picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 15 May 2022
Manuscript Category

OXFORD

Subject Section

Building a Hidden Markov Model to Predict Bpti/Kunitz Region in Unknown Proteins

Maryam Mohammadi [1], Emidio Capriotti [2,*]

[1] Department of Biotechnology and Pharmacy, University of Bologna, Bologna, Italy

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Kunitz domains are the active domains of proteins that inhibit the function of degrading enzymes. They are short in size with a length of about 50 to 70 amino acids and a molecular weight of 6 kDa. The trypsin inhibitor was initially extracted from soybeans, and now these Kunitz domains are used as a based to produce new pharmaceutical drugs.The famous proteins such as aprotinin protein (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI) has this domain; and this protein is also remarkable for its good stability of the molecule is due to the three disulfide bonds linking the six cysteine of the chain (Cys5-Cys55, Cys14-Cys38, and Cys30-Cys51).
Results: To predict the Kunitz domain in unknown proteins we generate a Hidden Markove Model (HMM) by available structural information of Kunitz type protein. The HMM is a kind of probabilistic model that can calculates the indels probabilities. The overall performance of the generated model in the threshold of 1e-6 with accuracy of 0.999 and Matthew Correlation Coefficient of 0.991.
Contact: Maryam.Mohammadi3@studio.unibo.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Bpti/Kunitz Bovine pancreatic trypsin inhibitor is an extensively studied model structure. This is a highly selective inhibitor of factor Xa in the blood coagulation pathways. Its molecules are highly dipolar and are arranged to form a twisted two-stranded antiparallel beta sheet followed by an alpha helix.[1] the Kunitz/Bovine pancreatic trypsin inhibitor family inhibits proteases of the S1 family[2] and are restricted to the metazoa with a single exception. They have a short length sequence (about 50 to 70 amino acid residues) with the weight of 6 kD and are included as alpha/beta proteins with few secondary structures with two antiparallel beta-sheets and one or two helical regions that are stabilized with three disulfide bridges which has an important rule in the protein stability (Figure 1). One of the famous proteins which contain Kunitz domain is 3TGI. It is a single amino acid residue chain with three disulfide bridges, solved through X-Ray diffraction with a resolution of 0.9 Å. [3], and it has a Pharmaceutical useage. Previously available under the name Trasylol (Mile).

They are used for inhibiting coagulation so as to reduce blood loss during bypass surgery. The family of BPTI[4] (or basic protease inhibitor) includes a lot of members,[5] such as snake venom protease; mammalian inter-alpha-trypsin inhibitors; trypsin, a rat mast cell inhibitor of trypsin; a domain found in an alternatively spliced form of Alzheimer's amyloid beta-protein; domains at the C-terminal of the alpha one and alpha three chains of type VI and type VII collagens; tissue factor pathway inhibitor precursor; and Kunitz STI protease inhibitor contained in legume seeds. Kunitz domains are stable as standalone peptides are able to recognize specific protein structures and also work as competitive protease inhibitors in their free form. These properties have led to attempts at developing pharmaceutical drugs of Kunitz domains. One of the drugs which have been made by this kind of protein is the kallikrein inhibitor ecallantide, which is used for the treatment of hereditary angioedema.[6] Another drug example that is made by this protein is depelestat, which is used as an inhibitor of neutrophil elastase and has undergone Phase II clinical trials for the treatment of acute respiratory distress syndrome [7] and has also been described as a potential inhalable cystic fibrosis treatment.

A crucial step in blood coagulation is the conversion of prothrombin into its active form, thrombin, and it leads to the formation of clots that impede blood loss. [1] The reverse process is fibrinolysis which prevents blood clots from becoming problematic, and they break down the fibrin clot. Its enzyme is in the blood, and it degrades many blood plasma

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

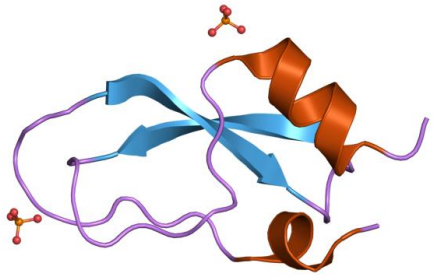

Fig. 1. Structure of BPTI. In this structure Alpha helices are shown in orange and Beta sheets blue. As can be seen in the picture, the conserved disulfide bonds are shown in back and bone structure

proteins and cuts the fibrin mesh. [8] because it is the main mechanism in cardiovascular disease, so an unregulated process can lead to undesirable effects in thrombotic and atherosclerotic and in the response and finally in injury. Therefore this protein which is famous as a protease inhibitor, play an important role in preventing fibrinolysis; for this reason, it has been investigated very well and is called protease inhibitors. This protein generally contain one or more functional regions which the Kunitz domains are the active domains of those proteins that prevent access of the serine protease to its physiological substrate through the insertion of a residue into the active site cleft(Arg-24 in BPTI Kunitz domain, that is frequently used as a model of this family.) The pancreatic trypsin inhibitor has a low relative molecular mass, a basic isoelectric point, and one or several inhibitory domains with a broad spectrum of activity toward serine proteases. [2]

