Kirthi Reddy
Biochemistry 218 Final Project

Analysis of *Caenorhabditis elegans* protein T09E8.2

Introduction

Meiosis is the cell division process by which sexually reproducing diploid organisms produce gametes containing a haploid number of chromosomes. Errors in meiotic chromosome segregation can lead to gametes containing an abnormal number of chromosomes, affecting development and viability (reviewed in Roeder, 1997). The nematode *C. elegans* provides an excellent system for studying meiosis, as this organism has many robust genetic, molecular, and cytological tools that can be used for the study of this process (Zetka and Rose, 1995). For instance, the complete genome sequence of *C. elegans* has permitted large-scale microarray expression analysis; this technology has been used to identify genes that are germline-enriched and therefore candidate meiosis genes (Reinke et al, 2000). RNA interference (RNAi) is also well-established in *C. elegans*; this powerful technique allows the rapid generation of functional knockouts of genes (Fire et al, 1998). Members of the Villeneuve lab have screened a subset of the germline-enriched genes using RNAi to find molecules that are components of the meiotic machinery in *C. elegans*.

One interesting gene identified with this screen is T09E8.2, which resides on chromosome V and encodes a predicted 954 amino acid protein. Loss of function of this gene (assayed by RNAi) results in a defect in meiotic chromosome segregation at the first meiotic division. Also, three recessive mutant alleles of this gene have been identified, each of which appears to confer a strong loss of function phenotype.

In the following sections, the sequence of the T09E8.2 protein is analyzed using several different methods. Potential motifs and related proteins are discussed, as well as possible structural configurations of this protein.

Methods

Sequence retrieval:

NCBI Entrez (http://www.ncbi.nlm.nih.gov/Entrez/)

Motif analysis:

PROSITE - ScanProsite at Expasy (http://expasy.cbr.nrc.ca/tools/scnpsit1.html)
BLOCKS (http://www.blocks.fhcrc.org/blocks/blocks_search.html)
eMOTIF (http://motif.stanford.edu/emotif/emotif-search.html)
eMATRIX (http://motif.stanford.edu/ematrix/ematrix-search.html)
eBLOCKS (http://eblocks.stanford.edu/eblocks/seqsearch.html)

Search for related proteins:

Smith-Waterman (http://decypher2.stanford.edu/algo-sw/SW_aa.html-ssi)
Gapped-BLAST (http://www.ncbi.nlm.nih.gov/blast/blast.cgi)
Human Genome (http://www.ensembl.org/)
PropSearch (http://www.embl-heidelberg.de/prs.html)

Structural analysis:

Paircoil (http://nightingale.lcs.mit.edu/cgi-bin/score)
GOR IV (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html)
Helix-turn-helix prediction (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_hth.html)

Note: full results for these searches can be found in the Appendix.

Motif Analysis of T09E8.2:

ScanProsite:

A search done excluding patterns with a high probability of occurrence found two motifs.

· PDOC00207 PS00234 GAS_VESICLE_A_1 Gas vesicles protein GVPa signature 1
936-950 VEEIVEQVVDDGVVV

This 15 aa-long motif is found in proteins involved with gas vesicles in micro-organisms. No sequences that do not belong to the gas vesicles protein family were detected in Swiss-Prot by this motif. Also, there are two signature patterns for the gas vesicles protein family, and the T09E8.2 protein only contains one of these motifs. When this motif was used to scan both Swiss-Prot and Trembl, 32 protein sequences were found (15 in Trembl); 31 of these 32 were bacterial gas vesicle proteins. The 32nd protein was the T09E8.2 protein. This result seems rather strange and unlikely, as the T09E8.2 protein is from the metazoan *C. elegans*.

· PDOC00018 PS00018 EF_HAND EF-hand calcium-binding domain
829-841 DDNGSGRLDYDDV

A large number of known calcium-binding proteins have been shown to contain this 13 aa-long EF-hand domain; however, another 52 proteins in Swiss-Prot have been found that contain this motif but are probably not calcium-binding proteins. It is therefore difficult to say with certainty that this domain is functioning in the T09E8.2 protein. There is no biological reason to believe that the T09E8.2 protein (which functions during meiosis) contains a calcium-binding domain, although it certainly is possible.

Another six motifs were found when patterns with a high probability of occurrence were included in the search. The common patterns found include glycosylation and phosphorylation sites, as well as an N-myristylation site. Searches done with several unrelated proteins all pulled up a similar assortment of these common sites. The common patterns were all fairly short, ranging from 3 to 8 amino acids in length. All of these patterns were found multiple times in the T09E8.2 protein – in fact, the 3 aa motif was found 19 times. As these patterns have a high probability of occurrence, further evidence (such as biological confirmation) that these are functional motifs in this protein is needed before believing these results.

BLOCKS:

The BLOCKS+ database was searched, with a cutoff expect value of 5. The following hits were obtained:

Family	Strand	Blocks	E-value
PR01110 C-C chemokine receptor type 5 signa	1	1 of 5	0.016
IPB001705 Ribosomal protein L33	1	1 of 2	0.076
IPB002041 GTP-binding nuclear protein Ran fam	1	1 of 4	0.46
IPB002038 Osteopontin	1	1 of 5	1.3
PF00951 Arterivirus GL envelope glycoprotei	1	1 of 4	1.3
IPB001990 Granins (chromogranin or secretogra	1	1 of 4	1.3
PR00708 Alpha-1-acid glycoprotein signature	1	1 of 5	1.9
IPB002048 EF-hand family	1	1 of 1	3
PF01896 DNA primase small subunit	1	1 of 4	4.6
PF01846 FF domain	1	1 of 4	4.6
PF01795 Domain of unknown function	1	1 of 5	4.9

The most statistically significant result found was a C-C chemokine receptor type 5 signature (combined E-value of 0.016). As the T09E8.2 protein matches only one of the five blocks, it is possible that the function of the T09E8.2 protein is entirely different from this receptor. As for the biological significance, this receptor domain is most often found in blood cells in mammals. As *C. elegans* does not have blood, it seems unlikely that there is any useful functional information to be obtained from this hit. However, as there is no biological evidence to support or contradict this finding, it remains possible that the T09E8.2 protein has a similar function as the C-C chemokine receptor.

