eMOTIF Maker: Nodally Awesome:

Comparing Results of eMOTIF Maker with Neighbor-Joining Trees

Biochemistry 218

Douglas L. Brutlag

Lee Kozar

6/6/02

Pei-Hsien Ren

pren@stanford.edu

Phylogenetic trees are useful in determining the relationship among proteins and in grouping proteins into their correct family. Protein families have been helpful in elucidating the function and structure of new protein members. In principle, the tree building programs that are distance-based generate pairwise alignments of each sequence against all other sequences in the set. The mutation distances between all pairs are the stored in a matrix. Two taxa are joined as neighbors if the pair has the least mutational distance. The optimal tree is finally generated after minimizing mutation distances at each step (1).

A quite different program from that of tree-building also makes use of alignment of closely related proteins. eMOTIF maker takes these sequence alignments and returns a set of motifs with various degree of sensitivity and specificity. This is a way to discover motifs that are conserved among a protein family (2). Using this group of motifs to perform a scan in the database returns hits that should include members used to generate that motif or additional members containing that motif. The compiled result would vary according to the sensitivity and the specificity of the motif. Because of the similarity of approach used in both tree-building and motif-building, albeit for different purposes, this project would seek to do a proof-of-concept experiment to investigate how well the results from these two programs match up.

Out of the distance-based methods to build a tree, neighbor-joining proves to be very efficient in generating the best tree for large data set (3). In addition, neighbor-joining does not require the data to be ultrameteric and produces less biased tree when given sequence data that have unequal evolutionary rate (4). These characteristics of neighbor-joining make it the suitable tree-building method used in this experiment.

In order to make motifs out of related proteins, ungapped alignment of their sequences must be generated so the alignment can be input into eMOTIF maker. Block Maker program is used in this experiment to produce these alignments (5). Block Maker is chosen because of its ease of use and manipulation of formatted result.

The obtained blocks of sequence alignment are then put into eMOTIF maker and the result is a graphical representation of motif enumeration, showing each motif positioned according to its specificity on the y-axis and the number of training sequences it covers on the x-axis (2). If one seeks for a motif that covers a certain number of sequences, there is only one motif that can give the best specificity and that is the one that lies lowest on the y-axis. If one seeks for a motif with a certain specificity, there is also only one motif that gives the best coverage and that is the one that lies on the most right on the x-axis. These dominating motifs can then be connected by a line called the Pareto-optimal curve (2). Motifs lying on the Pareto-optimal curve are then used for subsequent motif scan in this experiment.

Two training sets, each with 100 sequences or more, are used to compare the neighbor-joining tree with the results of a motif scan. Motif scan done with a motif with the highest specificity should have the least coverage and its hits should correspond to a small cluster under few nodes in the tree. This small cluster would contain sequences that are the most related to each other. Motif scan done with a low specificity motif should return high number of hits that correspond to sequences under more number of nodes.

One of the two training set is derived from the globin family, heme-containing orthologous proteins, all of them are vertebrate proteins. The set composes mostly of alpha chains, beta chains, and their variants with a few myoglobins. See Appendix for a

complete list of the set (6). The large number of alpha and beta chains should return trees with nodes where most, if not all, of the sequences clustered according to the type of chains.

The second training set is derived from a subfamily of the serine proteases family – trypsin family with the serine active site. The set composes of mast cell proteases, trypsins, and various forms of venom serine proteases. See Appendix for a complete list of the set (7). This paralogous family has various proteins that although acts to cleave proteins, do not share functions in the same context and would probably cluster according to their functions in context. There has been some difficulty in choosing members of this training set because each protein member of the family has diverged much. Even if the active sites are very similar, the global sequence alignment is impossible because the majority of the sequences are too different form each other. This training set attempts to include members that are very similar in both sequence and function within a subgroup but also to include three different, divergent functions of serine proteases.

RESULTS

With the orthologous set as the input, the neighbor-joining method produces a tree that places most of beta chains and their variants as the outgroup to alpha and myoglobin. See Figure 1. All and only the myoglobins fall under one main node and so do the alpha chains. The beta chains are dispersed and there is no one node under which all the beta chains fall and they have given rise to many other variants of hemoglobins. Some of them have evolved from the same ancestral sequence that gives rise to the alpha chains. One pair that is closest to alpha chain in distance is the hbb1_torma and hbb2_torma. Although they seem to be quite close in distance to the alpha cluster, biochemical and

structural data from the database have identified the two proteins as beta chains (6). The representation of the tree may seem as if the hbb1 and hbb2_torma evolve from alpha chains but the tree actually shows that they and the alpha chains split from the same ancestral sequence derived from the other beta clusters.

Looking at the tree again, one can also see certain isolated groups, such as the epsilon chains, that branch off close to the root between two subtypes. That could be an indication of recombination. Overall, the tree seems to be reasonable and there is no one particular pair that seems misplaced.

The same training set is submitted to Block Maker for alignment and two blocks are generated that covers all sequences in the training set (Table 3). EMOTIF maker uses these two blocks to make motifs. Graph enumerating all the motifs from various blocks are in the appendix. 7 random motifs are sampled for subsequent motif scan.

For the first block, the most specific motif returns most of the sequences that lie to the left of the red line marked on the tree (Fig 1). They are mostly beta and epsilon chains and fetal forms of hemoglobin. The beta-1 and beta-2 variants, bracketed by purple brackets in Fig. 1, are not picked up by the first motif. The second most specific motif returns some more sequences that are missed by the first motif. The sixth motif, ranked in terms of its specificity, returns more beta and epsilon chains from chick that are not picked up by the first motif. They lie between the two green vertical lines marked on the tree. It is not until the 30^{th} motif that the beta-1 and –2 chains are picked up. It is not until the 51^{st} motif that the alpha chains are picked up. The 51^{st} motif picks up all most of the alpha chains that lie between the two blue lines.

In an attempt to present quantitatively the correlation between the motif specificity and the tree clusters, a hypothetical set is devised. This hypothetical set is a group of sequences that fall under a particular node of the tree of which a majority of sequences are hits of a motif of a particular specificity. If roughly more than 50% of the sequences are picked up by a motif, every sequence under that node is included in the set. For example, all sequences between the two green lines are included in a hypothetical set belonging to the sixth motif. Similar method is applied to the sequences covered by the purple brackets and so on. Since there are numerous nodes and isolated branches between subtypes, the set is designed arbitrarily and by eye. The quantification should only be seen as a rough characterization from looking at the tree. Results are summarized in Table 1.

After repeating the same process with block 2, one can see that the results are quite different by comparing figure 1 to figure 2. Block 2 is an alignment more internal to the sequence since block 1 lies to the n-terminus of block 2. An alignment derived from more internal sequences may be a better correlation with the variant characteristics of the protein function. One can see that the block 2 picks up many more clusters, with each cluster having fewer nodes than block1. Hence, it is more specific. Looking at the results summarized in Table 2, the percentages of positive hits are higher than those in Table 1. Block 2 also does not have the characteristic of picking up sequences from a particular organism, such as block 1 on the chick globins that fall between the two green lines. The most specific motif generated from Block 2 picks up most of the beta chains.

The second training set containing serine proteases basically are processed in the same way. First, a neighbor-joining tree is generated (Fig 3). Looking at the tree, the

clusters seem to form according to the function assigned to the sequences. All the venom proteinases group to one side of the tree very early on and are the outgroup to all other members. Most of the trypsins pair up according to organisms' taxonomy and the type of trypsin. Another group is the mast cell proteases and they all cluster together under one main node. No other mast cell proteases are found among other clusters. Overall, the tree looks very reasonable.

The result of the motif scan done using block 1 shows that the most specific motif defines trypsin better than either venom proteinases or mast cell proteases (sequences with blue lines or brackets in Fig 3). A couple of trypsins that have longer branch length, such as fish (gadmo) and hawkmoth (manse) are not picked up until after 24th motif. This makes sense that if the sequences have diverged much through time, it would be less similar in their alignment. Motifs of 12th and 16th specificity return several groups of venom proteinases and several trypsinP. According to the tree, venom proteinases and trypsinP are quite distant taxa. The sources of these trypsin range from vertebrate, insects, fungus, to bacteria. Due to this variability, the context in which trypsinP works is not clear. The fact that the motif of this particular specificity picks up both typsinP and venome proteinases could be explained by convergent evolution in this block.

Motif of 20th specificity defines very well the node under which all the mast cell proteases fall since all of the mast cell proteases are returned with this motif (red bracket in Fig 3).

As for block 2, the most specific motif returns many venom proteinases along with most of the trypsin in the cluster closest to the venom taxa (blue lines and brackets in Fig 4). This block probably is the main reason why this group of trypsins is close in

mutational distance to the venom taxa. The purple brackets in Fig 4 denotes the group that is returned with a motif of specificity 30th. This group composes of mosquito trypsins (anoga) and mammalian trypsins secreted by mast cells. It is interesting to speculate that to digest blood proteins, the mosquito trypsins have evolved similarly to mammalian trypsins in a particular segment as both are functioning in the context of blood/lymph. Motif of specificity 36th recovers only a small group of mast cell proteases. This block may not define the characteristics of mast cell proteases very well.

DISCUSSION

At the first glance of figures 1-4, one can see that several nodes correspond quite well to a motif of a particular specificity. As the motif become less sensitive, more sequences are returned as hits. There would be time when the motif is not sensitive enough for this purpose and positive hits are returned with no particular cluster relationship to the nodes of the tree.

If one sees a colored marking that denotes a isolated sequence as a positive hit while no other neighbor sequence in its cluster are returned as a positive hit, it is very likely that the colored marking correspond to a motif of low specificity and high coverage. In this case, the number of false positives increases and this isolated sequence may simply match due to its consensus sequence and does not provide information for the goal of this experiment.

Various blocks generated from different segment of the sequences also have different effects on how well the motif scan results correspond to the nodes of a tree. For example, block 1 and block 2 from both training sets have produced different hits when a motif scan is done. If a particular segment used to generate the block can reflect the

mutation distance of a member to other members of the set, then the motif scan from that block would produce collaborating results. It is interesting to see that a specific motif generated from different block can pick up different combination of groups. For the Trypsin_SER trees, the first block generates a specific motif that picks up typsins & its variants, while the 2nd block returns the venom taxa along with one subtype of trypsin.

