Supplemental Figure 1. Fold Recognition Performance of Ada-BLAST with Different Settings Given Fold-specific Libraries. (a) Relative frequency of # of PSSMs generated from each TZ-SABmark queries. The average number of PSSMs generated from the queries is 3.6, with the most being 31 PSSMs which were generated from *d1tfra1*, a *SAM-domain like* fold. If a TZ-SABmark query was not in the version of PDB which we used to construct fold-specific libraries, PSSMs were not generated from the query. There are 60 such TZ-SABmark queries. (b) Comparison of ROC curves of Ada-BLAST at different coverage thresholds when e-value threshold is fixed at 10^{10}. (c) Comparison of ROC curves of Ada-BLAST of e-value 0.01 & no coverage, e-value 0.01 & 80% coverage, e-value 10^{10} & no coverage, and e-value 10^{10} & 80% coverage.

Supplemental Figure 2. Characterization of alignments used by Ada-BLAST at e-value 0.01 and 10^{10} thresholds. Pairwise identity of the alignments collected by Ada-BLAST using rps-BLAST with e-value 0.01, no coverage and e-value 10^{10}, 80% coverage. The analysis is done separately for the alignments between queries and the PSSMs of true fold-specific library and those between the queries and the PSSMs of false fold-specific library.

Supplemental Figure 3. Comparison of Ada-BLAST dendrograms of e-value 0.01 and 10^{10} thresholds. A portion of the e-value 0.01, no coverage dendrogram containing the queries mis-clustered (*top left box)* and a portion of the e-value 10^{10}, 80% coverage dendrogram containing the same queries. The queries in blue boxes (queries in red in the dendrogram of e-value 0.01, no coverage in Fig. 2b) were improperly clustered in the result of e-value 0.01, no coverage, but properly clustered in the result of e-value 10^{10}, 80% coverage.

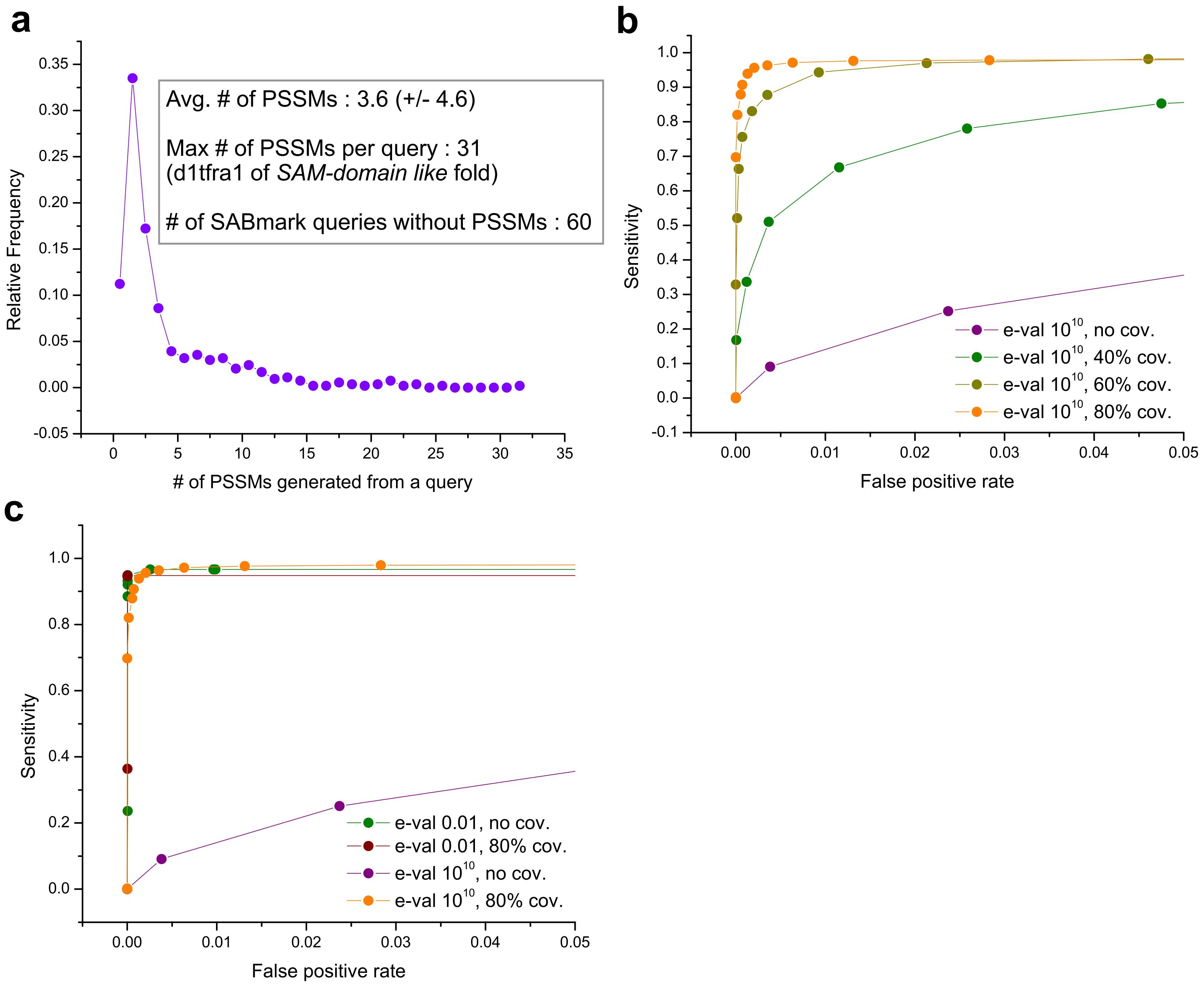
Supplemental Figure 4-5. Hierarchical clustering of transmembrane containing proteins with additional Ada-BLAST settings. 74 sequences representing multiple classes of transmembrane containing proteins were hierarchically clustered and visualized by Cluster and Treeview [21]. The horizontal lines represent the correlation scores derived from the analysis. Alignments for the ILB DB PSSMs were derived from embedded alignments (Fig. S4) and e=0.01 (Fig. S5) threshold.

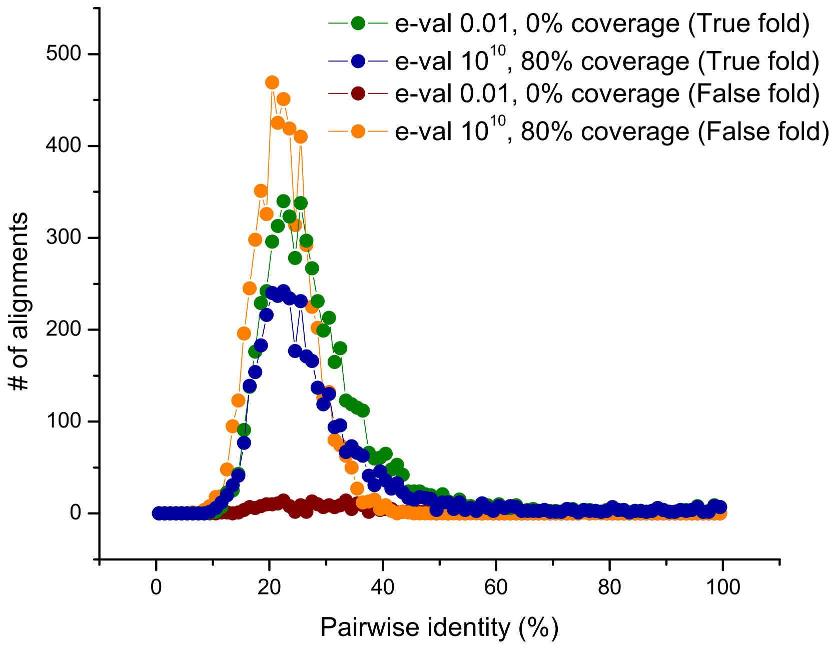
Supplemental Figure 6. rps-BLAST Analysis of 1F88. The primary amino acid sequence for 1F88 (gi|197107531) was screened by rps-BLAST at e-value=1. This figure shows the web-based output for domain identification.

