rhamnose	2
Aminosugars	Galactosamine	trace
	Glucosamine	14
Uronic acids	Galacturonic acid	6
	Glucuronic acid	14
	Mannuronic acid	7

TABLE S3. Proteins identified by proteomics analysis of the HF extract from the insoluble organic matrices (see corresponding file: **TableS3.xls**).

REFERENCES

1. Poulsen, N., and Kröger, N. (2004) Silica morphogenesis by alternative processing of ailaffins in the diatom *Thalassiosira pseudonana*. *J. Biol. Chem.* **279**, 42993–42999
2. Schägger, H., and von Jagow, G. (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal. Biochem.* **166**, 368–379