Despite the variations from block segments and isolated branches in a tree, this experiment demonstrates that motif sensitivity seems to correspond well to the nodes of the neighbor-joining trees overall. There are several things that can be improved so that the result can be more definitive than what is presented here. The members of the training set have not been chosen with as much care as they deserve. There are several sequences that are never returned as hits even if the motif specificity is very low.

For example, in the globin training set, beta-1 chains of Indian Cobra (Najna) and Electric Ray (Torma) are not picked up by any of the 7 sample motifs used although they are close neighbors to other beta chains that are picked up pretty early on. After doing some additional searches, it seems that electric ray beta-1 chain is pretty distant from all other beta-1 chains compare to other beta-1 chains. There is another fish beta-1 chain that is closely related to electric ray beta-1 chain named hbb1_Dasak, whose crystal structure has been solved. Hbb1_Torma is probably classified as hbb1 since it is related to hbb1_Dasak, whose structure confirms its identity as hbb1. Hbb1_Dasak seems to be able to group with other hbb1 better than hbb1_Torma. In this case, it would be helpful to compare the sequences of hbb1_Torma to hbb1_Dasak and other hbb1 to figure out why hbb1_Torma is dissimilar to other hbb1.

Comparing the globins training set and the Trypsin_SER training set reveals another weakness in this experiment. The motifs generated from the globins alignment have much less number of expected false positives than Trypsin_Ser (Tables 1-4). Although the function of trypsin is to cleave protein, the context of function has diverged much and to generate a motif that attempts to have a decent coverage, specificity is compromised. One way to improve this paralogous training set is to compile members that would generate blocks that are as specific as the ones in the globin training set.

In conclusion, this proof-of-concept experiment shows that the output of tree-building algorithm can be matched with motifs with various specificities. Looking at the tables, the percentages of matching are generally 80% or better. If there is a more quantitative method to represent this observation, the result can be analyzed with statistics. Nevertheless, the results of this experiment are in agreement and correlate with the principles behind both tree-building method and motif maker despite its shortcomings.

Fig. 1. Neighbor-Joining Phylogenetic Tree of Globins, Marked according to motif scans generated from Block 1

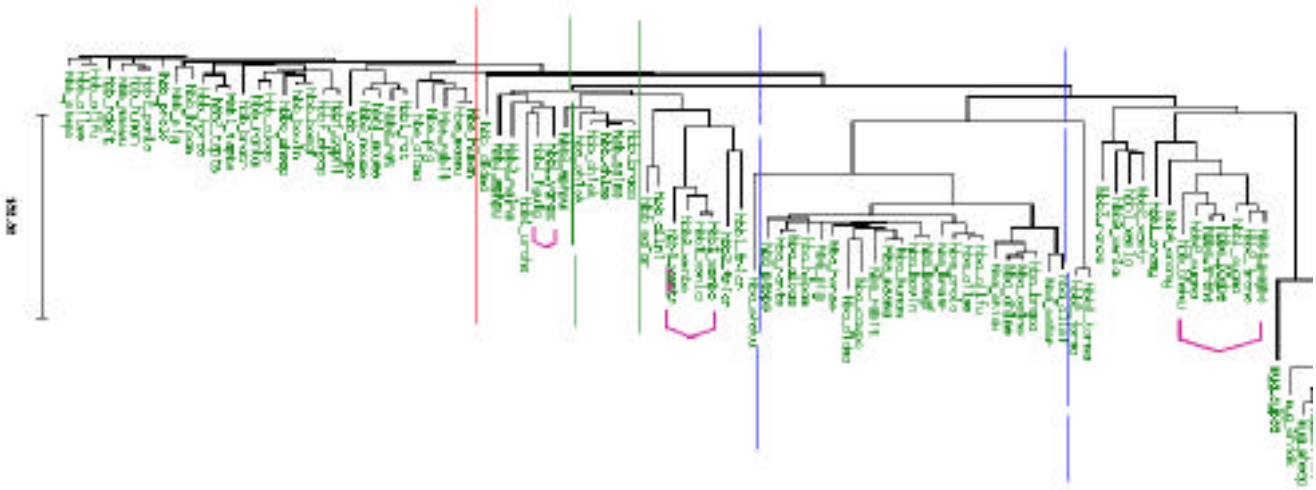


Fig. 2 Neighbor-Joining Phylogenetic Tree of Globins, Marked according to motif scans generated from Block 2

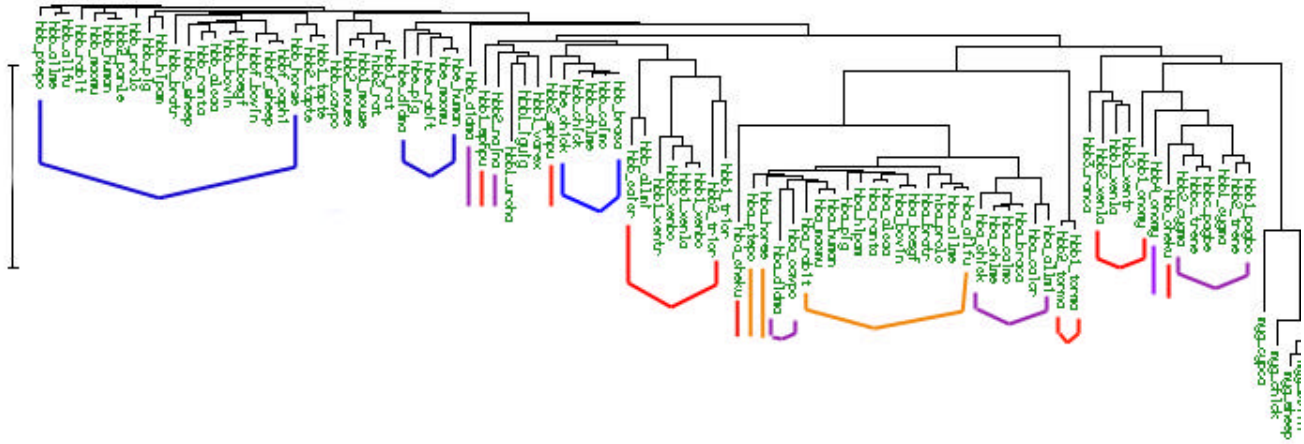


Table 1. Summarization of Tree and eMOTIF maker correlation : Globins Block 1

Motif Specificity* (# of expected FP)	Number of hits	Number of sequences in a hypothetical set	% of hits / sequences in the set
1 (10^{-26})	24	31	77
2 (10^{-25})	27	31	87
6 (10^{-24})	33	37	89
9 (10^{-23})	36	41	88
16 (10^{-19})	42	42	100
30 (10^{-12})	57	58	98
51 (10^{-3})	78	82	95

*with 1 being the most specific

Table 2. Summarization of Tree and eMOTIF maker correlation : Globins Block 2

Motif Specificity* (# of expected FP)	Number of hits	Number of sequences in a hypothetical set	% of hits / sequences in the set
1 (10^{-31})	27	32	84
2 (10^{-31})	31	32	97
6 (10^{-29})	35	37	95
10 (10^{-26})	38	40	95
20 (10^{-19})	53	58	91
40 (10^{-13})	70	72	97
58 (10^{-6})	86	91	95

* with 1 being the most specific

Fig 3. Neighbor-Joining Phylogenetic Tree of Trypsin_SER, Marked according to motif scans generated from Block 1

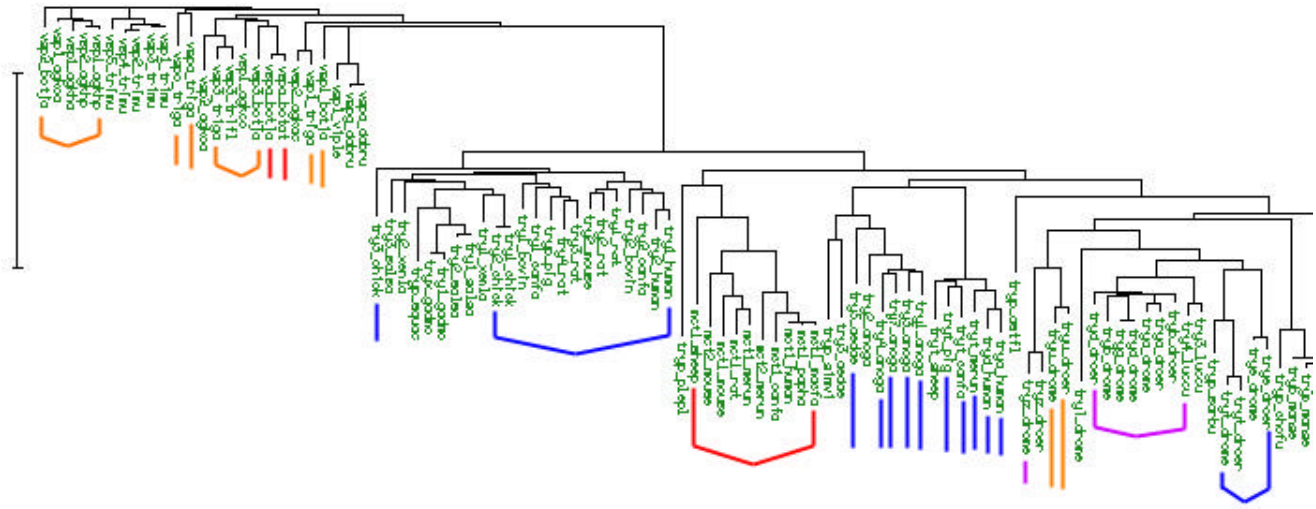


Fig 4. Neighbor-Joining Phylogenetic Tree of Trypsin_SER, Marked according to motif scans generated from Block 2