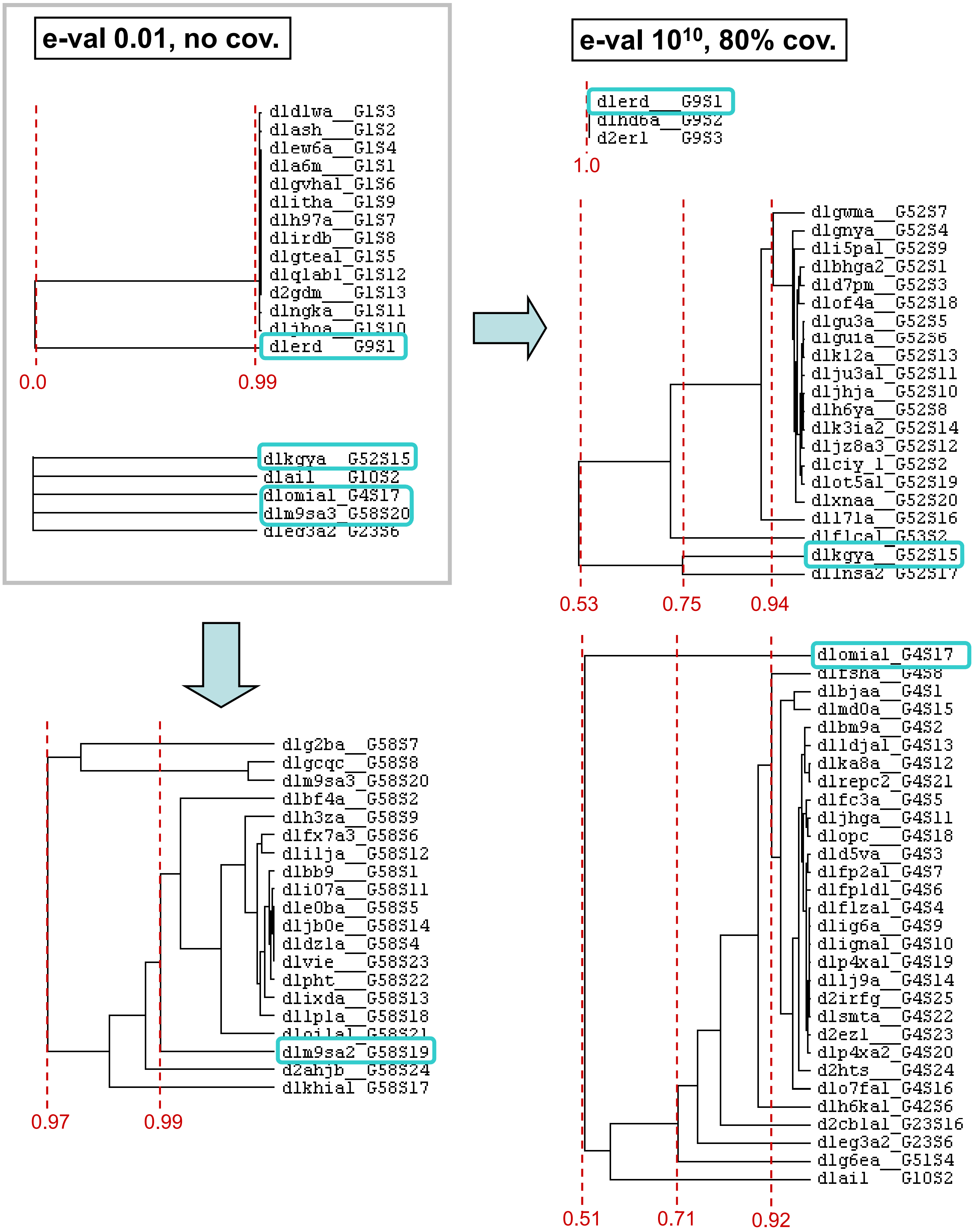
Supplemental Figure 7. The Characterization of Structural Elements in 1F88. This graph shows the performance of Hidden Markov Models (TMHMM) vs embedded Ada-BLAST in determining the membrane spanning domains in Bovine Rhodopsin as determined by X-ray Crystallography (green= Beta pleated sheets, red=helices, loops not shown). This protein was analyzed with an expanded set of PSSMs representing a large variety of transmembrane domains (~30K PSSMs). It is resonable to consider that the amino-acids within transmembrane spanning helicies will be more conserved than the intervening loop residues. The support of this hypothesis is presented herein. The structural features are annotated with droplines. The transmembrane probability determined by TMHMM is shown in teal. The other graph depicts a curve-fitted positional score for embedded Ada-BLAST (see Figure 4 for raw and smoothed data). The positional score was quantified in the following manner. For each positive PSSM, the alignment boundaries are determined by the overlapping alignments obtained from Ada-BLAST. These regions were extracted and realigned by the Smith-Waterman algorithm with a BLOSUM62 substitution matrix. Using the alignments, each residue was scored with substitution scores of BLOSUM62 if the residue is identically or positively (non-identical but conserved) aligned. This process was repeated for all positive PSSMs and the results were summed for each amino acid in the protein. The positional results were normalized to zero by subtracting the average positional score across the protein length from each point, and each amino acid position was then subjected to smoothing (Fast Fourier-transform point=8) and discontinous baselining using Origin Lab 7.5© . Baseline correction was performed by baselining to every local minimum across the entire curve.