The other matches found with BLOCKS were not statistically significant (using 95% confidence) and provide no obvious clues as to the query protein's function. The EF-hand calcium binding domain that was found with Prosite was number 8 on this list, with a combined E-value of 3 (1 of 1 blocks). This is not a statistically significant result, suggesting that the T09E8.2 protein is not likely a calcium-binding protein. The puzzling result found with Prosite, the gas vesicles protein signature, was not found with BLOCKs, correlating with the idea that this is a result to be ignored.

eMOTIFS:

An eMOTIF search was done with the T09E8.2 protein. At an expectation of 10 or less, 14 hits were obtained (see Appendix for results). The first hit on the list, C2HC-type zinc-finger signature I, has an expect value of 8.36e-01, which is not very significant. Therefore, no statistically significant hits were found using eMOTIF. As for biological significance: even though the zinc-finger domain was not a statistically significant hit, it could have biological relevance. This domain is found in proteins that bind RNA or DNA, and it is certainly possible that the T09E8.2 protein has this function. This would need to be confirmed experimentally. The other hits (which all had expect values greater than 2) did not seem to have a common biological function, and therefore did not provide any clues as to the function of the T09E8.2 function.

eMATRIX:

The T09E8.2 protein sequence was used to do an eMATRIX search with a desired significance threshold of 10e-8; two hits were found. The first, a C-C chemokine receptor domain, had a probability of 7.28e-9. To determine the expect value for this hit, this probability was multiplied times the size of the swiss-prot database (33,206,838 aa). This gives an expect of 0.24 , which is not highly significant. This motif was also found with the BLOCKS search, and the biological significance of this hit was discussed above.

The second hit found, GTP-binding nuclear protein Ran family, had a probability of 9.88e-9. The expect value calculated for this is 0.33 which, again, is not statistically significant. This domain has been found in lower organisms such as yeast as well as mammals. The T09E8.2 protein could be functioning as a GTPase, although there is no experimental evidence that suggests this.

eBLOCKS:

The T09E8.2 protein sequence was also used to do an eBLOCKS search with either eMOTIFs or eMATRICES made from the eBLOCKs.

When eMOTIFS was used as the search method, thirteen eBLOCKS were found. The only statistically significant hit is the bone marrow stromal antigen 2 eBLOCK, which does not seem to be biologically significant. Nearly all of the eBLOCKS found were pulled up with the C-terminus end of the T09E8.2 protein. Also, most of the eBLOCKS found contain a high proportion of glutamic acid residues (as does the C-terminal end of the T09E8.2 protein).

Many of the proteins included in the eBLOCKs pulled up with this search are myosins from various organisms. For example, one potentially interesting eBLOCK found is P06180G1B2 (histone-binding protein N1/N2); this could be biologically significant, as changes in the histone composition of chromosomes have been shown to be important in meiosis (Riggs, 1997). There are 62 proteins included in this eBLOCK, and the majority of these are myosin proteins from organisms such as human, rat, chick, and fruitfly. The predominance of myosin among these hits could be because myosin is a

very long protein, increasing the probability that a similarity to myosin could be found by chance.

When eMATRIX was used as the search method, 227 eBLOCKs were found. Again, the majority of these eBLOCKs match at the glutamic acid-rich C-terminus end of the T09E8.2 protein. Several potentially interesting hits were found, such as the mouse centromere protein B, a protein thought to be involved in kinetochore assembly. The DNA repair proteins Rad26, Rad21, and XRC1 were also found, which are intriguing as several DNA repair proteins have been shown to be involved in meiotic recombination (Roeder, 1997). It is quite possible that the T09E8.2 protein is functioning as a DNA repair protein or a centromere protein; however, there is of yet no experimental data to support these ideas.

Search for related proteins:

Smith-Waterman:

The T09E8.2 protein sequence was used to query the Swiss-Prot database using the Smith-Waterman algorithm (on the Decypher machine) with the default choices. No statistically significant hits were found. All of the alignments produced were very short and had many large gaps. As for biological significance, there was no common biochemical theme found among these hits.

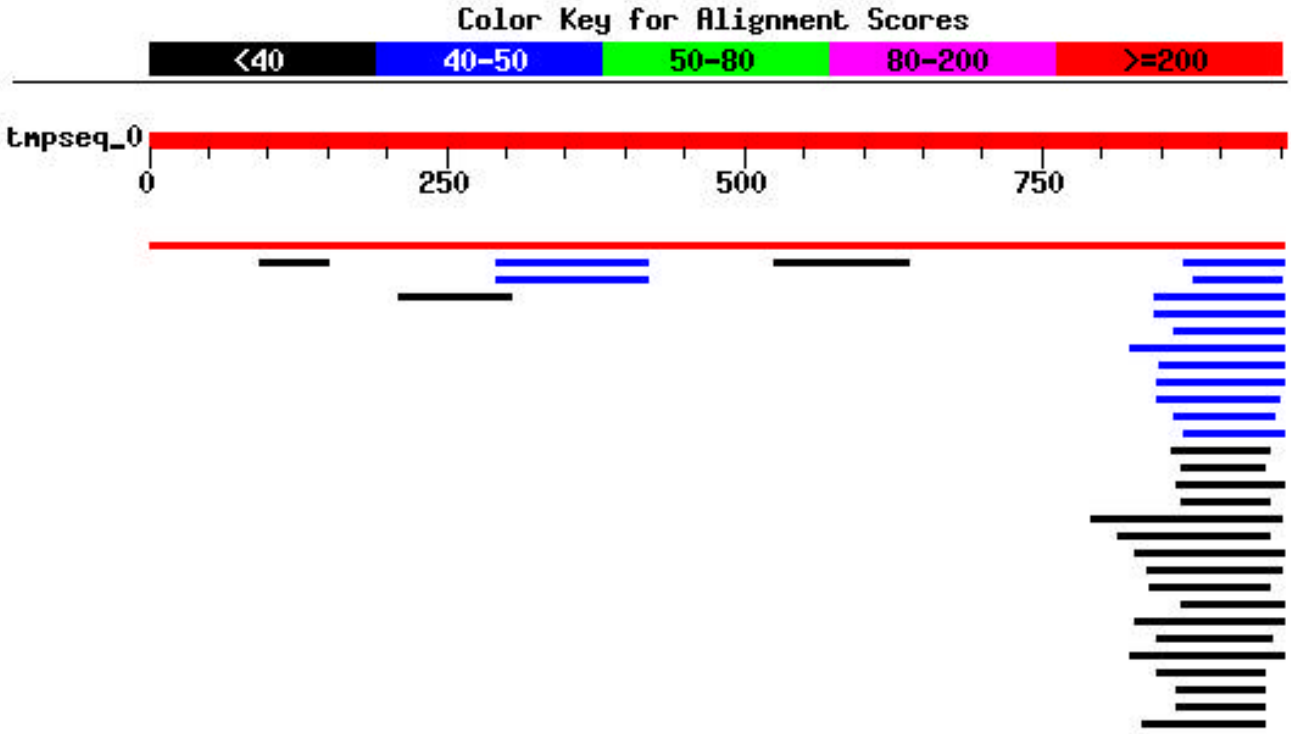
Note: this query did not pull up the T09E8.2 protein itself. A search at Expasy shows that this is because the T09E8.2 sequence is in TrEMBL, not Swiss-Prot.