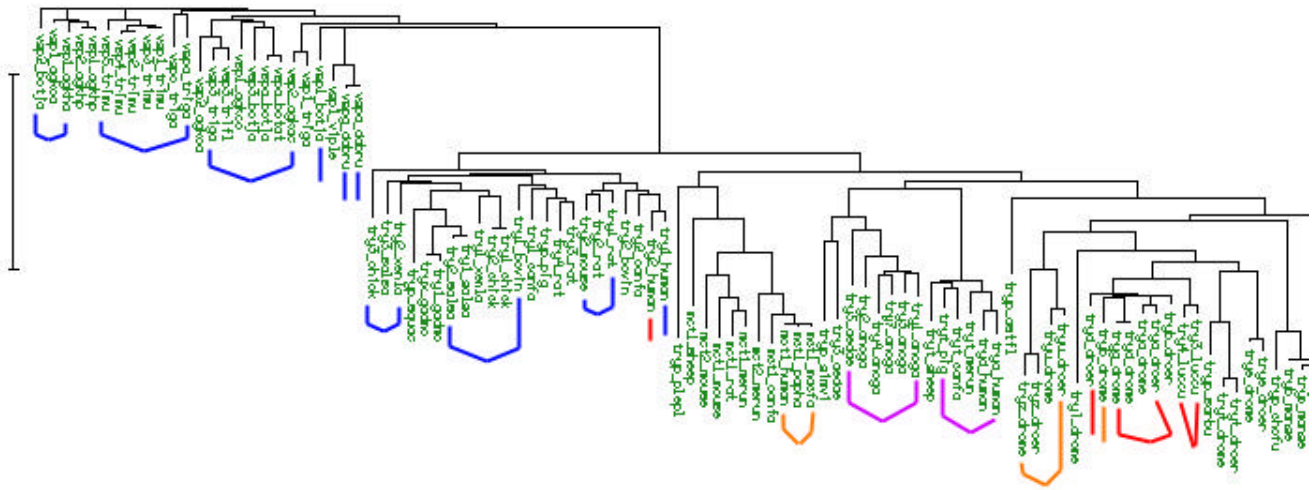


Table 3. Summarization of Tree and eMOTIF maker correlation: Trypsins_SER Block 1

Motif Specificity* (# of expected FP)	Number of hits	Number of sequences in a hypothetical set	% of hits / sequences in the set
1 (10^{-4})	30	42	71
2 (10^{-4})	39	53	74
8 (10^{-3})	42	53	79
12 (10^{-1})	49	60	82
16 (10^{0})	57	68	84
20 (10^{0})	69	82	84
24 (10^{2})	79	86	88
27 (10^{3})	93	100	93

*with 1 being the most specific

Table 4. Summarization of Tree and eMOTIF maker correlation: Trypsins_SER Block 2

Motif Specificity* (# of expected FP)	Number of hits	Number of sequences in a hypothetical set	% of hits / sequences in the set
1 (10^{-16})	33	48	69
6 (10^{-13})	42	48	88
9 (10^{-12})	43	48	90
16 (10^{-7})	52	57	91
25 (10^{-4})	61	62	98
30 (10^{-1})	73	76	96
36 (10^{0})	83	91	91

43 (10^{1})	85	91	93

*with 1 being the most specific

REFERENCES

1. Thompson, Higgins & Gibson, *Nucleic Acids Res* 1994 Nov 11;22(22):4673-80.
2. Nevill-Manning, Wu & Brutlag, Proc. Natl. Acad. Sci. 1998 May 95:5865-5817.
3. Saitou & Nei, Mol. Biol. Evol. 4(4):406-425.
4. Kuhner & Felsenstein, Mol. Biol. Evol. 1994, 11(3):459-468.
5. Henikoff, Henikoff, Alford, and Pietrokovski (1995), Gene 163:GC17-26.
6. http://us.expasy.org/cgi-bin/get-prosite-entry?PS01033
7. http://www.expasy.ch/cgi-bin/prosite-search-de?search=PS00135

Appendix

Table 1.

Entry Name: **GLOBIN**
Accession number: **PS01033**

HBB1_CYGMA (P23017), HBB1_IGUIG (P18987), HBB1_MOUSE (P02088),
HBB1_ONCMY (P02142), HBB1_PAGBO (O93348), HBB1_RAT (P02091),
HBB1_SPHPU (P10060), HBB1_TAPTE (P02064), HBB1_TORMA (P20246),
HBB1_TRICR (P10785), HBB1_UROHA (P18991), HBB1_VAREX (P18993),
HBB1_XENBO (P07432), HBB1_XENLA (P02132), HBB1_XENTR (P07429),
HBB2_CYGMA (P23018), HBB2_MOUSE (P02089), HBB2_NAJNA (P22743),
HBB2_PANLE (P18988), HBB2_RAT (P11517), HBB2_SPHPU (P10061),
HBB2_TAPTE (P02065), HBB2_TORMA (P20247), HBB2_TRENE (O93349),
HBB2_TRICR (P10786), HBB2_XENBO (P07433), HBB2_XENLA (P02133),
HBB2_XENTR (P08423), HBB3_RANCA (P02136), HBB4_ONCMY (P02141),

HBA_MACMU (P01925), HBA_DIDMA (P01976), HBA_AILFU (P18969),
HBA_AILME (P18970), HBA_ALCAA (P01971), HBA_ALLMI (P01999),
HBA_BOSGF (P01969), HBA_BOVIN (P01966), HBA_BRACA (P01991),
HBA_BRATR (P14525), HBA_CAICR (P02000), HBA_CAIMO (P01987),
HBA_CHEKU (P80270), HBA_CHICK (P01994), HBA_CHLME (P07034),
HBA_HIPAM (P19015), HBA_HORSE (P01958), HBA_HUMAN (P01922),
HBA_PTEPO (P14390), HBA_RABIT (P01948), HBA_RANTA (P21379),
HBA_PIG (P01965), HBA_PROLO (P18977), HBA_CAVPO (P01947),

HBB_MACMU (P02026), HBB_DIDMA (P02109), HBB_AILFU (P18982),
HBB_AILME (P18983), HBB_ALCAA (P02073), HBB_ALLMI (P02130),
HBB_BOSGF (P02071), HBB_BOVIN (P02070), HBB_BRACA (P02119),
HBB_BRATR (P14526), HBB_CAICR (P02131), HBB_CAIMO (P14260),
HBB_CHEKU (P80271), HBB_CHICK (P02112), HBB_CHLME (P07036),
HBB_HIPAM (P19016), HBB_HORSE (P02062), HBB_HUMAN (P02023),
HBB_PTEPO (P14392), HBB_RABIT (P02057), HBB_RANTA (P21380),
HBB_PIG (P02067), HBB_PROLO (P18989), HBB_CAVPO (P02095),

HBE_PIG (P02101), HBE_CHICK (P02128), HBE_HUMAN (P02100),
HBE_RABIT (P02103), HBE_MACMU (Q28507), HBE_DIDMA (P11025),

MYG_CHICK (P02197), MYG_CYPCA (P02204), MYG_DIDMA (P02193),
MYG_MOUSE (P04247), MYG_BOVIN (P02192), MYG_HORSE (P02188),
MYG_HUMAN (P02144), MYG_RABIT (P02170), MYG_SHEEP (P02190),

HBBC_PAGBE (P45722), HBBC_SHEEP (P02079), HBBC_TRENE (P45721),
HBBF_BOVIN (P02081), HBBF_CAPHI (P02082), HBBF_SHEEP (P02083),
HBBL_XENLA (P02137)

Table 2.

Entry Name: **TRYPSIN_SER**
Accession number: **PS00135**

TRY1_ANOGA (P35035), TRY1_BOVIN (P00760), TRY1_CANFA (P06871),
TRY1_CHICK (Q90627), TRY1_GADMO (P16049), TRY1_HUMAN (P07477),
TRY1_RAT (P00762), TRY1_SALSA (P35031), TRY1_XENLA (P19799),
TRY2_ANOGA (P35036), TRY2_BOVIN (Q29463), TRY2_CANFA (P06872),
TRY2_CHICK (Q90628), TRY2_HUMAN (P07478), TRY2_MOUSE (P07146),
TRY2_RAT (P00763), TRY2_SALSA (P35032), TRY2_XENLA (P70059),
TRY3_AEDAE (P29786), TRY3_ANOGA (P35037), TRY3_CHICK (Q90629),
TRY3_LUCCU (P35043), TRY3_RAT (P08426), TRY3_SALSA (P35033),
TRY4_ANOGA (P35038), TRY4_LUCCU (P35044), TRY4_RAT (P12788),
TRY5_AEDAE (P29787), TRY7_ANOGA (P35041), TRYA_DROER (P54624),
TRYA_DROME (P04814), TRYA_HUMAN (P15157), TRYA_MANSE (P35045),
TRYB_DROER (P54625), TRYB_DROME (P35004), TRYB_MANSE (P35046),
TRYC_MANSE (P35047), TRYD_DROER (P54626), TRYD_DROME (P42276),
TRYD_HUMAN (Q9BZJ3), TRYE_DROER (P54627), TRYE_DROME (P35005),
TRYG_DROME (P42277), TRYI_DROME (P52905), TRYP_ASTFL (P00765),
TRYP_CHOFU (P35042), TRYP_FUSOX (P35049), TRYP_PIG (P00761),
TRYP_PLEPL (P35034), TRYP_SACER (P24664), TRYP_SARBU (P51588),
TRYP_SIMVI (P35048), TRYP_SQUAC (P00764), TRYP_STRGA (Q54179),
TRYP_STRGR (P00775), TRYT_CANFA (P15944), TRYT_DROER (P54628),
TRYT_DROME (P42278), TRYT_MERUN (P50342), TRYT_PIG (Q9N2D1),
TRYT_SHEEP (Q9XSM2), TRYU_DROER (P54629), TRYU_DROME (P42279),
TRYX_GADMO (Q91041), TRYZ_DROER (P54630), TRYZ_DROME (P42280),
VSP1_AGKCA (Q91053), VSP1_AGKCO (P09872), VSP1_AGKHA (P81176),
VSP1_AGKHP (Q9YGJ2), VSP1_BOTJA (P81824), VSP1_TRIGA (O13059),
VSP1_TRIMU (Q91507), VSP1_VIPLE (Q9PT41), VSP2_AGKAC (Q9I8X1),
VSP2_AGKCA (O42207), VSP2_AGKHP (Q9YGI6), VSP2_BOTJA (O13069),
VSP2_TRIMU (Q91508), VSP3_BOTJA (Q9PTU8), VSP3_TRIFL (O13058),
VSP3_TRIGA (O13063), VSP3_TRIMU (Q91509), VSP4_TRIMU (Q91510),
VSP5_TRIMU (Q91511), VSPA_BOTAT (P04971), VSPA_BOTJA (P81661),
VSPA_DABRU (P18964), VSPA_TRIGA (O13060), VSPC_TRIGA (O13062),
VSPG_DABRU (P18965), MCT1_CANFA (P21842), MCT1_HUMAN (P23946),
MCT1_MACFA (P56435), MCT1_MERUN (P50340), MCT1_MOUSE (P11034),
MCT1_PAPHA (P52195), MCT1_RAT (P09650), MCT1_SHEEP (P80931),
MCT2_MERUN (P50341), MCT2_MOUSE (P15119)