2 Methods

2.1 Datasets

In order to generate a Hidden Markov model which is able to predict the proteins that contain the Kunitz domain, we need some proteins set, training set, and test set such as the following:

- A training set of the proteins which has the Kunitz domain or Pfam pf00014, with the length of 40-80 aminoacids and are manually reviewed in the Swissprot database with the total number of 153 has been created. (Table S1, supplementary material).
- A positive set of the proteins which has the Kunitz domain or Pfam pf00014, with the minimum length of 40 to * and are manually reviewed in the Swissprot database with the total number of 363 has been created.
- A negative of the proteins which doesn't have the Kunitz domain or Pfam pf00014, with the minimum length of 40 to * and are manually reviewed in the Swissprot database with the total number of 557287 has been created.

To build a model, we need a series of data about 30 proteins that represent proteins containing the Kunitz domain, and we call them sequences. These sequences are the best sequences with which we can build the Hidden Markov model. The data to produce the training set should extract from the protein data bank (PDB) with the total number of 153. The restriction which has been applied on the protein data are: having good quality or resolution, below 3 angstroms, the length and size of protein (Kunitz domain has the length of 50-60 amino acids) which we can apply a little larger length with polymer entity sequence length such as 40-80 (To avoid the existence of structure which contains multiple domains in proteins with large lenght). Finally by producing the custom report with selecting the options of PDB id/entity id (polymer identity)/auth asym chain/sequence/resolution /polymer entity sequence length, we generate clean-pdb.seq to use it for finding common identifiers (2.1 linux command, supplementary material). Then, we should start searching from a seed protein within a normal structure (3TGI) which has the Kunitz domain, and get back with a list of potential seed proteins, which are a clean set of proteins that contain contains the Kunitz domain in the PDBeFOLD website. By generating this two list of proteins which has Kunitz domain, now we should apply a comparison between them to identify the common proteins (153 common proteins or identifier) and finding the related chain with BPTI KUNITZ domain (because we don't know which domains of PDB proteins related to kunitz domain) (2.1 linux command, supplementary material) Now we need some proteins which be representative of all of these 153 identifiers. So, by using the cd-hit online website and applying some restrictions such as cut-off value 0.95, minimal alignment coverage for the shorter sequence 0.8(to avoid the redundancy), the clustering of proteins based on their similarities which lead to 37 clusters, has been done. Now, based on different options such as length, resolution, and etc. we can select the representative for every cluster. Here the representative proteins of every cluster have been selected by the best and lowest resolution, which has been shown in the table in supplementary materials. (Table S2 supplementary material) To obtain the cdhit-cluster List first multiple alignment (Figure 2) should be done. Once the representative proteins of each cluster and their chain extracted and we did the multiple alignment by yanglab online website by using the fasta file of multiple alignment we can produce the hmm model by hmmer package. Before building a model, it is better to cut out the C terminal and N terminal of the multiple alignment fasta files, which contains gaps, and then building the model. (2.6 linux command, supplementary material) So we derived a 37 clusters of 153 representative proteins that contain the Kunitz domain, with a similar shape to the one of 3TGI Kunitz protein, with the best resolution and all approximately of the same length. (Table S3, supplementary material).