Gapped-BLAST:

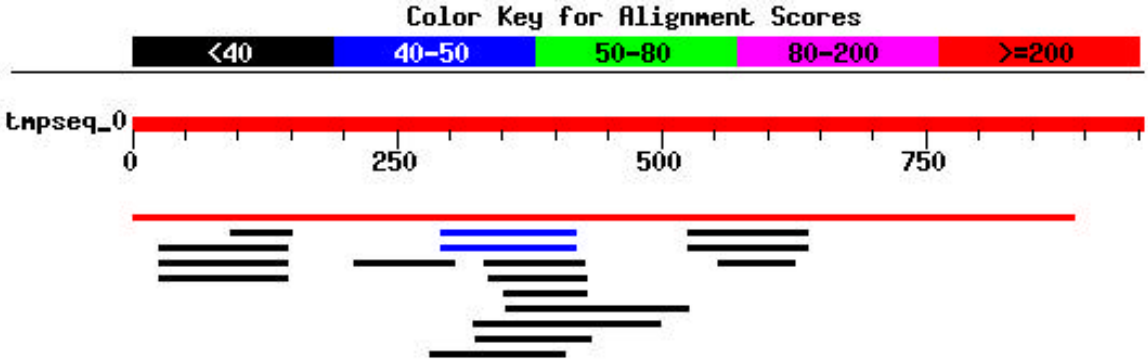
The T09E8.2 sequence was then used in an unfiltered Gapped-BLAST search of the non-restricted databases. The T09E8.2 sequence itself was found with this query, as expected. The most statistically significant match found (E value of 0.007) is the erythrocyte membrane-associated giant protein antigen of the malaria parasite. The alignment of this protein with T09E8.2 is not very long and has many gaps, which makes the significance of this alignment rather unconvincing. This malaria protein is rich in glutamic acid residues, as is the C-terminus of the T09E8.2 protein. It is unlikely that these proteins have similar biological functions. Most of the other hits (all of which were not statistically significant) found with this query also align at the C-terminal region of the T09E8.2 protein (see below) and are rich in glutamic acid residues. As with the Smith-Waterman search, there is no common biological theme found among these hits.

A filtered Gapped-BLAST search was also done with the T09E8.2 sequence. Eighteen hits were found, none of which were statistically significant (except for the T09E8.2 protein itself). None of the hits aligned at the C-terminal region of the query protein; this glutamic acid-rich region was likely filtered out for this search. The most significant hit (E value of 0.079) was hypothetical protein F07A11.6b from *Caenorhabditis elegans*. This predicted protein is of unknown function and could be biologically relevant. A first-pass analysis of this protein's function by using RNAi to create a functional knockout could quickly confirm or disprove this hypothesis.

Gapped BLAST search (unfiltered):



Gapped BLAST search (filtered):



Human Genome:

The human genome database was searched with the T09E8.2 protein using the Ensembl genome server. As with the Gapped-BLAST search described above, the majority of the hits found with the unfiltered search aligned at the C-terminal region of the query protein. None of the alignments produced were of great length, and all had many large gaps. The filtered option was not working, so this search could not be performed.

PropSearch:

A search using PropSearch was also done using the T09E8.2 protein sequence as the query. This search engine looks for functional or structural homologs using properties such as molecular weight, amino acid composition, and hydrophobicity. Seven proteins were found that have a 94% or better reliability (as calculated by PropSearch). These hits include the vav proto-oncogene, a tyrosine kinase, and a RNA-directed RNA polymerase subunit. The vav protein is thought to be only expressed in hematopoietic cells, and is therefore unlikely to be relevant. The RNA-directed RNA polymerase is an intriguing match; even though it is from the influenza virus, it could be that the T09E8.2 protein has a nucleotide-binding function.

Structural Analysis:

PAIRCOIL

The Paircoil program predicts regions of coiled-coil domains in amino acid sequences. As mentioned above, many myosin proteins (which are known to contain coiled-coil domains) were pulled up with the eBLOCKS search done with the T09E8.2 protein. Several of the other proteins pulled up with the eBLOCKS search also are likely to contain coiled-coils (as analyzed with the Paircoil program), such as the neurofilament triplet M protein, the bone marrow stromal antigen 2 protein, as well as the DNA repair proteins and the centromere proteins found. In addition, the regions of these proteins that were used to create the eBLOCKS found in this query are regions that contain putative coiled-coil domains. Therefore, it seemed possible that the T09E8.2 protein also contained a coiled-coil domain.

The T09E8.2 protein sequence was analyzed by the Paircoil program. No coiled-coil domains were found with a high probability; however, the results showed that the glutamic acid-rich C-terminal region of this protein has a low probability (0.3) of containing a coiled-coil domain. Therefore it is possible, though unlikely, that the T09E8.2 protein contains a coiled-coil domain.

GOR IV

The GOR IV program was used to predict the secondary structure of the T09E8.2 protein, with particular attention paid to the results for the C-terminus end of this protein (as this is the region that most often pulled up hits in the motif searches). The C-terminus end was predicted to contain mostly alpha helices, which fits with the possibility that this region could contain a coiled-coil domian.

HELIX-TURN-HELIX MOTIF PREDICTION

The helix-turn-helix prediction software at the Network Protein Sequence analysis site was used to search the T09E8.2 protein sequence for this DNA-binding motif. No significant helix-turn-helix motifs were found in this protein, although a sequence in the central region of the protein had a score of 0.52. Proteins that contain the helix-turn-helix motif generally bind DNA in a sequence-specific manner. It is not known how the T09E8.2 protein is affecting chromosome behavior during meiosis, but it is possible that it binds DNA in a non-specific manner. The absence of the helix-turn-helix motif correlates with this hypothesis.

Discussion

Experimental analysis of the T09E8.2 protein has previously shown that it functions in meiosis and is required for proper meiotic chromosome segregation. To elicit further information about the function of this protein, various sequence analyses were performed.

Several motif searches were done using this protein, but very few statistically significant hits were found; also, no obvious biological relevance was seen amongst these significant hits. The statistically insignificant results found with these searches contained some potentially interesting proteins, such as DNA repair proteins and centromere proteins. Experimental results are needed to make any correlation between the function of these proteins and the T09E8.2 protein.