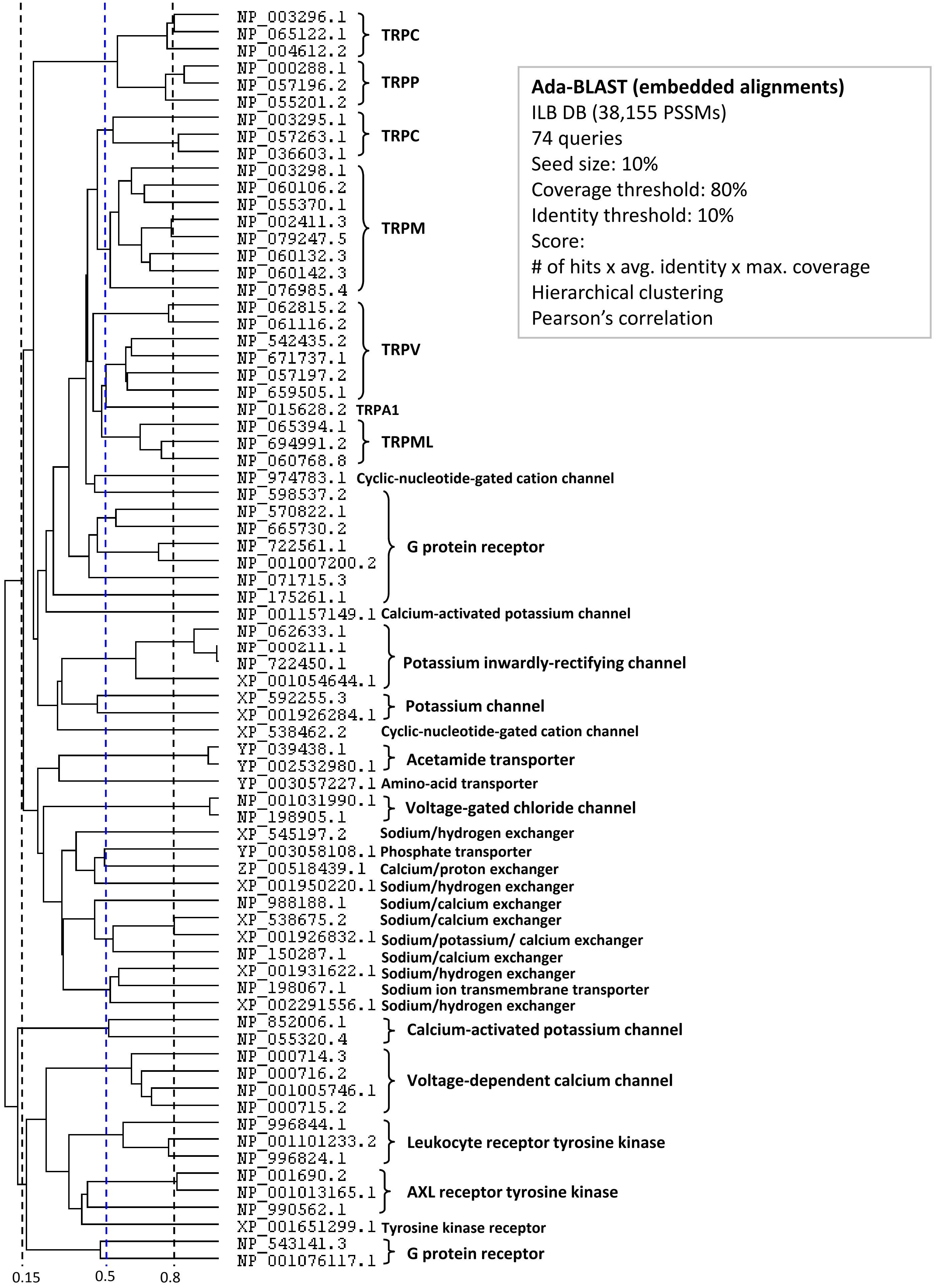
Supplemental Figure 8. Ada-BLAST comparison with TMHMM and rps-BLAST for TRPC3. The top graphic depicts the domain organization predicted by rps-BLAST. Below this is a graph containing TM probabilities predicted by hidden-markov model TMHMM (magenta, right y-axis) compared with postitional Ada-BLAST data from embedded alignments with >30K TM PSSMs.

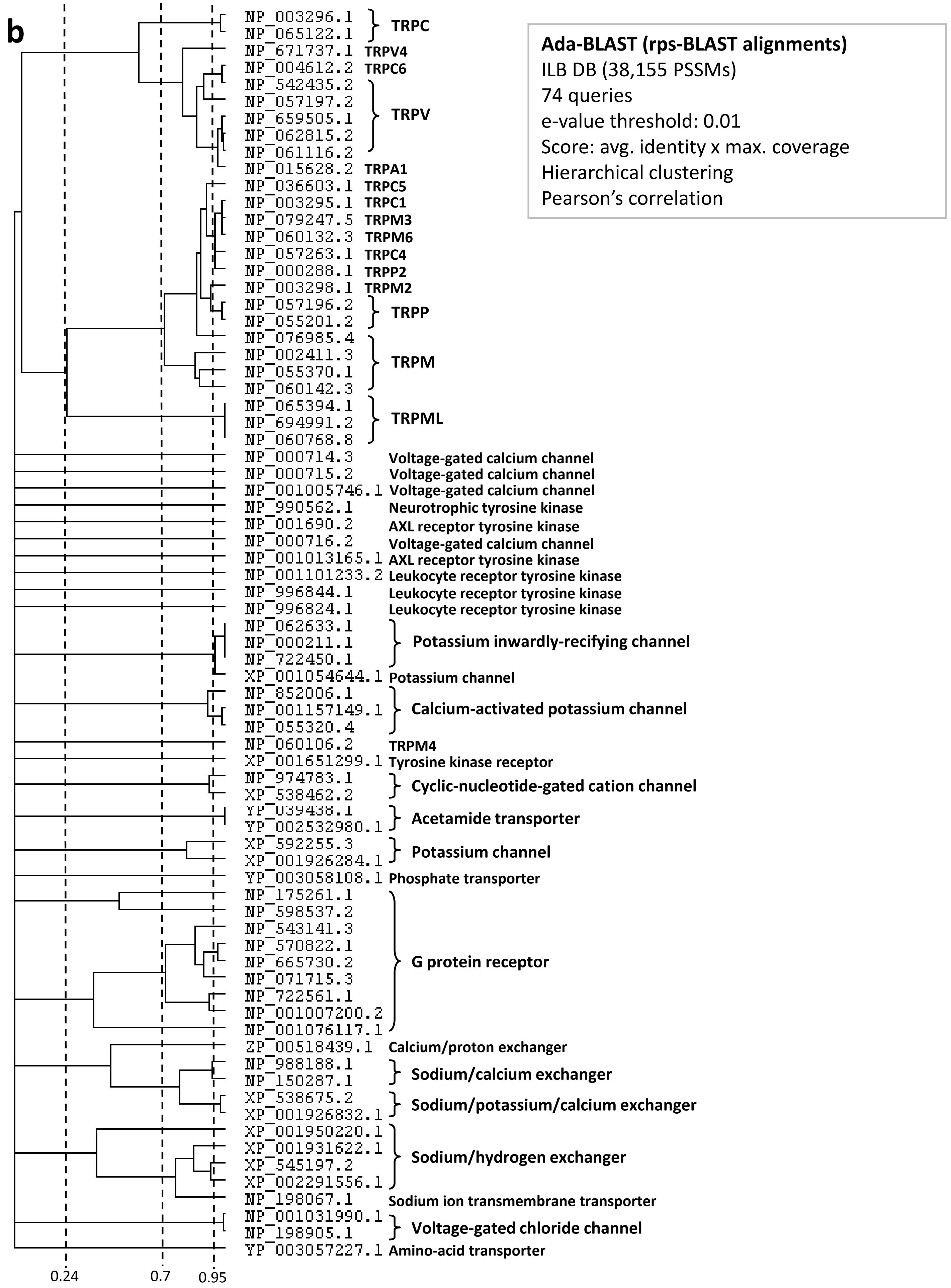
Supplemental Figure 9. Ada-BLAST comparison with TMHMM and rps-BLAST for TRPV5. The top graphic depicts the domain organization predicted by rps-BLAST. Note that rps-BLAST does not show a domain for the known channel region. Below this is a graph containing TM probabilities predicted by hidden-markov model TMHMM (magenta, right y-axis) compared with postitional Ada-BLAST data from embedded alignments with >30K TM PSSMs.


a
Relative Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
0
5
10
15
20
25
30
35
# of PSSMs generated from a query
Avg. # of PSSMs : 3.6 (+/- 4.6)
Max # of PSSMs per query : 31
(d1tfra1 of *SAM-domain like* fold)
# of SABmark queries without PSSMs : 60
b
Sensitivity
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0.00
0.01
0.02
0.03
0.04
0.05
False positive rate
e-val 10^{10}, no cov.
e-val 10^{10}, 40% cov.
e-val 10^{10}, 60% cov.
e-val 10^{10}, 80% cov.
c
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.01
0.02
0.03
0.04
0.05
False positive rate
e-val 0.01, no cov.
e-val 0.01, 80% cov.
e-val 10^{10}, no cov.
e-val 10^{10}, 80% cov.