Table 3. Globins Blocks from Block Maker

```
unknownA, width = 37
 gi|1170175    24 GPATLARCLVVYPWTQRYFGKFGNLYNATAIAENAMV
 gi|1170176    24 GPATLARCLVVYPWTQRYFGKFGNLYNAAAIAQNAMV
  gi|122341    25 GGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGKK
  gi|122342    25 GGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGKK
  gi|122343    25 GAEALERMFLSFPTTKTYFPHFDLSHGSAQVKAHGEK
  gi|122344    25 GAEALERMFCAYPQTKIYFPHFDMSHNSAQIRAHGKK
  gi|122360    25 GAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGAK
  gi|122362    25 GAETLERMFVAYPQTKTYFPHFDLQHGSAQIKAHGKK
  gi|122363    25 GGEALERTFLSFPTTKTYFPHFDLSPGSAQVKAHGKK
  gi|122364    25 GAEALERMFCAYPQTKIYFPHFDMSHNSAQIRGHGKK
  gi|122365    26 GAETLERMFIAYPQTKTYFPHFDLQHGSAQIKAHGKK
  gi|122372    25 VAEGLTRMFTSFPTTKTYFHHIDVSPGSGDIKAHGKK
  gi|122378    26 GAETLERMFTTYPPTKTYFPHFDLSHGSAQIKGHGKK
  gi|122379    25 GAETLERMFIAYPQTKTYFPHFDLHHGSAQIKAHGKK
  gi|122395    25 MGEALYRTFLSFPTTKTYFPNYDFSAGSAQIKTQGQK
  gi|122410    25 GAEALERMFLSFPTTKTYFPHFDLSHGSSQVKAHGKK
  gi|122411    26 GAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHGKK
  gi|122412    26 GAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKK
```

```
gi|122465    25 GAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHGQK
gi|122470    25 GGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHGKK
gi|122474    25 GAEALERMFLSFPTTKTYFPHFDLAHGSSQVKAHGKK
gi|122475    26 GAEAVERMFLGFPTTKTYFPHFDFTHGSEQIKAHGKK
gi|122476    25 GAEALERMFLSFPTTKTYFPHFDLSHGSAQVKAHGEK
gi|122512    24 GGETLACLLVVYPWTQRFFPDFGNLSNAAAICGNAKV
gi|122513    25 GGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNAKV
gi|122514    25 GGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKV
gi|122515    24 GPLALARVLIVYPWTQRYFGSFGNVSTPAAIMGNPKV
gi|122516    24 GGEALGRLLVVYPWTQRFFADFGNLSSATAICGNPRV
gi|122517    24 GGEALGRLLVVYPWTQRFFDSFGDLSTAAAVMGNPKV
gi|122518    24 TAKALERVFYVYPWTTRLFTSFNHNFKASDKQVHDHA
gi|122519    23 GAEALGRLILVNPWTRRYFKSFGDLSSAEAIQHNPKV
gi|122520    24 GGETLANLLVVYPWTQRFFEDFGNLSTPSAILNNPKX
gi|122521    24 GGETLAGLLVIYPWTQRQFSHFGNLSSPTAIAGNPRV
gi|122522    24 GKEALGRLLWTYPWTQRYFSSFGNLNSADAVFHNEAV
gi|122523    24 GQEALGRLLWTYPWTQRYFSSFGNLNSADAVFHNEAV
gi|122524    25 GKQALGSMLYTYPWTQRYFSSFGNLSSIEAIFHNAAV
gi|122525    24 GPATLARCLVVYPWTQRYFGKFGNLYNAAAIAENAMV
gi|122526    25 GGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKV
gi|122527    24 GAATLGKMMVMYPWTQRFFAHFGNLSGPSALCGNPQV
gi|122528    24 GGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
gi|122529    25 GAEALGRLLVVYPWTQRYFSKFGDLSSASAIMGNPQV
gi|122530    24 GGEALGRLLIVYPWTQRFFSSFGNLSSSTAICGNPRV
gi|122531    24 GGEALGRLLVVYPWTQRFFDSFGDLSTAAAVMGNPKV
gi|122532    24 TAKALERVFYVYPWTTRLFTSFNHNFKASDKGVHDHA
gi|122533    24 GGQCLARLIVVNPWSRRYFHDFGDLSSCDAICRNPKV
gi|122535    24 GKEALGRLLNTFPWTQRYFSSFGNLGSAEAIFHNEAV
gi|122536    25 GHDALGRLLIVYPWTQRYFSNFGNLSNSAAVAGNAKV
gi|122537    25 GHDALSRLLVVYPWTQRYFSSFGNLSNVSAVSGNVKV
gi|122538    24 GPQALARLLIVSPWTQRHFSTFGNLSTPAAIMGNPAV
gi|122543    20 GAEALGRLLVVYPWTQRFFEHFGDLSTADAVLGNAKV
gi|122544    23 GGEALGRLLVVYPWTQRFFESFGDLSSADAILGNPKV
gi|122545    23 GGEALGRLLVVYPWTQRFFEHFGDLSSADAILGNPKV
gi|122546    23 GGEALGRLLVVYPWTQRFFEHFGDLSSADAILGNPKV
gi|122548    25 GHDALTRLLVVFPWTQRYFSSFGNLSNVAAISGNAKV
gi|122553    24 GGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKV
gi|122554    24 GGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKV
gi|122555    23 GGEALGRLLVVYPWTQRFFEHFGDLSTADAVMHNAKV
gi|122556    24 GADALSRMLIVYPWKRRYFEHFGKMCNAHDILHNSKV
gi|122570    23 GGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKV
gi|122572    23 GGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKV
gi|122573    24 GAEALARLLIVYPWTQRFFSSFGNLSSPTAILGNPMV
gi|122574    24 GGEALGRLLVVYPWTSRFFESFGDLSSADAVFSNAKV
gi|122575    24 GGDALSRMLIIYPWKRRYFEHFGKLSTDQDVLHNEKI
gi|122576    25 GAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMV
gi|122581    24 GAEALGRLLVVYPWTQRFFEKFGDLSSASAIMSNAHV
gi|122587    25 GAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMV
gi|122588    24 GAEALARLLIVYPWTQRFFASFGNLSSPTAISGNPMV
gi|122601    25 GGEALGRMLVVYPWTTRFFGSFGDLSSPGAVMSNSKV
gi|122613    24 GGEALGRLLVVYPWTQRFFESFGDLSSADAVMNNPKV
gi|122614    24 GGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKV
gi|122615    25 GGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
gi|122634    24 GGEALGRLLLVYPWTQRFFESFGDLSSPDAVMGNPKV
gi|122671    24 GGEALGRLLVVYPWTQRFFESFGDLSSADAIMGNPKV
gi|122675    24 GGEALGRLLVVYPWTQRFFDSFGDLSSAPAVMGNPKV
gi|122676    25 GGEALGRLLVVYPWTQRFFESFGDLSSANAVMNNPKV
```

```
          gi|122678    23 GAEALGRLLVVYPWTQRFFEHFGDLSSADAIMHNDKV
          gi|122723    25 GAEALARLLIVYPWTQRFFASFGNLSSPTAIMGNPRV
          gi|122724    25 GGESLARLLVVYPWTQRFFDSFGNLSSASAVMGNPKV
          gi|122726    25 GGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPKV
          gi|122731    25 GGEALGRLLVVYPWTQRFFDNFGNLSSSSAIMGNPKV
          gi|127638    26 GQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL
          gi|127647    21 GGEVLTRLFKQHPETQKLFPKFVGIASNELAGNAAVK
          gi|127648    25 GQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL
          gi|127661    26 GQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL
          gi|127676    26 GQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSEDL
          gi|127691    25 GQEVLIRLFHTHPETLEKFDKFKHLKSEDEMKASEDL
          gi|127694    25 GQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL
        gi|13634094    26 GAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGAK
        gi|14194774    25 GPKALSRCLIVYPWTQRHFSGFGNLYNAESIIGNANV
        gi|14194794    25 GPKALSRCLIVYPWTQRHFSGFGNLYNAEAIIGNANV
        gi|14195588    25 GPKALSRCLIVYPWTQRHFSGFGNLYNAEAIIGNANV
         gi|1708122    25 GHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPKV
         gi|2506462    25 GQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL
         gi|3041678    25 GGEALGRLLVVYPWTQRFFESFGDLSNADAVMGNPKV
         gi|3041679    25 GGQAVGRLLVVYPWTQRFFDSFGNMSSPSAIMGNPKV
          gi|462246    25 GAEALARMLTVYPQTKTYFTHWTDLSPSSTSVKNHGK
          gi|462247    24 GPATLTRTVIVYPWTLRYFAKFGNICSTAAILGNKEI
          gi|462677    25 GHEVLMRLFHDHPETLDRFDKFKGLKTPDQMKGSEDL
         gi|6016192    25 GGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPKV
         gi|6166198    26 GAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKK

             unknownB, width = 44
gi|1170175  (  13)    74 AVKNMDDIKNTYAELSVLHSEKLHVDPDNFKLLADCLTIVVAAR
gi|1170176  (  13)    74 AVKNMDDITNTYAELSVLHSEKLHVDPDNFKLLADCLTIVVAAR
gi|122341  (   7)    69 AVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACH
gi|122342  (   7)    69 AVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASH
gi|122343  (   7)    69 AVGHLDDLPGTLSDLSDLHAHKLRVDPVNFKLLSHTLLVTLAAH
gi|122344  (   7)    69 AVNHIDDLPGALCRLSELHAHSLRVDPVNFKFLAHCVLVVFAIH
gi|122360  (   7)    69 AVGHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASH
gi|122362  (   7)    69 AVNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIH
gi|122363  (   7)    69 AVGHLDDLPGALSDLSDLHAHKLRVDPVNFKLLGHCVLVTLALH
gi|122364  (   7)    69 AVNHIDDLAGALCRLSDLHAHNLRVDPVNFKFLSQCILVVFGVH
gi|122365  (   7)    70 AVNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIH
gi|122372  (   7)    69 AVGHLDDLPTALSTLSDVHAHKLRVDPVNFKFLNHCLLVTLAAH
gi|122378  (   7)    70 AANHIDDIAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIH
gi|122379  (   7)    69 AVNHIDDITGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIH
gi|122395  (   7)    69 AVAHLDDMPTALSSLSDLHAHELKVDPVNFKFLCHNVLVTMAAH
gi|122410  (   7)    69 AVGHLDDLPGALSDLSDLHAHKLRVDPVNFKLLSHCLLVTLAAH
gi|122411  (   7)    70 AVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVH
gi|122412  (   7)    70 AVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH
gi|122465  (   7)    69 AVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH
gi|122470  (   7)    69 AVGHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACH
gi|122474  (   7)    69 AVGHMDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLANH
gi|122475  (   7)    70 AVGHLDDLPGALSTLSDLHAHKLRVDPVNFKLLSHCLLVTLANH
gi|122476  (   7)    69 AVGHLDDLPGTLSDLSDLHAHKLRVDPVNFKLLSHTLLVTLASH
gi|122512  (  13)    74 AVKNLDNIKDTFAKLSELHCDKLHVDPVNFRLLGNVMITRLAAH
gi|122513  (  13)    75 GLNHLDSLKGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHH
gi|122514  (  13)    75 GLKHLDNLKGTFAHLSELHCDKLHVDPENFRLLGNMIVIVLGHH
gi|122515  (  13)    74 AVKNMGNILATYKSLSETHANKLFVDPDNFRVLADVLTIVIAAK
gi|122516  (  13)    74 ALKHLDNLKETFASLSELHCDKLHVDTENFKLLGNLVIVVLAAR
gi|122517  (  13)    74 GVHHLDDLKVTFAQLSELHCDKLHVDPENFRLLGNVLVVVLAQQ
gi|122518  (   9)    70 AIGDLHDINKNFSALSTKHQKKLGVDTSNFMLLGQAFLVELAAL
```

```
gi|122519  (  13)    73 AVKHLDDLKAYYADLSTIHCKKLYVDPANFKLFGGIVSIVTGMH
gi|122520  (  13)    74 ALKNLDNVXXXXXKLSEYHCNKLHVDPVNFRLLGDVLITLSAAN
gi|122521  (  13)    74 AIKNLDNIKDTFAKLSELHCDKLHVDPTNFKLLGNVLVIVLADH
gi|122522  (  13)    74 AIKHMDDIKGYYAQLSKYHSETLHVDPCNFKRFGGCLSISLARQ
gi|122523  (  13)    74 AIKHMDDIKGYYAQLSKYHSETLHVDPLNFKRFGGCLSIALARH
gi|122524  (  13)    75 AIKHMDDIKGYYAQLSKYHSETLHVDPYNFKRFCSCTIISMAQT
gi|122525  (  13)    74 AVKNMDDIKNTYAELSVLHCDKLHVDPDNFQLLAECLTIVLAAQ
gi|122526  (  13)    75 GLKNLDNLKGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHH
gi|122527  (  13)    74 ALKHLDNVKETFAKLSELHFDKLHVDPENFKLLGNVLIIVLAGH
gi|122528  (  13)    74 GLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH
gi|122529  (  13)    75 GLKHLDNLKGTFAHLSELHCDKLHVDPENFRLLGNMIVIVLGHH
gi|122530  (  13)    74 AVKNLDNIKATYAKLSELHCEKLHVDPQNFNLLGDIFIIVLAAH
gi|122531  (  13)    74 GVHHLDDLKVTFAQLSELHCDKLHVDPENFRLLGNVLVVVLAQQ
gi|122532  (   9)    70 AIGDLHNVNKNFSALSTKHQKKLGVDTSNFMLLGQAFLVELAAF
gi|122533  (  13)    74 ATKHLDNLREYYADLSVTHSLKFYVDPENFKLFSGIVIVCLALT
gi|122535  (  13)    74 AIKHMDDIKGYYAELSKYHSETLHVDPNNFKRFGGCLSITLGHH
gi|122536  (  13)    75 AISHIDSVKSSLQQLSKIHATELFVDPENFKRFGGVLVIVLGAK
gi|122537  (  13)    75 AIQHLDDVKSHLKGLSKSHAEDLHVDPENFKRLADVLVIVLAAK
gi|122538  (  13)    74 AVQNLDDIKNTYATLSVMHSEKLHVDPDNFRLLADCITVCVAAK
gi|122543  (  13)    70 GVQHLDDLKGTFAQLSELHCDKLHVDPENFRLLGNVLVVVLARH
gi|122544  (  13)    73 GLKQLDDLKGAFASLSELHCDKLHVDPENFRLLGNVLVVVLARR
gi|122545  (  13)    73 GLKQLDDLKGAFASLSELHCDKLHVDPENFRLLGNVLVVVLARR
gi|122546  (  13)    73 GLKQLDDLKGAFASLSELHCDKLHVDPENFRLLGNVLVVVLARR
gi|122548  (  13)    75 SIHHLDDIKNFLSVLSTKHAEELHVDPENFKRLADVLVIVLAGK
gi|122553  (  13)    74 GLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHH
gi|122554  (  13)    74 GLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHH
gi|122555  (  13)    73 GLKHLDDLKGAFAKLSELHCDKLHVDPENFRLLGNVLVVVLARH
gi|122556  (  13)    74 AVKHLDNIKGHFANLSKLHCEKFHVDPENFKLLGDIIIIVLAAH
gi|122570  (  13)    73 GMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARH
gi|122572  (  13)    73 GMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARN
gi|122573  (  13)    74 AVKNLDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIIVLAAH
gi|122574  (  13)    74 GLKHLDDLKGTYAHLSELHCDKLHVDPENFKLLGNVLVIVLARH
gi|122575  (  13)    74 AVKHLDNIKGHFAHLSKLHFEKFHVDCENFKLLGDIIIVVLGMH
gi|122576  (  13)    75 AVKNLDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIIVLAAH
gi|122581  (  13)    74 GLKHLQDLKGTFAKLSELHCDKLHVDPENFRLLGNMIVIALAHH
gi|122587  (  13)    75 AVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAH
gi|122588  (  13)    74 AVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAH
gi|122601  (  13)    75 AVKHLDNLKGTYAKLSELHCDKLHVDPENFKMLGNIIVICLAEH
gi|122613  (  13)    74 GLKHLDNLKGTFAALSELHCDQLHVDPENFRLLGNELVVVLART
gi|122614  (  13)    74 GVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARH
gi|122615  (  13)    75 GLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH
gi|122634  (  13)    74 GLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHH
gi|122671  (  13)    74 GLKNLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVCVLAHH
gi|122675  (  13)    74 GLQHLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVCVLARH
gi|122676  (  13)    75 GLSHLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVIVLSHH
gi|122678  (  13)    73 GLKHLDDLKGAFAKLSELHCDKLHVDPENFRLLGNVLVVVLARH
gi|122723  (  13)    75 AVKNLDNIKNTYAKLSELHCDKLHVDPENFRLLGDILIIVLASH
gi|122724  (  13)    75 GVKNMDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLIIVLASR
gi|122726  (  13)    75 AIKNMDNLKPAFAKLSELHCDKLHVDPENFKLLGNVMVIILATH
gi|122731  (  13)    75 AIKNMDNLKGAFAKLSELHCDKLHVDPENFKLLGNVLLIVLATH
gi|127638  (  39)   102 VKYLEFISDAIIHVLHAKHPSDFGADAQAAMSKALELFRNDMAA
gi|127647  (  12)    70 LLKARGDHAAILKPLATTHANTHKIALNNFRLITEVLVKVMAEK
gi|127648  (  39)   101 VQFLEFISEAIIQVIQSKHPGDFGGDAQAAMGKALELFRNDMAA
gi|127661  (  39)   102 VKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMAS
gi|127676  (  39)   102 VKYLEFISEIIIEVLKKRHSGDFGADAQGAMSKALELFRNDIAA
gi|127691  (  39)   101 VKYLEFISEAIIHVLHSKHPGDFGADAQAAMSKALELFRNDIAA
gi|127694  (  39)   101 VKYLEFISDAIIHVLHAKHPSNFGADAQGAMSKALELFRNDMAA
```

```
gi|13634094  (   7)    70 AVEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASH
gi|14194774  (  13)    75 GLKNMDNIEATYADLSTLHSEKLHVDPDNFKLLADCITIVLAAK
gi|14194794  (  13)    75 GMKNMDNIADAYTDLSTLHSEKLHVDPDNFKLLSDCITIVLAAK
gi|14195588  (  13)    75 GLKNMDNIVDAYAELSTLHSEKLHVDPDNFKLLSDCITIVLAAK
gi|1708122  (  13)    75 AIHNLDDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAK
gi|2506462  (  39)   101 IKYLEFISDAIIHVLHSKHPGDFGADAQGAMTKALELFRNDIAA
gi|3041678  (  13)    75 GLKHLDNLKGTFAKLSELHCDQLHVDPENFRLLGNVIVVVLARR
gi|3041679  (  13)    75 AVKNMDNLKGTFAKLSELHCDKLHVDPENFRLLGNMIVIILASH
gi|462246  (   8)    70 AVSKMDDLTAGLLELSEKHAFQLRVDPANFKLLSHCLLVVISIM
gi|462247  (  13)    74 GVKNMDDIKNTYAELSKLHSEKLHVDPDNFRLLSDCLTIVVAAK
gi|462677  (  39)   101 VKYLEFISEVIIKVIAEKHAADFGADSQAAMKKALELFRNDMAS
gi|6016192  (  13)    75 AIKNMDNLKITFAKLSELHCDKLHVDPENFKLLGNVMVIILATH
gi|6166198  (   7)    70 AVGHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH
```