Searches to find related proteins were also done using the T09E8.2 protein sequence as a query. Again, very few significant results were found, and none of these results were of clear biological relevance. In addition, no homolog to the T09E8.2 protein was found in the complete human genome sequence (or in any other databases). This indicates that this protein is novel and non-conserved. This is somewhat surprising, as meiosis itself is a highly conserved process, and several key meiotic proteins are known to be strongly conserved. However, *C. elegans* chromosomes are rather unique in that they are holocentric; perhaps this property of the chromosomes requires a specialization of the meiotic machinery that involves the T09E8.2 protein.

Interestingly, many of the putative related proteins found with the motif searches are likely to contain coiled-coil regions. An analysis of the T09E8.2 protein was done using the Paircoil program to determine if this protein also contains a putative coiled-coil domain. The results indicated that there is a possibility that there is a coiled-coil domain at the C-terminus end of this protein. This result is intriguing, as many proteins involved in meiosis that affect chromosome behavior have been shown to contain coiled-coil domains (as in Sym et al, 1993).

One of the strong loss-of-function alleles of this protein contains an early stop codon that truncates the last 93 amino acids, which is the glutamic acid-rich region of this protein. It therefore could be that this C-terminus region contains an important functional domain; of course, it could also be that loss of this region of the protein alters the tertiary structure of the protein in a way that makes it unable to perform its function. Several other proteins that contain glutamic acid-rich regions were identified with the searches presented here. No common biological function could be assigned to the proteins containing these stretches of glutamic acid residues; however, several of these low complexity regions fall within predicted coiled-coil domains, an interesting result which is discussed above.

In conclusion, the T09E8.2 protein appears to be novel and non-conserved. None of the searches described here have yielded definitive results as to the function of the T09E8.2 protein, though several interesting possibilities have been discovered, some of which may lead to new lines of experimental investigation. Biological evidence is needed before these results can be confirmed or disproven.

References

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 1998 Feb 19;391(6669):806-11.

Reinke V, Smith HE, Nance J, Wang J, Van Doren C, Begley R, Jones SJ, Davis EB, Scherer S, Ward S, Kim SK. A global profile of germline gene expression in C. elegans. *Mol Cell* 2000 Sep;6(3):605-16.

Riggs CD. Meiotin-1: the meiosis readiness factor? *Bioessays* 1997 Oct;19(10):925-31.

Roeder GS. Meiotic chromosomes: it takes two to tango. *Genes Dev* 1997 Oct 15;11(20):2600-21.

Sym M, Engebrecht JA, Roeder GS. ZIP1 is a synaptonemal complex protein required for meiotic chromosome synapsis. *Cell* 1993 Feb 12;72(3):365-78.

Zetka M, Rose A. The genetics of meiosis in Caenorhabditis elegans. *Trends Genet* 1995 Jan;11(1):27-31.

Appendix

Protein Sequence

>gi|7507544|pir||T24748 hypothetical protein T09E8.2 - Caenorhabditis elegans
MAKRPSTSKNPAEDDELIIDECRRKLDDDESNDVGAYFEDDNQQGTSYRTPFMPSHLGLI
QDQSPDLTMKDLFNRKRSRPQSTGFIDHSQARYYDRYKQQEYTRRNDFMKYCHVCRKK
TVGKLRVLPGDLRMRKVWILRSNLDEERSAELWLREMACEGSHGGEFCEAHFPAGSQH
MKGNQLIPLDVRPEEGRHSVVEDIQFDFDAIHLQCVFCGRDGPLATMLPFIRNRAKRLRW
IEQLASGNETYKNRLLHALRGGVTQFLCDYHISDSSFEINGFGEWRLLKNALPDPRLVAS
DKRGERKYLVDKCRDEMFWERAMWKSADLLSQNGTADDEDIIVGDDIRQFLTNSVPSE
AEQVLHPILQIKQNVEKKLNEADLQPGTSGTSGESNATADNLKIGISEEGVEYSDDSDEE
NELIEKMGDIPYSKRLCQVCSAVEPIGNEFPNNFPYKFSIRTWPFDECRHKKWLEIMDWP
PEFEESMKTLWQKRKTEGALSDSYHFCPINVCQSHLDFRQLPERMEEWHQTFCLLCDTC
MSDKSFLVGIPHNFETRTKWANSLFPVKEGNKFQLKKISWLRRRFLRSKPTRYRICVYHF
NRKAFQVNQDDKIVLDAEALPLPIDSDDFDLTPRGPNSVCKCVLCDDWKKVEDMVPFR
APHSEAERSFLIDIVIHSDKMTIKKALASLAKTNRPALICNIHFPDGIDPFSIIAERRLMYGV
QSECVLCAHANDCTAMIPFPGPDDEKLRTKWINSMCREPWIYRYLSTRLEKPGRHYLCA
SHFNRNSLRYHAGLGLWRRAAACPVLACTTDEERQEVWDLSKSQPLYHPLILETFDDNG
SGRLDYDDVLNFLGTERMRQIETELNFHGRNDSLLRRENRKKNTGRSFYADMPNIYEHG
ELQEVEDQMEGEEEEVVQEECIIYEEEHHLDEAEELEVIEEEEVVEEIVEQVVDDGVVVG
EEM

ScanProsite - Protein against PROSITE

[1] PDOC00001 PS00001 ASN_GLYCOSYLATION
N-glycosylation site

Number of matches: 5
1 244-247 NETY
2 329-332 NGTA
3 390-393 NATA
4 831-834 NGSG
5 863-866 NDSL

[2] PDOC00004 PS00004 CAMP_PHOSPHO_SITE
cAMP- and cGMP-dependent protein kinase phosphorylation site

Number of matches: 6
1 3-6 KRPS
2 75-78 RKRS
3 116-119 RKKT
4 487-490 KRKT
5 569-572 KKIS

6 873-876 KKNT

[3] PDOC00005 PS00005 PKC_PHOSPHO_SITE
Protein kinase C phosphorylation site

Number of matches: 19
1 7-9 TSK
2 47-49 SYR
3 68-70 TMK
4 103-105 TRR
5 246-248 TYK
6 296-298 SDK
7 535-537 SDK
8 669-671 SDK
9 427-429 SKR
10 453-455 SIR
11 480-482 SMK
12 296-298 SDK
13 535-537 SDK
14 669-671 SDK
15 625-627 TPR
16 296-298 SDK
17 535-537 SDK
18 669-671 SDK
19 673-675 TIK