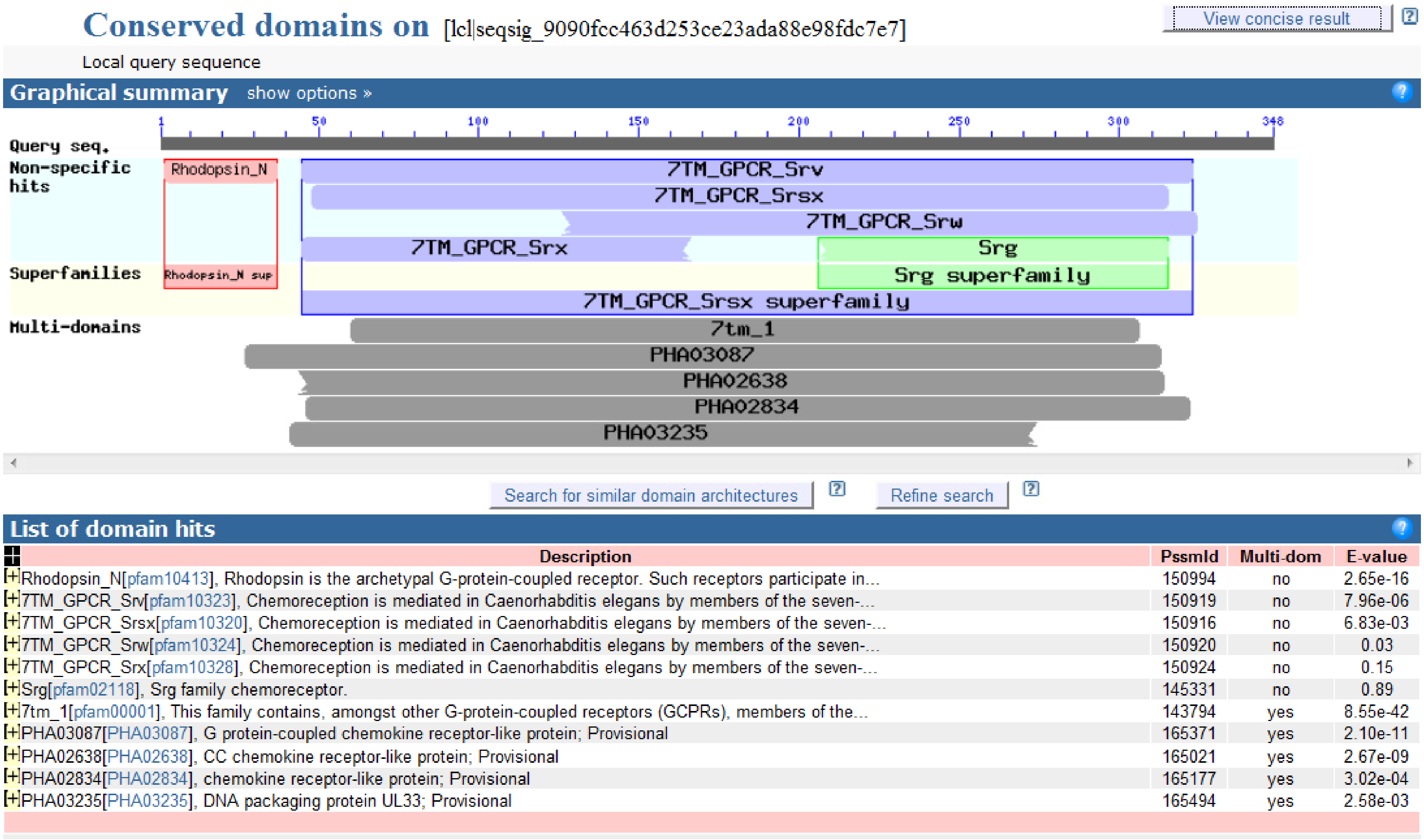

● e-val 0.01, 0% coverage (True fold)
● e-val 10^{10}, 80% coverage (True fold)
● e-val 0.01, 0% coverage (False fold)
● e-val 10^{10}, 80% coverage (False fold)
# of alignments
500
400
300
200
100
0
0
20
40
60
80
100
Pairwise identity (%)


e-val 0.01, no cov.
d1dlwa__G1S3
d1ash___G1S2
d1ew6a__G1S4
d1a6m___G1S1
d1gvha1_G1S6
d1itha__G1S9
d1h97a__G1S7
d1irdb__G1S8
d1gtea1_G1S5
d1qlab1_G1S12
d2gdm___G1S13
d1ngka__G1S11
d1jhoa__G1S10
d1erd___G9S1
0.0
0.99
d1kgya__G52S15
d1ail___G10S2
d1omia1_G4S17
d1m9sa3_G58S20
d1eg3a2_G23S6
e-val 10^{10}, 80% cov.
d1erd___G9S1
d1hd6a__G9S2
d2erl___G9S3
1.0
d1gwma__G52S7
d1gnya__G52S4
d1i5pa1_G52S9
d1bhga2_G52S1
d1d7pm__G52S3
d1of4a__G52S18
d1gu3a__G52S5
d1gu1a__G52S6
d1k12a__G52S13
d1ju3a1_G52S11
d1jhja__G52S10
d1h6ya__G52S8
d1k3ia2_G52S14
d1jz8a3_G52S12
d1ciy_1_G52S2
d1ot5a1_G52S19
d1xnaa__G52S20
d1l7la__G52S16
d1f1ca1_G53S2
d1kgya__G52S15
d1lnsa2_G52S17
0.53
0.75
0.94
d1g2ba__G58S7
d1gcqc__G58S8
d1m9sa3_G58S20
d1bf4a__G58S2
d1h3za__G58S9
d1fx7a3_G58S6
d1ilja__G58S12
d1bb9___G58S1
d1i07a__G58S11
d1e0ba__G58S5
d1jb0e__G58S14
d1dzla__G58S4
d1vie___G58S23
d1pht___G58S22
d1ixda__G58S13
d1lpla__G58S18
d1oi1a1_G58S21
d1m9sa2_G58S19
d2ahjb__G58S24
d1khia1_G58S17
0.97
0.99
d1omia1_G4S17
d1fsha__G4S8
d1bjaa__G4S1
d1md0a__G4S15
d1bm9a__G4S2
d1ldja1_G4S13
d1ka8a__G4S12
d1repc2_G4S21
d1fc3a__G4S5
d1jhga__G4S11
d1opc___G4S18
d1d5va__G4S3
d1fp2a1_G4S7
d1fp1d1_G4S6
d1flza1_G4S4
d1ig6a__G4S9
d1igna1_G4S10
d1p4xa1_G4S19
d1lj9a__G4S14
d2irfg__G4S25
d1smta__G4S22
d2ez1___G4S23
d1p4xa2_G4S20
d2hts___G4S24
d1o7fa1_G4S16
d1h6ka1_G42S6
d2cbla1_G23S16
d1eg3a2_G23S6
d1g6ea__G51S4
d1ail___G10S2
0.51
0.71
0.92