Table 4. Trypsin_SER Blocks from Block Maker

BLOCK 1

```
VIGGDECDINEHPFLAFM
IVGGSATTISSFPWQISL
IVGGSATTISSFPWQISL
IVGGEDTTIGGDPYQVSL
IVGGADTSSYYTKYVVQL
IIGGTESKPHSRPYMAHL
IIGGTECKPHSRPYMAYL
IIGGVEARPHSRPYMAHL
VVGGDECNINEHPFLVAL
VVGGDECNINEHPFLVAL
IVGGFEIDVSETPYQVSL
IVGGYTCSRNSVPYQVSL
IVGGYNCEENSVPYQVSL
IVGGYTCPEHSVPYQVSL
IVGGYTCEENSVPYQVSL
IVGGYICEENSVPYQVSL
IVGGYTCQKNSLPYQVSL
IVGGYTCPKHLVPYQVSL
IVGGTDAVLGEFPYQLSF
IVGGSATTISSFPWQISL
IVGGYTCRESSVPYQVSL
IVGGYTCAANSIPYQVSL
IVGGEDANVQDHPFTVAL
IVGGYECPKHAAPWTVSL
IIGGATCAKSSVPYIVSL
IVGGREAPGSKWPWQVSL
VVGGDECNINEHRSLVAI
VIGGHPCNINEHPFLVLV
VIGGDECNINEHRSLVVL
IIGGRPCDINEHRSLALV
VIGGDECNINEHRFLALV
VIGGDECNINEHRFLVAL
VIGGDECNINEHPFLVLV
VIGGDECNINEHPFLVLV
VIGGDECNINEHPFLVLV
VIGGDECNINEHPFLVLV
IIGGDECNINEHPFLVLV
VIGGNECDINEHRFLVAF
VVGGDECDINEHPFLVAL
```