[4] PDOC00006 PS00006 CK2_PHOSPHO_SITE
Casein kinase II phosphorylation site

Number of matches: 16
1 68-71 TMKD
2 140-143 SNLD
3 195-198 SVVE
4 242-245 SGNE
5 271-274 SSFE
6 331-334 TADD
7 353-356 SEAE
8 655-658 SEAE
9 385-388 TSGE
10 411-414 SDEE
11 435-438 SAVE
12 508-511 SHLD
13 353-356 SEAE
14 655-658 SEAE
15 761-764 TRLE
16 802-805 TTDE

[5] PDOC00007 PS00007 TYR_PHOSPHO_SITE
Tyrosine kinase phosphorylation site

Number of matches: 3
1 104-111 RRNDFMKY
2 298-304 KRGERKY
3 420-426 KMGDIPY

[6] PDOC00008 PS00008 MYRISTYL
N-myristoylation site

Number of matches: 6
1 159-164 GSHGGE
2 162-167 GGEFCE
3 381-386 GTSGTS
4 404-409 GVEYSD
5 542-547 GIPHNF
6 947-952 GVVVGE

[7] PDOC00018 PS00018 EF_HAND
EF-hand calcium-binding domain

829-841 DDNGSGRLDYDDV

[8] PDOC00207 PS00234 GAS_VESICLE_A_1
Gas vesicles protein GVPa signature 1

936-950 VEEIVEQVVDDGVVV

BLOCKS Hits

Query=Unknown Unknown
Size=954 Amino Acids
Blocks Searched=11117
Alignments Done= 10877671
Cutoff combined expected value for hits= 5
Cutoff block expected value for repeats/other= 5
==========

Family	Strand	Blocks	Combined E-value
PR01110 C-C chemokine receptor type 5 signa	1	1 of 5	0.016
IPB001705 Ribosomal protein L33	1	1 of 2	0.076
IPB002041 GTP-binding nuclear protein Ran fam	1	1 of 4	0.46
IPB002038 Osteopontin	1	1 of 5	1.3
PF00951 Arterivirus GL envelope glycoprotei	1	1 of 4	1.3
IPB001990 Granins (chromogranin or secretogra	1	1 of 4	1.3
PR00708 Alpha-1-acid glycoprotein signature	1	1 of 5	1.9
IPB002048 EF-hand family	1	1 of 1	3
PF01896 DNA primase small subunit	1	1 of 4	4.6
PF01846 FF domain	1	1 of 4	4.6
PF01795 Domain of unknown function	1	1 of 5	4.9

eMOTIF

Rank	Expectation		Description
1.	8.36e-01	43.4%	PR00939A C2HC-type zinc-finger c..cg[ekqr].g 211-CVFCGRDG-218
2.	2.25e+00	10.6%	PF01709D Domain of unknown function e[de].[ilv][ilmv][de]..[ilv]e 932-EEEVVEEIVE-941
3.	3.25e+00	3.9%	IPB001485B Phosphoglucomutase and phosphomannomutase family [iv][filmv][iv]g.d.r 337-IIVGDDIR-344
4.	4.56e+00	1%	PR00939A C2HC-type zinc-finger signature I c..cg..g 211-CVFCGRDG-218
5.	4.66e+00	0.9%	PF01844B HNH endonuclease n....c..c...[kr] 630-NSVCKCVLCDDWK-642
6.	4.99e+00	0.7%	IPB001707A Chloramphenicol acetyltransferase [fwy].r[kr]..f..[fy] 102-YTRRNDFMKY-111
7.	5.39e+00	0.5%	PR00469E Pyridine nucleotide disulphide reductase class-II signature V [iv]..c..cd 632-VCKCVLCD-639
8.	5.75e+00	0.3%	PF01873D Domain found in IF2B/IF5 [kqr].[filvy].[fly].c..c 524-QTFCLLCDTC-533
9.	6.39e+00	0.2%	PF01409C tRNA synthetases class II (F) [kr].....p[fwy]..p..[eq] 642-KKVEDMVPFRAPHSE-656
10.	6.75e+00	0.1%	IPB000183C Orn/DAP/Arg decarboxylases family 2 [ilmv].d....cp[ilmv] 494-LSDSYHFCPI-503

11. 7.49e+00 0.1% PR00403B WW domain signature II
[fy][fy].....[st].[fwy]..p
37-YFEDDNQQGTSYRTP-51

12. 8.29e+00 0% IPB000158D Cell division protein FtsZ
[ilv]d[fy].d[ilmv]...[filmv]
836-LDYDDVLNFL-845

13. 8.87e+00 0% DM00099B 4 kw A55R REDUCTASE TERMINAL
y....[kqr]tw
450-YKFSIRTW-457

14. 9.53e+00 0% IPB001444B Flagella basal body rod protein
[ilv].....[dn].i..q..[fy]
197-VEDIQFDFDAIHLQCVF-213

eMATRIX

Rank Significance Description

1. 7.28e-09 PR01110B C-C chemokine receptor type 5 signature II
761-TRLEKPGRHYLCASHFNRNSL-781

2. 9.88e-09 IPB002041D GTP-binding nuclear protein Ran family
904-EEEVVQEECIIYEEEHHLDEAEELEVIEEEEV-935

eBLOCKS

eMotif Search Result For Your Query Sequence

E-Value eBLOCK Motif

1. 1.21e-03 Q10589G2B1 eaee.[de]...[eq]ee...e
Segment 923--938 EAEELEVIEEEEVVE
BONE MARROW STROMAL ANTIGEN 2 (BST-2).

2. 4.56e-02 P03204G1B10 e.eee[de].....[ekq][de].....e
Segment 901--923 EGEEEEVVQEECIIYEEEHHLD
EBNA-6 NUCLEAR PROTEIN (EBNA-3C) (EBNA-4B).

3. 6.29e-02 Q09822G1B12 e...e.e.ee..[ekq][de][ekq]
Segment 897--911 EDQMEGEEEEVVQE
CELL DIVISION CONTROL PROTEIN 15.

4. 1.54e-01 P32448G3B8 eeee..[eq]ee
Segment 903--911 EEEEVVQE
ANTI-SILENCING PROTEIN 2.

5. 5.54e-01 Q09822G1B14 e....[ekq]ee....[ekq].ee
Segment 911--926 ECIIYEEEHHLDEAE
CELL DIVISION CONTROL PROTEIN 15.

6. 7.92e-01 Q53654G1B9 [ekq]e..v[ilv][ekq]..[ilv]e....[de]
Segment 931--946 EEEEVVEEIVEQVVD
COLLAGEN ADHESIN PRECURSOR.