NP_003296.1
NP_065122.1
NP_004612.2
TRPC
NP_000288.1
NP_057196.2
NP_055201.2
TRPP
NP_003295.1
NP_057263.1
NP_036603.1
TRPC
NP_003298.1
NP_060106.2
NP_055370.1
NP_002411.3
NP_079247.5
NP_060132.3
NP_060142.3
NP_076985.4
TRPM
NP_062815.2
NP_061116.2
NP_542435.2
NP_671737.1
NP_057197.2
NP_659505.1
TRPV
NP_015628.2 TRPA1
NP_065394.1
NP_694991.2
NP_060768.8
TRPML
NP_974783.1 Cyclic-nucleotide-gated cation channel
NP_598537.2
NP_570822.1
NP_665730.2
NP_722561.1
NP_001007200.2
NP_071715.3
NP_175261.1
G protein receptor
NP_001157149.1 Calcium-activated potassium channel
NP_062633.1
NP_000211.1
NP_722450.1
XP_001054644.1
Potassium inwardly-rectifying channel
XP_592255.3
XP_001926284.1
Potassium channel
XP_538462.2 Cyclic-nucleotide-gated cation channel
YP_039438.1
YP_002532980.1
Acetamide transporter
YP_003057227.1 Amino-acid transporter
NP_001031990.1
NP_198905.1
Voltage-gated chloride channel
XP_545197.2 Sodium/hydrogen exchanger
YP_003058108.1 Phosphate transporter
ZP_00518439.1 Calcium/proton exchanger
XP_001950220.1 Sodium/hydrogen exchanger
NP_988188.1 Sodium/calcium exchanger
XP_538675.2 Sodium/calcium exchanger
XP_001926832.1 Sodium/potassium/ calcium exchanger
NP_150287.1 Sodium/calcium exchanger
XP_001931622.1 Sodium/hydrogen exchanger
NP_198067.1 Sodium ion transmembrane transporter
XP_002291556.1 Sodium/hydrogen exchanger
NP_852006.1
NP_055320.4
Calcium-activated potassium channel
NP_000714.3
NP_000716.2
NP_001005746.1
NP_000715.2
Voltage-dependent calcium channel
NP_996844.1
NP_001101233.2
NP_996824.1
Leukocyte receptor tyrosine kinase
NP_001690.2
NP_001013165.1
NP_990562.1
AXL receptor tyrosine kinase
XP_001651299.1 Tyrosine kinase receptor
NP_543141.3
NP_001076117.1
G protein receptor
0.15
0.5
0.8
Ada-BLAST (embedded alignments)
ILB DB (38,155 PSSMs)
74 queries
Seed size: 10%
Coverage threshold: 80%
Identity threshold: 10%
Score:
# of hits x avg. identity x max. coverage
Hierarchical clustering
Pearson’s correlation


b
Ada-BLAST (rps-BLAST alignments)
ILB DB (38,155 PSSMs)
74 queries
e-value threshold: 0.01
Score: avg. identity x max. coverage
Hierarchical clustering
Pearson's correlation
NP_003296.1
NP_065122.1
TRPC
NP_671737.1 TRPV4
NP_004612.2 TRPC6
NP_542435.2
NP_057197.2
NP_659505.1
NP_062815.2
NP_061116.2
TRPV
NP_015628.2 TRPA1
NP_036603.1 TRPC5
NP_003295.1 TRPC1
NP_079247.5 TRPM3
NP_060132.3 TRPM6
NP_057263.1 TRPC4
NP_000288.1 TRPP2
NP_003298.1 TRPM2
NP_057196.2
NP_055201.2
TRPP
NP_076985.4
NP_002411.3
NP_055370.1
NP_060142.3
TRPM
NP_065394.1
NP_694991.2
NP_060768.8
TRPML
NP_000714.3 Voltage-gated calcium channel
NP_000715.2 Voltage-gated calcium channel
NP_001005746.1 Voltage-gated calcium channel
NP_990562.1 Neurotrophic tyrosine kinase
NP_001690.2 AXL receptor tyrosine kinase
NP_000716.2 Voltage-gated calcium channel
NP_001013165.1 AXL receptor tyrosine kinase
NP_001101233.2 Leukocyte receptor tyrosine kinase
NP_996844.1 Leukocyte receptor tyrosine kinase
NP_996824.1 Leukocyte receptor tyrosine kinase
NP_062633.1
NP_000211.1
NP_722450.1
Potassium inwardly-recifying channel
XP_001054644.1 Potassium channel
NP_852006.1
NP_001157149.1
NP_055320.4
Calcium-activated potassium channel
NP_060106.2 TRPM4
XP_001651299.1 Tyrosine kinase receptor
NP_974783.1
XP_538462.2
Cyclic-nucleotide-gated cation channel
YP_039438.1
YP_002532980.1
Acetamide transporter
XP_592255.3
XP_001926284.1
Potassium channel
YP_003058108.1 Phosphate transporter
NP_175261.1
NP_598537.2
NP_543141.3
NP_570822.1
NP_665730.2
NP_071715.3
NP_722561.1
NP_001007200.2
NP_001076117.1
G protein receptor
ZP_00518439.1 Calcium/proton exchanger
NP_988188.1
NP_150287.1
Sodium/calcium exchanger
XP_538675.2
XP_001926832.1
Sodium/potassium/calcium exchanger
XP_001950220.1
XP_001931622.1
XP_545197.2
XP_002291556.1
Sodium/hydrogen exchanger
NP_198067.1 Sodium ion transmembrane transporter
NP_001031990.1
NP_198905.1
Voltage-gated chloride channel
YP_003057227.1 Amino-acid transporter
0.24
0.7
0.95