IIGGDECNINEHRFLVAL
VIGGDECNINEHRFLVAL
VIGGDECDINEHPFLAFM
IIGGVESKPHSRPYMAHL
IIGGTECKPHSRPYMAYL
IIGGTECKPHSRPYMAYL
IIGGVEAKPHSRPYMAYL
IVGGSATTISSFPWQISL
IVGGTATTISSFPWQISL
IVGGTATTISSFPWQISL
IVGGYETSIDAHPYQVSL
IIGGSDQLIRNAPWQVSI
IVNGVDTTIEAHPYQVPL
IVGGEDTTIRAHPYQVSL
IVGGQEAPGNKWPWQVSL
IVGGADTTNYHTKYVVQL
IVGGYVTDIAQVPYQITL
IVGGQEAPRSKWPWQVSL
IVGGKEAPGHKWPWQVSL
IIGGKEAPGSRWPWQVSL
IVGGQEAPRSKWPWQVSL
IVGGYTCAENSVPYQVSL
IVGGYSCARSAAPYQVSL
IVGGYSCARSAAPYQVSL
IVGGYTCPEHSVPYQVSL
IVGGFTCAKNAVPYQVSL
VIGGKPAAQNEFPFMVHL
IVGGYECTRHSQAHQVSL
IVGGYTCGANTVPYQVSL
IVGGYECTKHSQAHQVSL
IVGGFQIDIAEVPHQVSL
IVGGFEVPVEEVPFQVSL
IIGGVESRPHSRPYMAHL
IVGGYTCQENSVPYQVSL
IIGGTECKPHSRPYMAYL
IIGGDECNINEHRFLVAL
IIGGHEAKPHSRPYMAFL
IVGGFEIDVSDAPYQVSL
IVGGYECKAYSQTHQVSL
VVGGFQIDVSDAPYQVSL
IVGGYECKAYSQPHQVSL
IVGGYECRKNSASYQASL
IVGGFEIDVAETPYQVSL
IVGGVATTISSFPWQISL
IVGGFEINVSDTPYQVSL
IVGGSTTTIQQYPTIVAL
IVGGTATTISSFPWQISL
IVGGSTTTIQQYPTIVAL
IVGGSTTTIQQYPTIVAL
IVGGYETSIDAHPYQVSL
IVGGSVTTIEQWPSGSAL
IVGGTSASAGDFPFIVSI
IIGGHECAAHSRPFMASL
IVGGEMTDISLIPYQVSV
VVGGTRAAQGEFPFMVRL

BLOCK2

DKDIMLIRLDRPVKNSEHIAPLSLPSNPPSVGSVCRIMGWGAI
VNDIVIIKINGALTFSSTIKAIGLASSNPANGAAGSVSGWGTL
VNDIVIIKINGALTFSSTIKAIGLASSNPANGAAGSVSGWGTL
EYDVGILKLDEKVKETENIRYIELATETPPTGTTAVVTGWGSK
DIALVVVDPPLPLDSFSTMEAIVIASEQPPVGVQATISGWGYT
NDIAILFVDPPLALNNFTIKGIKLASEQPIEGTVSKVSGWGTT
DIMLLKLKEKANLTLAVGTLPLSPQFNFVPPGRMCRVAGWGKR
DIMLLKLKEKASLTLAVGTLPFPSQFNFVPPGRMCRVAGWGRT
DIMLLKLEEKAELTPTVDVIPLPGPSDFIDPGKMCWTAGWGKT
DKDIMLIRLRRPVTYSTHIAPVSLPSRSRGVGSRCRIMGWGKI
DKDIMLIRLRRPVTYSTHIAPVSLPSRSRGVGSRCRIMGWGKI
DFSLMELETELTFSDVVQPVSLPEQDEAVEDGTMTTVSGWGNT
DNDIMLIKLSSPATLNSRVSAIALPKSCPAAGTQCLISGWGNT
NNDIMLIKLSSRAVINARVSTISLPTAPPATGTKCLISGWGNT
NNDIMLIKLSSPVKLNARVAPVALPSACAPAGTQCLISGWGNT
DNDIMLIKLSSPAVLNARVATISLPRACAAPGTQCLISGWGNT
DNDILLIKLSSPAVINSRVSAISLPTAPPAAGTESLISGWGNT
DNDIMLIKLNSPATLNSRVSTVSLPRSCGSSGTKCLVSGWGNT
DNDIMLIKLKSPAVLNSQVSTVSLPRSCASTDAQCLVSGWGNT
DNDISLLKLSGSLTFNNNVAPIALPAQGHTATGNVIVTGWGTT
VNDIAVIRLSSSLSFSSSIKAISLATYNPANGASAAVSGWGTQ
DNDIMLIKLASPVTLNARVASVPLPSSCAPAGTQCLISGWGNT
DNDIMLIKLSSPATLNSRVATVSLPRSCAAAGTECLISGWGNT
KGFDVSVLTLEAPVKEAPIELAKADDAGYAPDTAATILGWGNT
DNDIMLIKLSKPAALNRNVDLISLPTGCAYAGEMCLISGWGNT
DNDIMLIKLSSPASLNAAVNTVPLPSGCSAAGTSCLISGWGNT
DIALLELEDPVNVSAHVQPVTLPPALQTFPTGTPCWVTGWGDV
DKDIMLIKLDSSVSNSEHIAPLSLPSSPPSVGSVCRIMGWGSI
GKDIMLIRLNRSVNNSTHIAPLSLPSSPPSQNTVCNIMGWGTI
DKDIMLIRLNRSVNNSVHIAPLSLPSSPPRLGSVCRVMGWGAI
DKDIMLIRLDSPVKNSAHIAPISLPSSPPIVGSVCRIMGWGTI
DKDIMLIRLDSPVNNSAHIAPLNLPFNPPMLGSVCRIMGWGAI
DKDIMLIRLDSPVSNSEHIAPLSLPSSPPSVGSVCRIMGWGRI
NKDIMLIRLDRPVRKSAHIAPLSLPSSPPSVGSVCRVMGWGTI
NKDIMLIRLDRPVRKSAHIAPLSLPSSPPSVGSVCRVMGWGTI
NKDIMLIRLDRPVRKSAHIAPLSLPSSPPSVGSVCRVMGWGTI
NKDIMLIRLDRPVRKSAHIAPLSLPSSPPSVGSVCRVMGWGTI
NKDIMLIRLNRPVRKSAHIAPLSLPSSPPSVGSVCRIMGWGTI
DKDIMLIKLDKPISNSKHIAPLSLPSSPPSVGSVCRIMGWGSI
DKDIMLIRLRRPVKNSAHIAPISLPSSPSSPRSRCRIMGWGKI
DKDIMLIRLDSPVKNSAHIAPLSLPSSPPSVGSVCRTMGWGRI
DKDIMLIRLDSPVKNSAHIAPLSLPSSPPSVGSDCRTMGWGRI
DKDIMLIRLNRPVKNSTHIAPISLPSNPPSVGSVCRIMGWGAI
DIMLLKLQKKAKVTASVDVISLPSPSDFINPGKVCRAAGWGRT
DIMLLKLKEKASLTLAVGTLPFPSQFNFVPPGRMCRVAGWGRT
DIMLLKLKEKAKLTLAVGTLPLPAKFSFIPPGRVCRAVGWGKT
DIMLLKLQKKAELNSDVDVISLPSSSDFIKPGKMCWTAGWGKT
VNDIAVIRLSSSLSFSSSIKAIALATYNPANGAAAAVSGWGTQ
VNDIAVIRLSSSLGFSSTIKSISLASSNPANGAAASVSGWGTQ
VNDIAVIRLSSSLSFSSTIKSISLASSNPPNGAAASVSGWGTQ
VNDIAIVRIESDLSFRSSIRAVRIADHNPREGATAVVSGWGTT
HYDIAVLRLSTPLTFGLSTRAINLASTSPSGGTTVTVTGWGHT
VNDVALIKLATPVRESSKIRYIRLADRTPPTGTPAVVTGWGTK
ENDVGILKLAEKVKETDDIRYIELATETPPTGTTAVVTGWGSK
DIALLELKNPVNISSHVHPVSLPPASETFPSGTLCWVTGWGNI

```
DIALVIVDPPLPLASSSTMEAIEIAAEQPAVGVQATISGWGYT
NDIAVLFVDPPLPLNNFTIKAIKLATEPPLDGAPSKISGWGST
DIALLELEEPVNISSHIHTVTLPPASETFPPGMPCWVTGWGDV
DIALLELEDPVNLSSHVQPVTLPPASETFPKGTRCWVTGWGDV
DIALLQLEEPVSISRHVQPVTLPPASETFPPESQCWVTGWGDV
DIALLELEEPVNISSRVHTVMLPPASETFPPGMPCWVTGWGDV
DNDILLIKLSTPAVINARVSTLLLPSACASAGTECLISGWGNT
NNDIMLIKLSKAATLNSYVNTVPLPTSCVTAGTTCLISGWGNT
NNDIMLIKLSKAATLNSYVNTVPLPTSCVTAGTTCLISGWGNT
NNDIMLIKLASAVEYSADIQPIALPSSCAKAGTECLISGWGNT
DNDIMLIKLSTTARLSANIQSVPLPSACASAGTNCLISGWGNT
YDGVGKDWALIKLAKPIDRPTLKIATTAKYNRGTFTIAGWGDV
DNDIMLIKLTEPATLNQYVHAVALPTECAADATMCTVSGWGNT
NNDIMLIKLKSAASLNSRVASISLPTSCASAGTQCLISGWGNT
NNDIMLIKLTKPATLNQYVHAVALPTECAADATMCTVSGWGNT
DFSLLELDESIGFSRSIEAIALPDASETVADGAMCTVSGWGDT
DFALLELEETVTFSDSCAPVKLPQKDTPVNEGTCLQVSGWGNT
DIMLLKLQKKAKVTPAVDVIPLPQPSDFLKPGKMCRAAGWGQT
NNDIMLIKLSSPVKLNARVATVALPSSCAPAGTQCLISGWGNT
DIMLLKLKEKASLTLAVGTLPFPSQFNFVPPGRMCRVAGWGRT
DKDIMLIRLDSPVKNSTHIEPFSLPSSPPSVGSVCRIMGWGRI
DIMLLQLTRKAEMSDAVSPINLPRSLEKVKPGMMCSVAGWGQL
DYSLLELEDELTFSDSVQPVGLPKQDETVKDGTMTTVSGWGNT
DNDIMLIKLSKPATLNTYVQPVALPTSCAPAGTMCTVSGWGNT
DFSLMELETELTFSDLVQPVELPEHEEPVEPGTMATVSGWGNT
DNDIMLIKLSKPATLNTYVQPVALPTSCAPAGTMCTVSGWGNT
DNDIMLIKLSKPASLNSYVSTVALPSSCASSGTRCLVSGWGNL
DYSLLELESVLTFSNKVQPITLPEQDEAVEDGIMTIVSGWGST
VNDIAVIRLSSSLTMSSTIKAIALTTAAPANGAAATVSGWGTT
DYALLELESELTFSDVVQPVALPEQDEAVDAGTMTIVSGWGST
DIAIMRTTSNIAFNNAAQPARIAGANYNLGDNQVVWAAGWGAI
TSDIAVLNLSSSLSFSSTIKAIGLASSNTANGAAASVSGWGTE
DIAIMRTSSNIAFNNAAQPARIAGANYNVGDNQVVWAAGWGDI
DIAIMRTASNIAFNNAAQPARIAGANYNLGDNQVVWAAGWGAI
VNDIAIIRIESDLSFRSSIREIRIADSNPREGATAVVSGWGTT
DIAILRSATTIAQNNQARPASIAGANYNLADNQAVWAIGWGAT
DLAILKLSTSIPSGGNIGYARLAASGSDPVAGSSATVAGWGAT
DLTSCSSSSTILWKVTHAVAPIPLPTSCPVAGTPCSVSGWGNT
DVALLELAEPIVMNYKTAAIELAEVGEEVETDAMAIVSGWGDT
YNGTGKDWALIKLAQPINQPTLKIATTTAYNQGTFTVAGWGAN
```