7. 8.11e-01 P06180G1B2 q[de][ilv][de].....e.[ekqr]...[ekqr][de]
Segment 894--911 QEVEDQMEGEEEEVVQE
HISTONE-BINDING PROTEIN N1/N2.

8. 1.27e+00 P53692G2B1 [ekqr]l[eq].[ilmv]...[ilmv]..[ekqr].e[ekqr]..[ekq]
Segment 892--909 ELQEVEDQMEGEEEEVV
DNA REPAIR PROTEIN RAD18.

9. 1.46e+00 P23003G1B2 [ekqr][iv][ilv]..[de].....v...[kqr].n[ilv]
Segment 124--142 RVLPGDLRMRKVWILRSN
TRNA (URACIL-5-)-METHYLTRANSFERASE (EC 2.1.1.35) (TRNA(M-5-U54)-METHYLTRANSFERASE) (RUMT).

10. 1.51e+00 P06153G1B2 e....e..e..e..e...e
Segment 923--941 EAEELEVIEEEEVVEEIV
IMMUNITY REPRESSOR PROTEIN.

11. 1.51e+00 Q02228G2B8 ee....eeee
Segment 925--934 EELEVIEEE
GAS VESICLE PROTEIN C.

12. 1.64e+00 P50478G3B14 e.[ekq]eee...[ekq][ekq]
Segment 901--911 EGEEEEVVQE
AMPHIPHYSIN.

13. 2.02e+00 P22793G4B2 [dn]g.g.[ilv][dn][fy].[de]
Segment 831--840 NGSGRLDYD
TRICHOHYALIN.

DeCypher Results for: Smith-Waterman Similarity Search

RANK	SCORES	QF	TARGET LOCUS NAME	ACCESSION#	TF	TARGET	P_SCORE	DESCRIPTION
1	98.00	1	IE68_HSVSA		1	SWISSPRO	0.000110	Q01042 herpesvirus saimiri (strain 11). immediate-ea
2	91.00	1	GARP_PLAFF		1	SWISSPRO	0.001021	P13816 plasmodium falciparum (isolate fc27 / papua n
3	90.00	1	YCF2_OENPI		1	SWISSPRO	0.001403	P31568 oenothera picensis (oenothera odoarata). hypo
4	85.00	1	YCF2_OENVI		1	SWISSPRO	0.006862	P31569 oenothera villaricae. hypothetical protein (o
5	83.00	1	RNA1_YEAST		1	SWISSPRO	0.012922	P11745 saccharomyces cerevisiae (baker's yeast). ran
6	83.00	1	NIL2_HUMAN		1	SWISSPRO	0.012922	P37275 homo sapiens (human). nil-2-a zinc finger pro
7	82.00	1	PHA1_HUMAN		1	SWISSPRO	0.017717	P39687 homo sapiens (human). potent heat-stable prot
8	80.00	1	CENB_CRIGR		1	SWISSPRO	0.033202	P48988 cricetulus griseus (chinese hamster). major c
9	80.00	1	SKI3_YEAST		1	SWISSPRO	0.033202	P17883 saccharomyces cerevisiae (baker's yeast). sup
10	79.00	1	YSE2_CAEEL		1	SWISSPRO	0.045346	Q09936 caenorhabditis elegans. hypothetical 47.4 kda
11	79.00	1	RESA_PLAFF		1	SWISSPRO	0.045346	P13830 plasmodium falciparum (isolate fc27 / papua n
12	79.00	1	P2CG_HUMAN		1	SWISSPRO	0.045346	O15355 homo sapiens (human). protein phosphatase 2c
13	78.00	1	TRHY_RABIT		1	SWISSPRO	0.061789	P37709 oryctolagus cuniculus (rabbit). trichohyalin.
14	78.00	1	RESA_PLAFP		1	SWISSPRO	0.061789	Q26005 plasmodium falciparum (isolate palo alto / ug
15	78.00	1	NFM_CHICK		1	SWISSPRO	0.061789	P16053 gallus gallus (chicken). neurofilament triple
16	78.00	1	TRT_DROME		1	SWISSPRO	0.061789	P19351 drosophila melanogaster (fruit fly). troponin
17	76.00	1	YM72_YEAST		1	SWISSPRO	0.113501	Q05021 saccharomyces cerevisiae (baker's yeast). hyp
18	76.00	1	INVO_CANFA		1	SWISSPRO	0.113501	P18174 canis familiaris (dog). involucrin. 2/1996
19	76.00	1	FKB3_YEAST		1	SWISSPRO	0.113501	P38911 saccharomyces cerevisiae (baker's yeast). fk5
20	76.00	1	NUCL_XENLA		1	SWISSPRO	0.113501	P20397 xenopus laevis (african clawed frog). nucleol
21	75.00	1	CEC1_CAEEL		1	SWISSPRO	0.152597	P34618 caenorhabditis elegans. cec-1 protein. 12/199
22	75.00	1	NFL_BOVIN		1	SWISSPRO	0.152597	P02548 bos taurus (bovine). neurofilament triplet l
23	75.00	1	NFL_HUMAN		1	SWISSPRO	0.152597	P07196 homo sapiens (human). neurofilament triplet l

24 75.00 1 RYNR_RABIT 1 SWISSPRO 0.152597 P11716
oryctolagus cuniculus (rabbit). ryanodine rec
25 75.00 1 S230_PLAFO 1 SWISSPRO 0.152597 Q08372
plasmodium falciparum (isolate nf54), and pla
26 74.00 1 CENB_MOUSE 1 SWISSPRO 0.203531 P27790 mus
musculus (mouse). major centromere autoan
27 74.00 1 RYNR_PIG 1 SWISSPRO 0.203531 P16960 sus scrofa
(pig). ryanodine receptor, skeleta
28 73.00 1 CENB_SHEEP 1 SWISSPRO 0.268577 P49451 ovis aries
(sheep). major centromere autoanti
29 73.00 1 MSL1_DROME 1 SWISSPRO 0.268577 P50535
drosophila melanogaster (fruit fly). male-spe
30 72.00 1 DPB3_YEAST 1 SWISSPRO 0.349396 P27344
saccharomyces cerevisiae (baker's yeast). dna