1F88

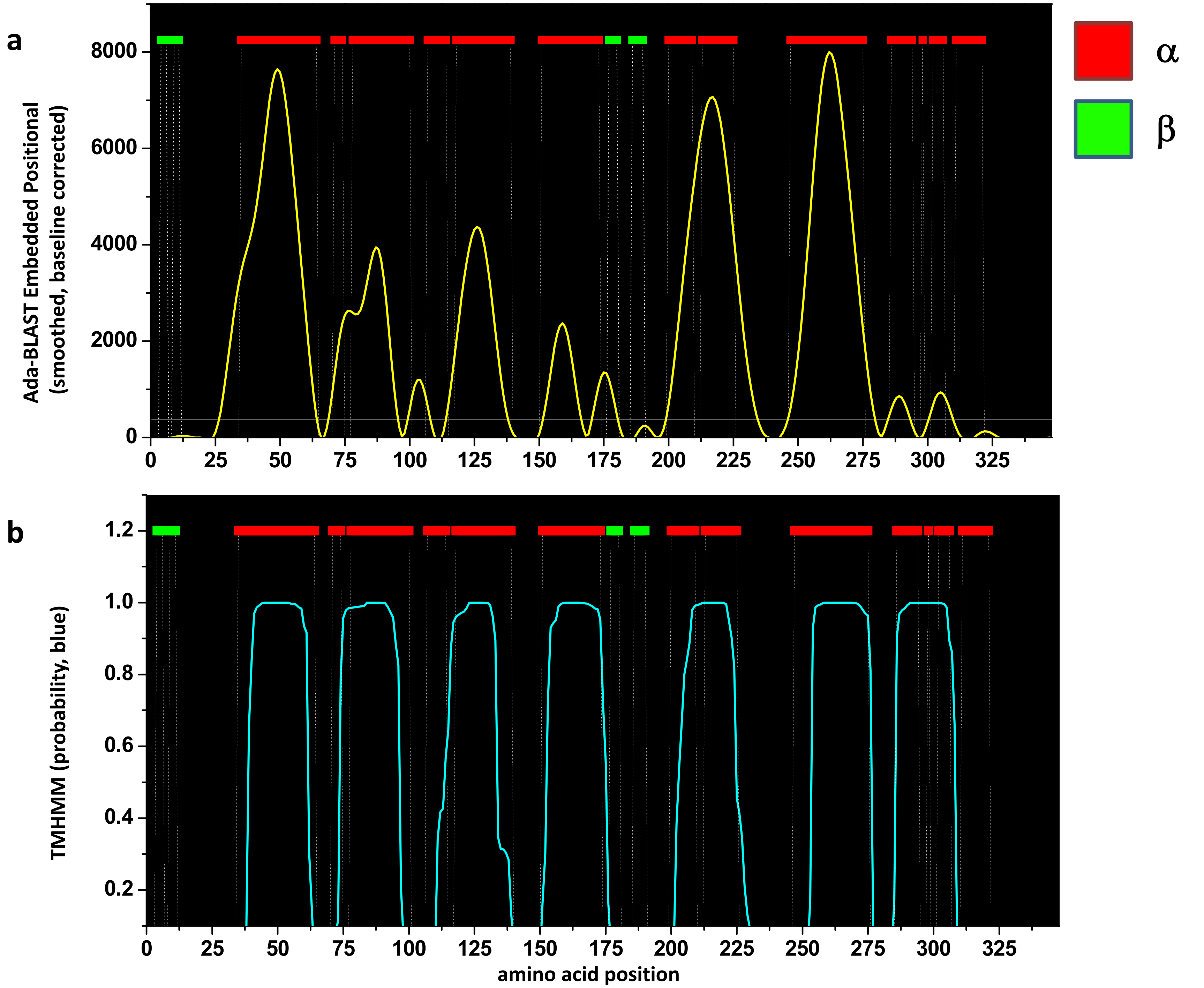


Conserved domains on [lcl|seqsig_9090fcc463d253ce23ada88e98fdc7e7]

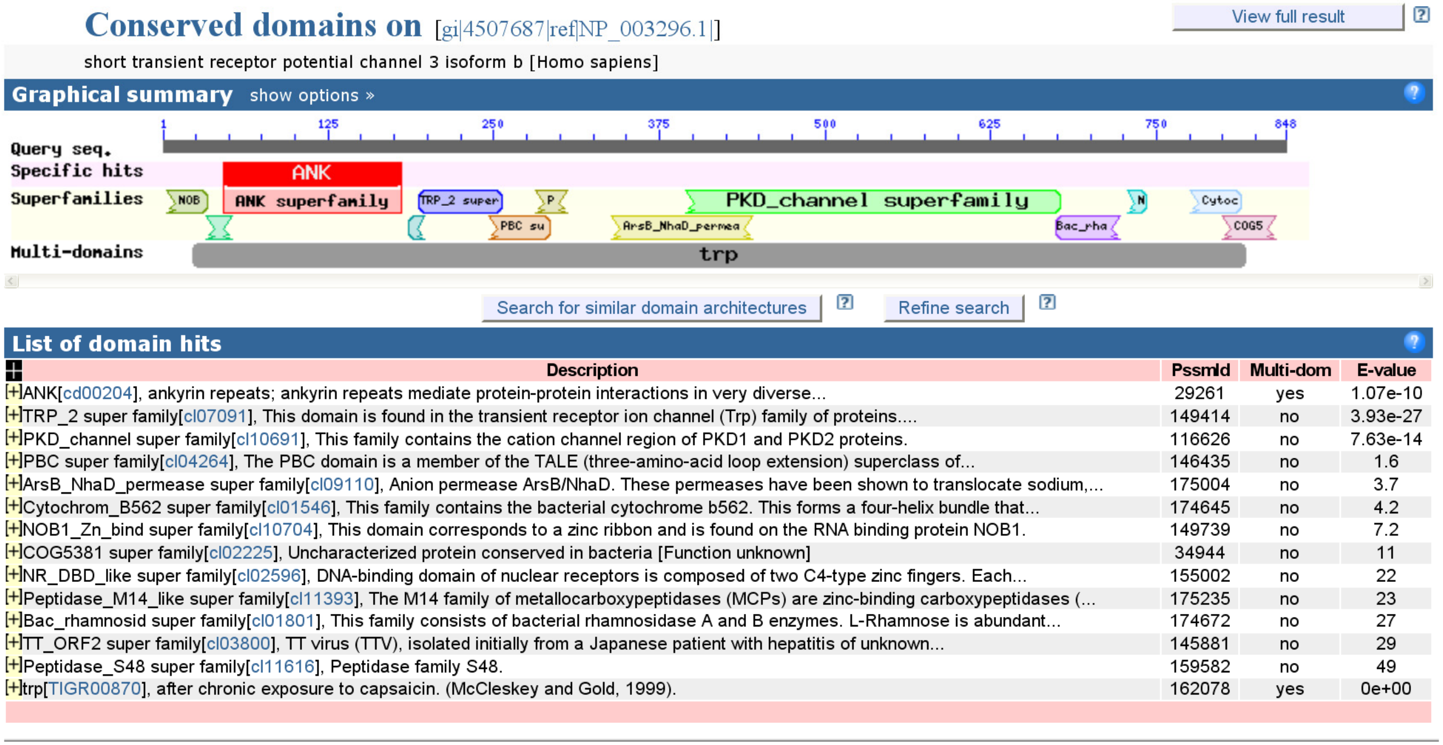
Local query sequence

List of domain hits

Description	PssmId	Multi-dom	E-value
[+]Rhodopsin_N[pfam10413], Rhodopsin is the archetypal G-protein-coupled receptor. Such receptors participate in...	150994	no	2.65e-16
[+]7TM_GPCR_Srv[pfam10323], Chemoreception is mediated in Caenorhabditis elegans by members of the seven-...	150919	no	7.96e-06
[+]7TM_GPCR_Srsx[pfam10320], Chemoreception is mediated in Caenorhabditis elegans by members of the seven-...	150916	no	6.83e-03
[+]7TM_GPCR_Srw[pfam10324], Chemoreception is mediated in Caenorhabditis elegans by members of the seven-...	150920	no	0.03
[+]7TM_GPCR_Srx[pfam10328], Chemoreception is mediated in Caenorhabditis elegans by members of the seven-...	150924	no	0.15
[+]Srg[pfam02118], Srg family chemoreceptor.	145331	no	0.89
[+]7tm_1[pfam00001], This family contains, amongst other G-protein-coupled receptors (GCPRs), members of the...	143794	yes	8.55e-42
[+]PHA03087[PHA03087], G protein-coupled chemokine receptor-like protein; Provisional	165371	yes	2.10e-11
[+]PHA02638[PHA02638], CC chemokine receptor-like protein; Provisional	165021	yes	2.67e-09
[+]PHA02834[PHA02834], chemokine receptor-like protein; Provisional	165177	yes	3.02e-04
[+]PHA03235[PHA03235], DNA packaging protein UL33; Provisional	165494	yes	2.58e-03