Fig. 1 Motif Enumeration: Block 1 of Globins
Identified *2358117* motifs
Score ranges from 765 to 3844.
Number of expected false positives 10^{-32} to 10^{5}

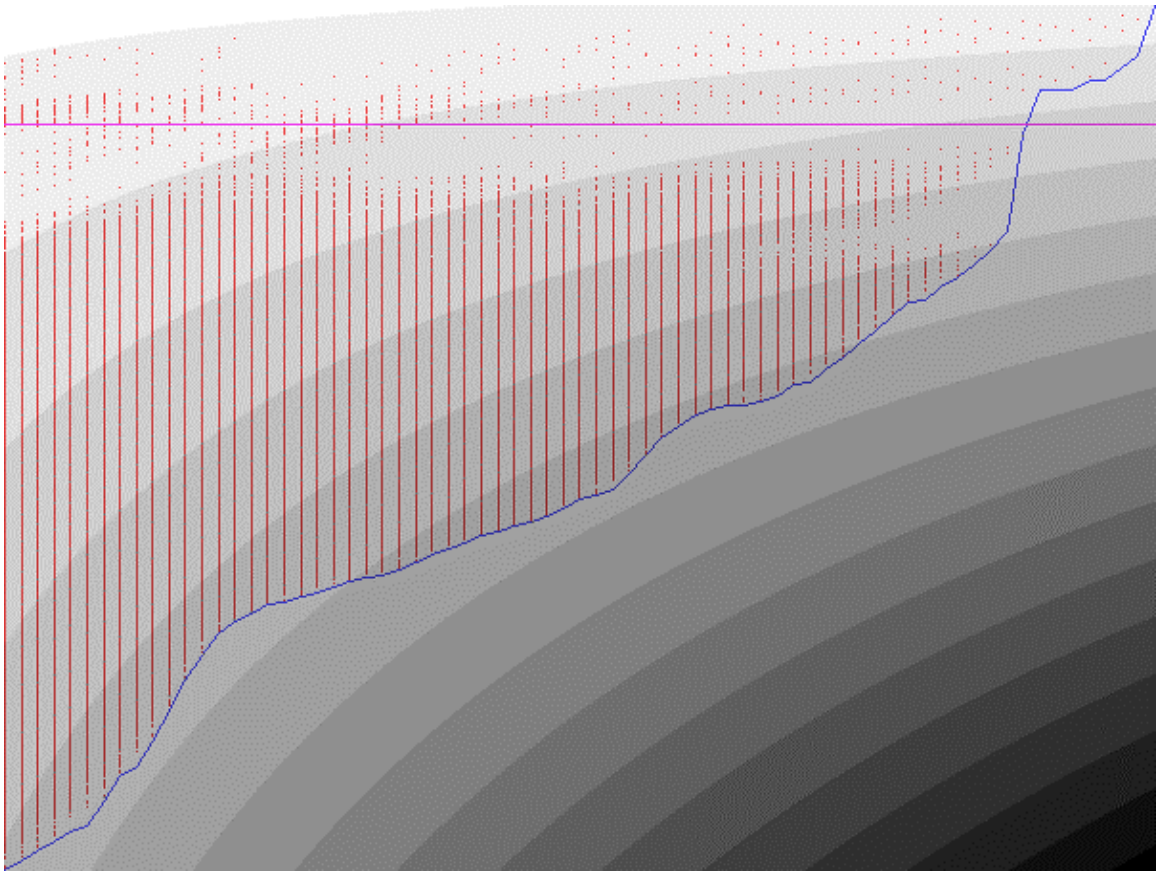


Fig 2. Motif Enumeration: Block 2 from Globins
Identified *115096* motifs
Score ranges from 3301 to 1295.
Number of expected false positives 10^{-26} to 10^{4}

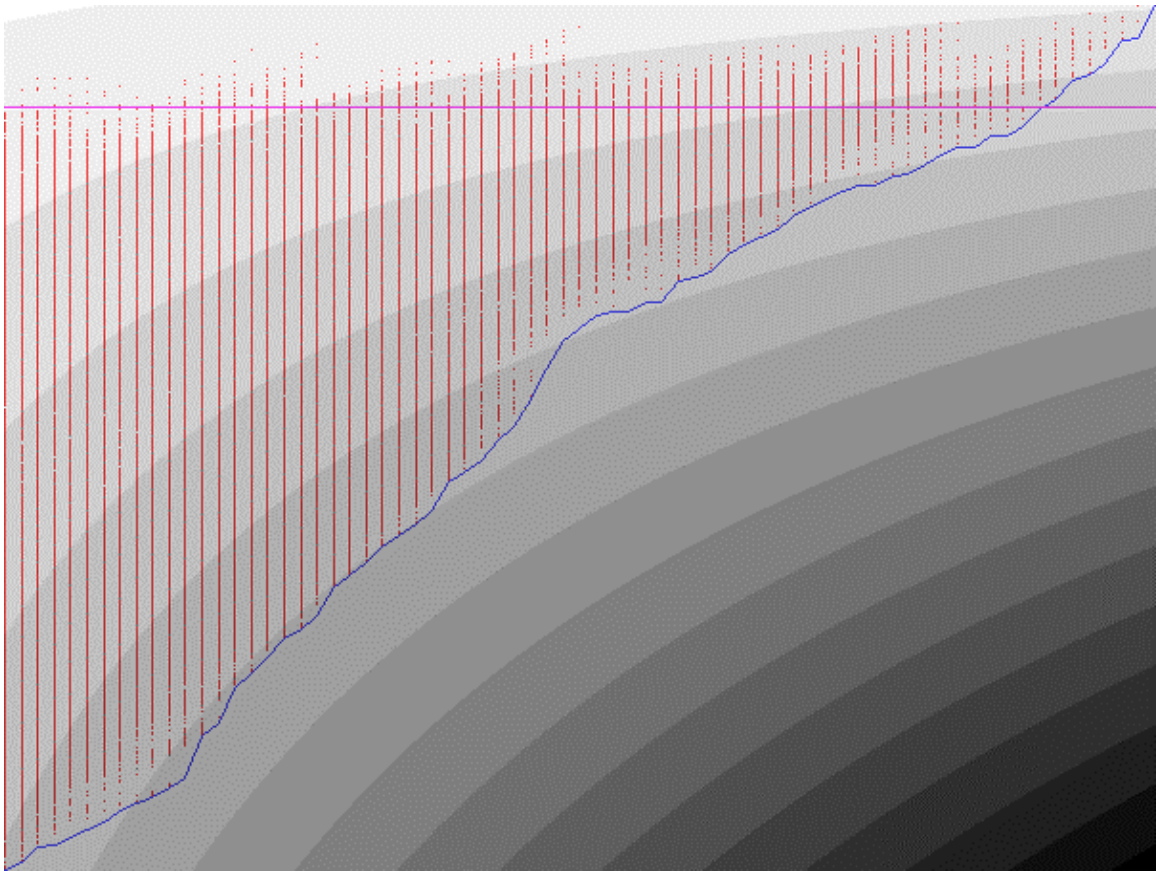


Fig 3. Motif Enumeration: Block 1 of Trypsin_SER
Identified *2764* motifs
Score Ranges from 1183 to 1098
Number of False Positives ranges from 10^{-5} to 10^{3}