Human Genome search

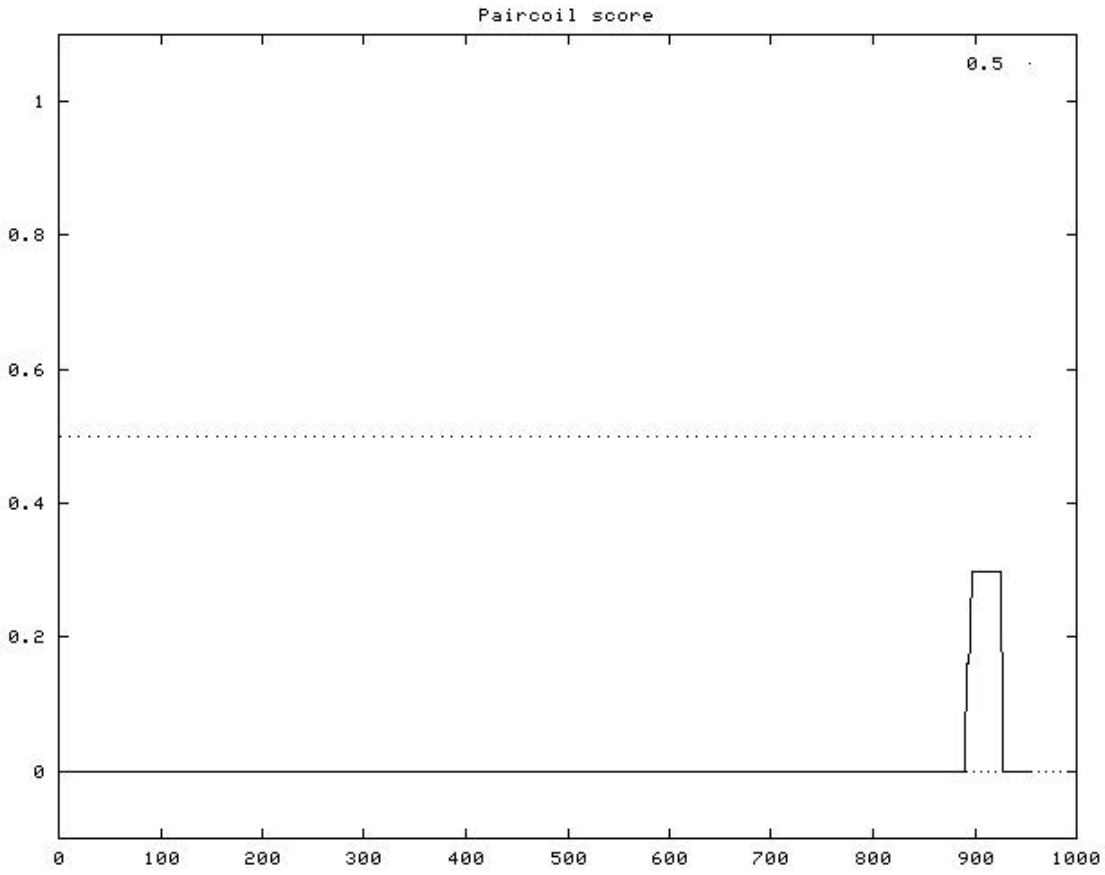
Database: ensembl.genscan.fa
191,925 sequences; 31,280,582 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Sum Probability P(N)	N
AC010873.00015.	99	6.1e-06	2
AC026017.00016.	111	1.6e-05	1
AL355861.00011.	108	3.3e-05	1
AL133345.00001.	100	0.00024	1
AC024057.00001.	100	0.00024	1
AC007491.00026.	97	0.00050	1
AL357372.00022.	96	0.00064	1
AC078873.00007.	118	0.00070	1
AC079141.00012.	95	0.00081	1
AC021183.00020.	95	0.00081	1
AC068203.00017.	95	0.00081	1
AC024034.00027.	94	0.0010	1
AC010184.00003.1	94	0.0010	1
AC025108.00015.	94	0.0010	1
AC073377.00026.	94	0.0010	1
AC009785.00004.	92	0.0017	1
AL139235.00007.	88	0.0018	2
AC068110.00024.	91	0.0022	1
AC026992.00030.	91	0.0022	1
AC023461.00010.1	91	0.0022	1
AC002317.00002.	89	0.0035	1
AC016063.00004.	112	0.0038	2
AL354984.00001.1	88	0.0045	1
AL139009.00003.	92	0.0051	1
AC004015.00001.1	87	0.0058	1
AC068640.00018.	87	0.0058	1
AC024233.00019.	87	0.0058	1
AC011791.00007.	75	0.0086	2
AL139411.00005.	85	0.0094	1
AC012588.00002.1	99	0.010	1

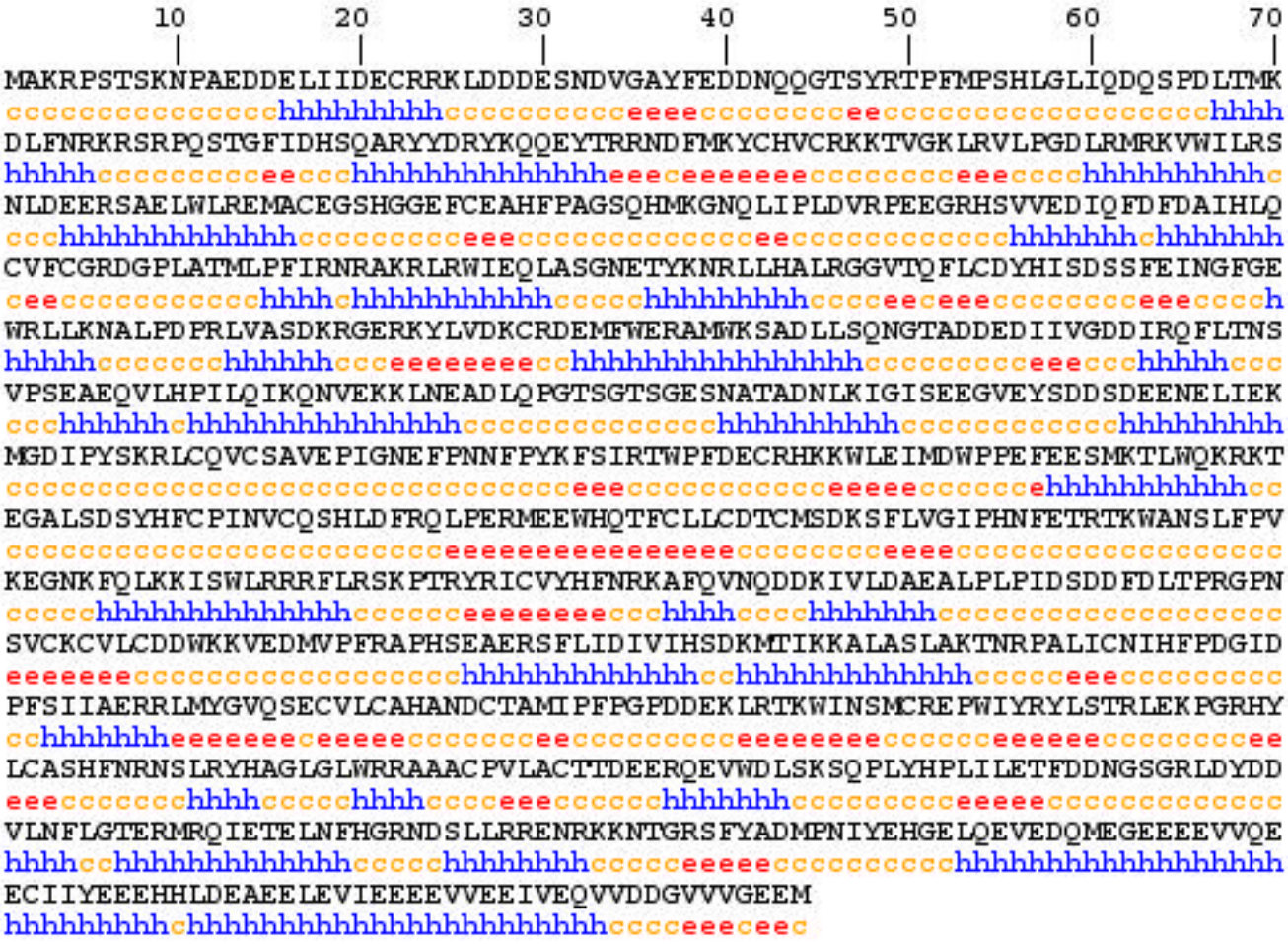
PropSearch results:

Rank	ID	DIST	LEN2	POS1	POS2	pI	DE
1	vav_human	7.70	846	1	846	6.59	VAV PROTO-ONCOGENE.
2	vav_rat	7.91	843	1	843	6.51	VAV PROTO-ONCOGENE (P95).
3	vav2_mouse	8.30	868	1	868	6.82	VAV2 PROTEIN.
4	vav_mouse	8.33	845	1	845	6.48	VAV PROTO-ONCOGENE.
5	fps_drome	8.63	803	1	803	6.72	TYROSINE-PROTEIN KINASE FPS
6	vav2_human	8.73	878	1	878	7.05	VAV2 PROTEIN.
7	rrp2_iaann	8.74	716	1	716	5.35	RNA-DIRECTED RNA POLYMERASE
8	vp41_human	8.88	854	1	854	5.79	VACUOLAR ASSEMBLY PROTEIN VPS41
9	v2a_bmv	8.88	822	1	822	5.30	RNA-DIRECTED RNA POLYMERASE
10	rrp2_iahte	8.96	716	1	716	5.49	RNA-DIRECTED RNA POLYMERASE
11	rrp2_iasin	9.07	716	1	716	5.30	RNA-DIRECTED RNA POLYMERASE
12	ubp7_human	9.07	1102	1	1102	5.33	UBIQUITIN CARBOXYL-TERMINAL
13	ysx7_caeel	9.08	978	1	978	5.64	HYPOTHETICAL 113.1 KD PROTEIN T28D9.7
14	v2a_ccmv	9.08	808	1	808	5.42	PROBABLE RNA-DIRECTED RNA polymerase
15	rrp2_iale2	9.10	716	1	716	5.26	RNA-DIRECTED RNA POLYMERASE
16	rrp2_iale1	9.12	716	1	716	5.21	RNA-DIRECTED RNA POLYMERASE
17	v2a_bbmv	9.12	810	1	810	5.19	RNA-DIRECTED RNA POLYMERASE
18	rrp2_iagu2	9.13	716	1	716	5.31	RNA-DIRECTED RNA POLYMERASE
19	rrp2_iawil	9.14	716	1	716	5.42	RNA-DIRECTED RNA POLYMERASE
20	can3_rat	9.14	821	1	821	6.02	CALPAIN P94, LARGE [CATALYTIC]
21	can3_mouse	9.16	821	1	821	6.06	CALPAIN P94, LARGE [CATALYTIC]
22	rrp2_iakor	9.17	716	1	716	5.20	RNA-DIRECTED RNA POLYMERASE
23	eg15_caeel	9.18	1040	1	1040	6.05	MYOBLAST GROWTH FACTOR RECEPTOR
24	rrp2_iafpr	9.19	716	1	716	5.82	RNA-DIRECTED RNA POLYMERASE
25	rrp2_iapi0	9.20	716	1	716	5.31	RNA-DIRECTED RNA POLYMERASE
26	rrp2_iapue	9.22	716	1	716	5.30	RNA-DIRECTED RNA POLYMERASE
27	rrp2_iahlo	9.23	716	1	716	5.36	RNA-DIRECTED RNA POLYMERASE
28	rrp2_iazh3	9.24	716	1	716	5.36	RNA-DIRECTED RNA POLYMERASE
29	rrp2_iant6	9.24	716	1	716	5.16	RNA-DIRECTED RNA POLYMERASE
30	jak1_human	9.24	1142	1	1142	7.61	TYROSINE-PROTEIN KINASE JAK1 (EC
31	pip5_human	9.25	1252	1	1252	6.63	1-PHOSPHATIDYLINOSITOL-4,5
32	rrp2_iatkm	9.26	716	1	716	5.31	RNA-DIRECTED RNA POLYMERASE
33	nof_drome	9.26	984	1	984	6.68	112 KD PROTEIN IN NOF-FB
34	rrp2_iavi7	9.27	716	1	716	5.25	RNA-DIRECTED RNA POLYMERASE
35	rrp2_iarud	9.27	716	1	716	5.42	RNA-DIRECTED RNA POLYMERASE
36	cnrb_mouse	9.28	856	1	856	5.17	ROD CGMP-SPECIFIC 3',5'-CYCLIC
37	rrp2_iadh2	9.28	716	1	716	5.26	RNA-DIRECTED RNA POLYMERASE
38	rrp2_iazh2	9.30	716	1	716	5.31	RNA-DIRECTED RNA POLYMERASE
39	jak1_mouse	9.33	1153	1	1153	7.53	TYROSINE-PROTEIN KINASE JAK1

Paircoil


Paircoil score
0.5
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
600
700
800
900
1000

GOR IV (structure prediction)



Helix-turn-helix prediction

The score is 0.52 at position 569.
The sequence at this position is KKISWLRRRFLRSKPTRYRICV.
This score is not significant.