a
Ada-BLAST Embedded Positional (smoothed, baseline corrected)
0
2000
4000
6000
8000
0
25
50
75
100
125
150
175
200
225
250
275
300
325
α
β
b
TMHMM (probability, blue)
0.2
0.4
0.6
0.8
1.0
1.2
0
25
50
75
100
125
150
175
200
225
250
275
300
325
amino acid position

rps-BLAST

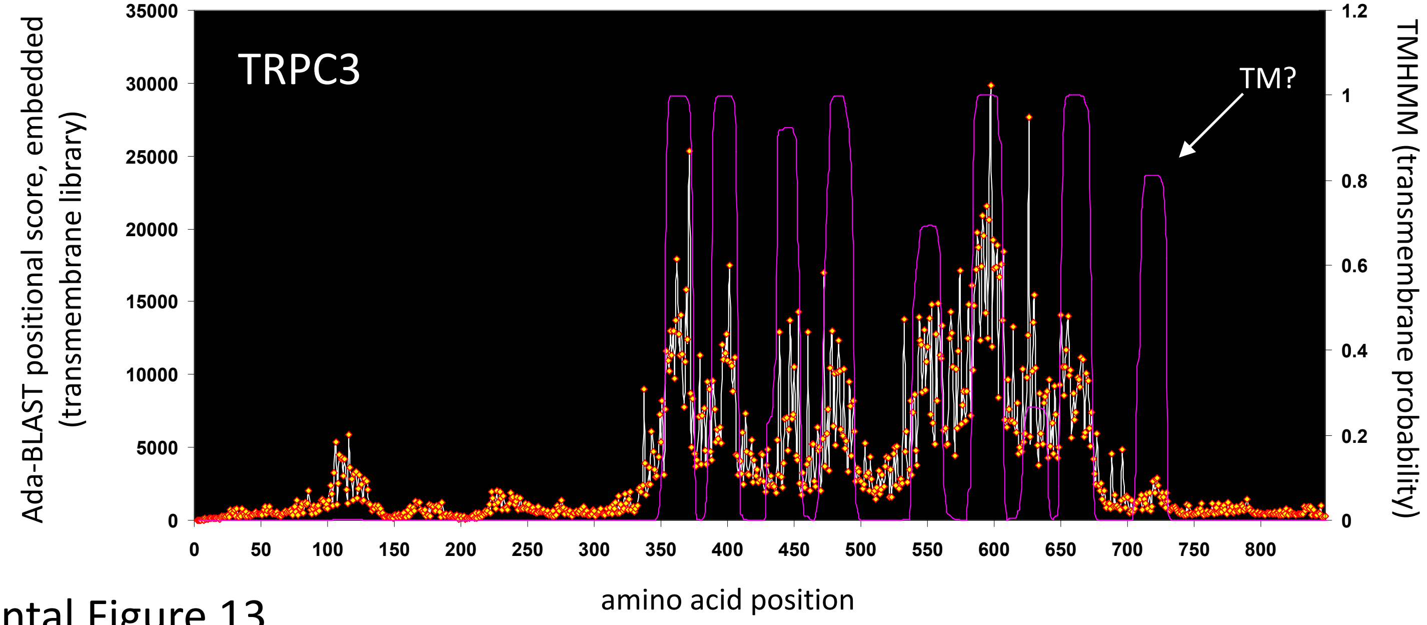


Conserved domains on [gi|4507687|ref|NP_003296.1|]

short transient receptor potential channel 3 isoform b [Homo sapiens]