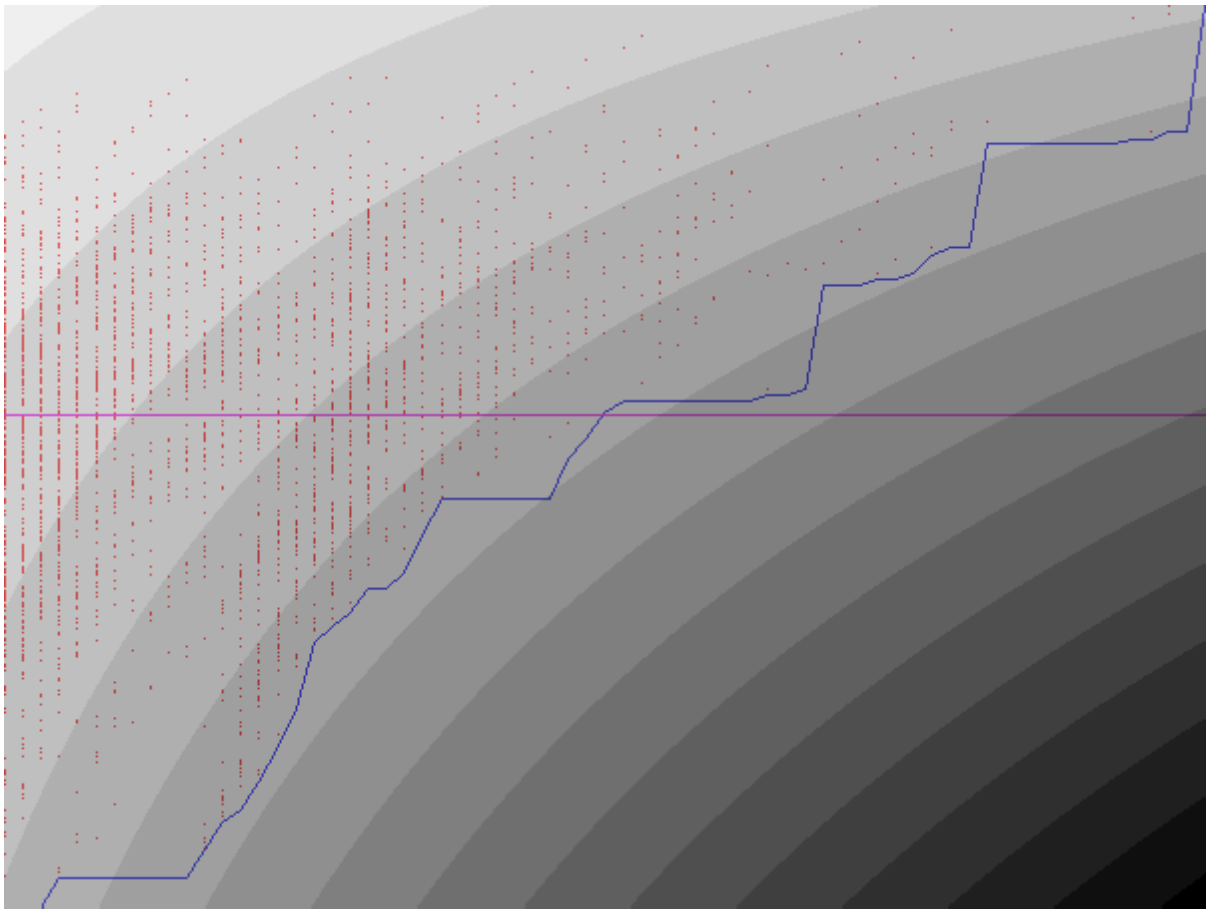
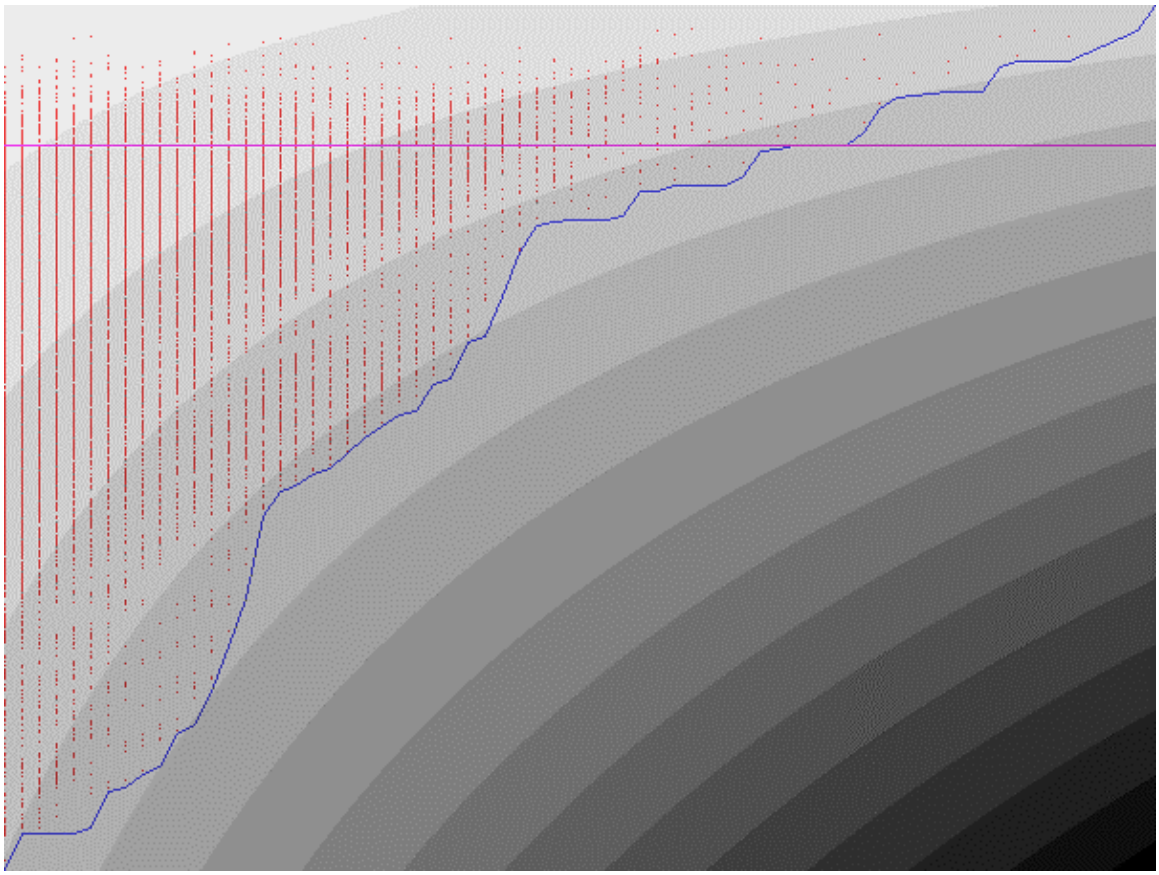


Fig 4. Motif Enumeration: Block 2 of Trypsin_SER
Identified motifs
Score Ranges from 1339 to 2144
Number of False Positives ranges from 10^{-16} to 10^{3}



MOTIFS Sampled for Globin Block 1 (with 1 being the most specific)

1. `g.ealgrll[ilv]vypwtqr[fy]f..fg[dn]ls...a[iv]..n.[kqr]v`
2. `g.[de]algrll[ilv]vypwtqr[fy]f..fg[dn]ls...a[iv]..n.[kqr]v`
6. `g.eal.rl[fly][ilv]vypwtqr[fy]f..fg[dn]ls[st]..a[iv]..n..v`
8. `g.[de]al.rl[fly][iv]vypwtqr[fy]f..fg[dn]ls...a[iv]..n..v`
16. `g..[as][ilv].r[ilmv][fly][ilv]v[fy]pwt.r[fy]f..fg[dn][ilmv]s...a[iv]..n..v`
30. `g..[ast][ilv]...[filmv]..[fy]pwt.r.f..fg[dn][ilmv]....[as][ilv]..n..[iv]`
51. `g..[ast][ilv]...[filmv]..[fy]p.t...f..f........[iv]......`

MOTIFS Sampled for Globin Block 2

1. `.[ilmv]..[ilmv]d[dn][ilv]k.[ast][fy][as].lselhcdklhvdpenf[kr]llg[dn][ilmv][ilmv][ilv].[iv]l[as]..`
2. `.[ilmv]..[ilmv]d[dn][ilv]k.[ast][fy][as].lselhcd[kqr]lhvdpenf[kr]llg[dn][ilmv][ilmv][ilv].[iv]l[as]..`
6. `.[ilmv]..[ilmv]d[dn][ilv]k.[ast][fy][as].lselhcd[kqr]lhvdpenf[kr]llg[dn].[ilmv][ilv].[iv]l...`

10. [ilmv]..[ilmv]d.[ilv]k.[ast][fy][as].lselhcd[kqr]lhvdp.nf[kr][ilmv]lg[dn].[ilmv][ilv]..l...
20. .[ilmv]..[ilmv]d[dn][ilmv]..[ast][fly][as].ls[de]lh..[kqr]l.vdp.nf[kr]ll...[ilmv][ilv]..la..
40. ....[ilmv]d[dn][ilmv]..[ast][fly]..ls.[ilmv]h...l.vdp.nf[kqr][filmv]l...[filmv]...[filmv]...
58. ....[ilmv]..[ilmv]...[fly]..ls..h...[fly].vd..nf..[fly]...[filmv]...[filmv]...

MOTIFS Sampled for Trypsin_SER Block 1

1. ivgg........p[fwy]qvsl
2. ivgg........p[fwy]q[iv]sl
8. [iv][iv]gg........p[fwy]q[ilv][st][filmv]
12. [iv][iv]gg........p[fwy].[iv][as][ilv]
16. [iv][iv]gg........p[fwy].[ilv].[film
20. [iv][iv]gg........p[fwy]...[ilmv]
24. [iv][iv]gg........p....[filmv]
27. [iv][iv]gg............[ilv]

MOTIFS Sampled for Trypsin_SER Block 2

1. [dn].dimli[kr]l.........[iv]..[filmv].lp[st].....g[st].c.[iv].gwg..
6. [dn].dimli[kr]l.........[iv]..[filmv].lp.......[st].c.[iv].gwg..
9. [dn].dimli[kr]l.........[iv]..[filmv].lp.......[ast].c...gwg..
16. ..di.[ilv]i[kr][ilv].........[iv]..[ilv].l........[ast]...[iv].gwg..
25. ..d[iv].[ilv][ilv][kr][ilv].........[iv]..[ilv].[ilv].............gwg..
30. ...[ilv]..[ilv][ekqr].............[ilv].........g....[iv].gwg..
36. ......[filmv].............[filmv].........g....[iv].gwg..
43. ......[filmv].............[filmv].............[iv].gwg..