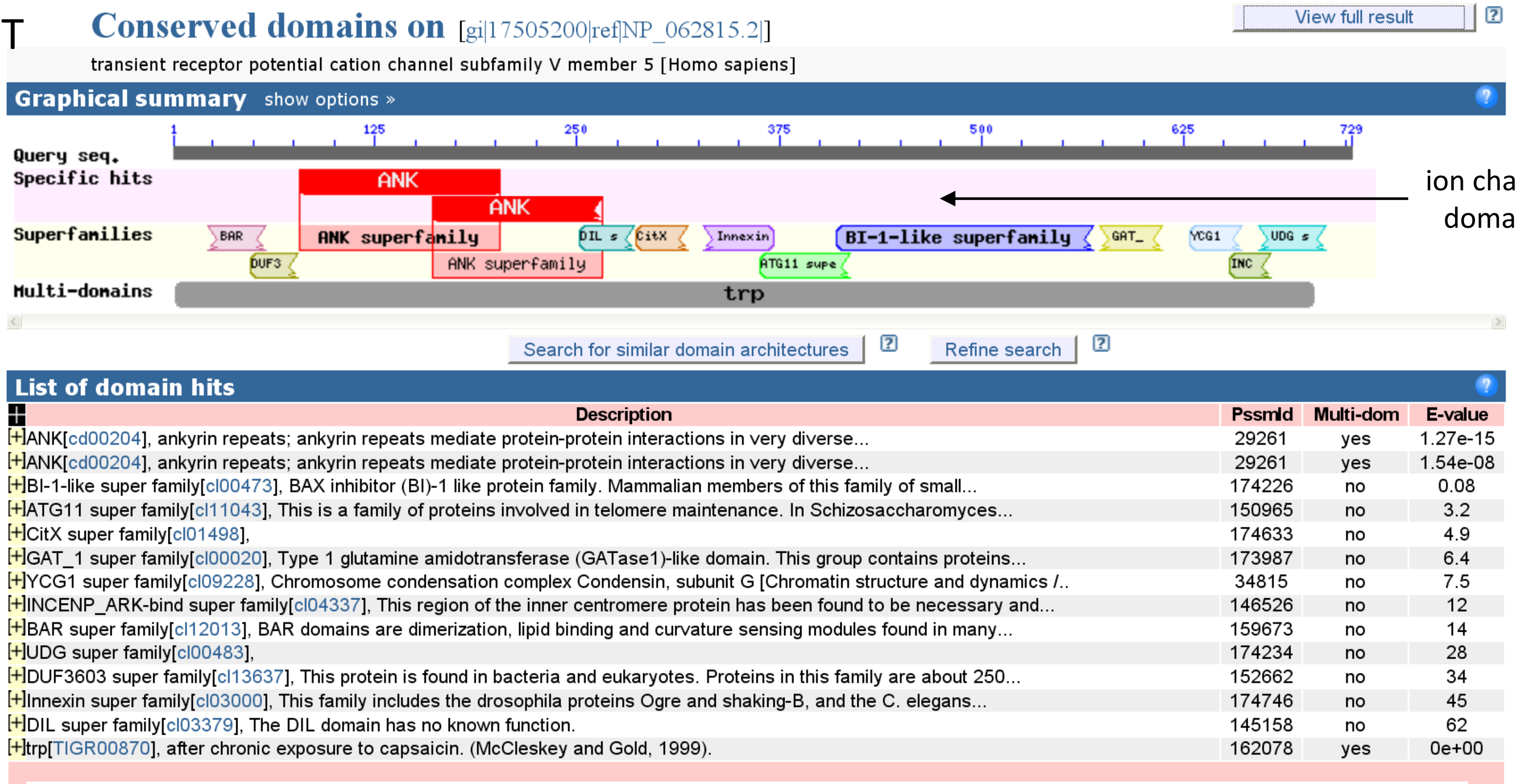
List of domain hits

Description	PssmId	Multi-dom	E-value
[+]ANK[cd00204], ankyrin repeats; ankyrin repeats mediate protein-protein interactions in very diverse...	29261	yes	1.07e-10
[+]TRP_2 super family[cl07091], This domain is found in the transient receptor ion channel (Trp) family of proteins....	149414	no	3.93e-27
[+]PKD_channel super family[cl10691], This family contains the cation channel region of PKD1 and PKD2 proteins.	116626	no	7.63e-14
[+]PBC super family[cl04264], The PBC domain is a member of the TALE (three-amino-acid loop extension) superclass of...	146435	no	1.6
[+]ArsB_NhaD_permease super family[cl09110], Anion permease ArsB/NhaD. These permeases have been shown to translocate sodium,...	175004	no	3.7
[+]Cytochrom_B562 super family[cl01546], This family contains the bacterial cytochrome b562. This forms a four-helix bundle that...	174645	no	4.2
[+]NOB1_Zn_bind super family[cl10704], This domain corresponds to a zinc ribbon and is found on the RNA binding protein NOB1.	149739	no	7.2
[+]COG5381 super family[cl02225], Uncharacterized protein conserved in bacteria [Function unknown]	34944	no	11
[+]NR_DBD_like super family[cl02596], DNA-binding domain of nuclear receptors is composed of two C4-type zinc fingers. Each...	155002	no	22
[+]Peptidase_M14_like super family[cl11393], The M14 family of metallocarboxypeptidases (MCPs) are zinc-binding carboxypeptidases (...	175235	no	23
[+]Bac_rhamnosid super family[cl01801], This family consists of bacterial rhamnosidase A and B enzymes. L-Rhamnose is abundant...	174672	no	27
[+]TT_ORF2 super family[cl03800], TT virus (TTV), isolated initially from a Japanese patient with hepatitis of unknown...	145881	no	29
[+]Peptidase_S48 super family[cl11616], Peptidase family S48.	159582	no	49
[+]trp[TIGR00870], after chronic exposure to capsaicin. (McCleskey and Gold, 1999).	162078	yes	0e+00



Supplemental Figure 13

rps-BLAST



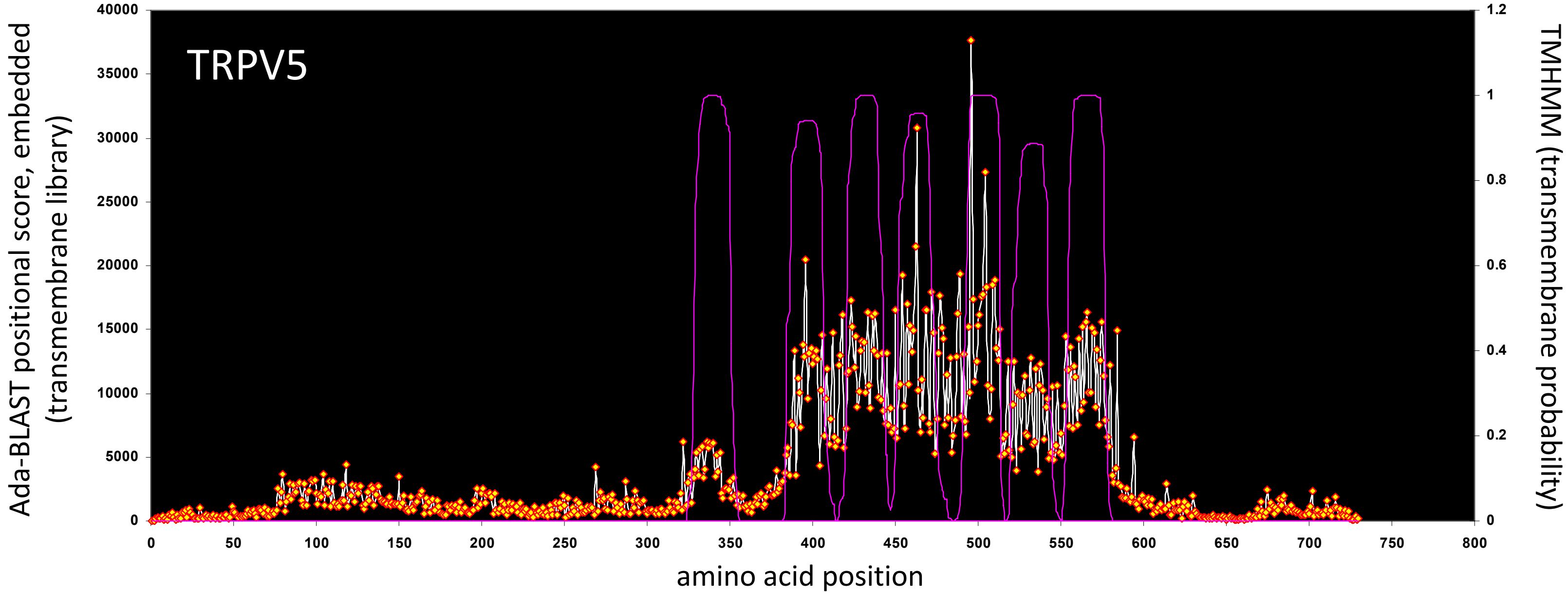
Conserved domains on [gi|17505200|ref|NP_062815.2|]

transient receptor potential cation channel subfamily V member 5 [Homo sapiens]

List of domain hits

Description	PssmId	Multi-dom	E-value
[+]ANK[cd00204], ankyrin repeats; ankyrin repeats mediate protein-protein interactions in very diverse...	29261	yes	1.27e-15
[+]ANK[cd00204], ankyrin repeats; ankyrin repeats mediate protein-protein interactions in very diverse...	29261	yes	1.54e-08
[+]BI-1-like super family[cl00473], BAX inhibitor (BI)-1 like protein family. Mammalian members of this family of small...	174226	no	0.08
[+]ATG11 super family[cl11043], This is a family of proteins involved in telomere maintenance. In Schizosaccharomyces...	150965	no	3.2
[+]CitX super family[cl01498],	174633	no	4.9
[+]GAT_1 super family[cl00020], Type 1 glutamine amidotransferase (GATase1)-like domain. This group contains proteins...	173987	no	6.4
[+]YCG1 super family[cl09228], Chromosome condensation complex Condensin, subunit G [Chromatin structure and dynamics /..	34815	no	7.5
[+]INCENP_ARK-bind super family[cl04337], This region of the inner centromere protein has been found to be necessary and...	146526	no	12
[+]BAR super family[cl12013], BAR domains are dimerization, lipid binding and curvature sensing modules found in many...	159673	no	14
[+]UDG super family[cl00483],	174234	no	28
[+]DUF3603 super family[cl13637], This protein is found in bacteria and eukaryotes. Proteins in this family are about 250...	152662	no	34
[+]Innexin super family[cl03000], This family includes the drosophila proteins Ogre and shaking-B, and the C. elegans...	174746	no	45
[+]DIL super family[cl03379], The DIL domain has no known function.	145158	no	62
[+]trp[TIGR00870], after chronic exposure to capsaicin. (McCleskey and Gold, 1999).	162078	yes	0e+00

ion channel domain?


TRPV5
Ada-BLAST positional score, embedded (transmembrane library)
TMHMM (transmembrane probability)
amino acid position