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

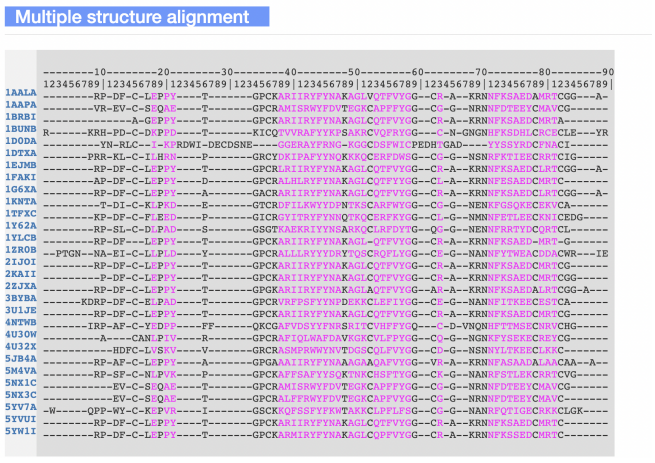

Fig. 2. mutliple alignment by MTM-aln website

2.2 Generating Hidden Markov Model (Training Set)

In order to build a Hidden Markov Model we used HMMER3.3.2. package [11] HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models.This useful package is used to search sequence databases to find homolog sequences. It implements methods by using probabilistic models which called hidden Markov models profiles. HMMER is often used with a profile database, such as Pfam or many of the databases that participate in Interpro. [12] In the past, this strength came at a significant computational expense, but now with thanks to this good package, finding homologs has become easier. By using hmmbuild command in HMMER is possible to build a profile HMM from a multiple sequence alignment, which means it reads a multiple sequence alignment file and builds a new profile HMM, and saves the HMM in hmmfile.(2.8 linux command, supplementary material) Then, by using the hmmsearch command, the model can search a sequence database with a profile HMM means search profile HMMs against a sequence database, and by using hmmlogo, it can be generated an HMM logo from an HMM file (Figure 3)

3 Prediction

In order to predict the performance of the hidden Markov model first we should create the set-all.res group, which contains all of the negative and positive results, and it has been built by hmmsearch (to verify whether our model will be able to predict the positive and negative set of the Kunitz family correctly) (2.13 linux command, supplementary material).

Once we obtained the positive and the negative testing sets with the name of negatives.ids and postives.ids first we sort all of them (2.10 linux command, supplementary materials) and then we divide both these groups to half with the name of pos1.ids post2.ids neg1.ids and neg2.ids (2.11 linux command supplementary material). It the time that we should get the fasta format of this four groups with the python scripts and create the files of set-pos1.fasta set-pos2.fasta set-neg1.fasta neg2.fasta. (2.12,2.13 linux commands, supplementary materials) With having the Fasta format of these groups, we performed the hmmsearch, and it returned an output where for each protein sequence, there were reported three different values: E-value, score and bias, for both the full sequence and the domain which in the supplementary material we have reported the negative groups and its sequence E-value and domain E-value. Then we created a file (set-all1.res, 2.14, 2.15 linux commands, supplementary materials) with all the hints about the positive set and the negative set with the respective E-value of the sequence. Finally, we associated class 1 with all the proteins belonging to the positive set and class 0 with all the proteins belonging to the negative one and we create the set-all1.res and set-all2.res. In the set-all.res we also added all the proteins included in the negative sets and positive sets that were not scored by the Hmm search. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

3.1 Calculating the performance of the training model

In order to measure the performance of the method, we use a python script (python script 3, supplementary materials) which can measure all of the parameters of the Matthew correlation coefficient, accuracy, confusion matrix (Figure 3, confusion matrix), true positive rate, and false-positive rate. All of the calculated parameters are shown in the supplementary materials in different e-value. (Table S6, S7,S8, supplementary material)

- MCC: Matthew Correlation Coefficient is a statistical parameter that produces a high score only if the prediction gives good results. While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the Matthews correlation coefficient is generally regarded as being one of the best such measures. TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. MCC returns a value between 1 and 1. A coefficient of +1, which represents a perfect prediction, 0 no better than random prediction (wrong prediction), and 1 indicates total disagreement between prediction and observation (completely wrong prediction).
- Accuracy is the degree of closeness to the true value. As we know, in this matrix, the sensitivity and accuracy in comparison to the overall result are so important. Accuracy calculates based on this formula: ACC = (TP +TN)/(TP +FN +TN +FP)
- Confusion Matrix (CM): The confusing matrix, despite its simple logic and structure, is a powerful concept that can provide comprehensive information on how classification works in a variety of research alone. two clusters in comparison of the whole set More commonly, it is a description of systematic errors and measures the statistical bias; low accuracy causes a difference between a result and a "true" value. that is, the accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.
- TPR: True positive rate or sensitivity is a statistical parameter that calculates how many of the positive predictions are correctly predicted

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

	Predicted Condition	
Actual Condition	True Positive Correct Accept	False Negative Type 2 error Understimation
	False Positive Type 1 error overestimation	True Negative Correct Rejection

Fig. 3. Confusion Matrix (CM)

in real. The True positive rate (TPR) gives the proportion of correct predictions in predictions of the positive class. TPR= TP /(TP+FN)

- FPR The false positive rate is another statistical parameter that calculates the ratio between the number of negatives that wrongly predicted positive (false positives) the Konitz domain. In this parameter, the number of proteins that have been mistakenly identified as containing the Konitz domain is not included in the calculations and has no effect on this parameter.

4 Result and Discussion

Today, using of hidden Markov models (HMM) has been widely applied with the aim of representing protein clusters. To find the homolog proteins which has Kunitz domain, we should use a few related sequences to build a hidden Markov model, which finally we optimize the model in special threshold with good Mathew correlation coefficient and good accuracy to use it for annotating. The generated Kunitz training model is able to perfectly find the sequences belonging to the Kunitz family and neglect the non-Kunitz proteins. Finally, we evaluated the performance of our results in terms of accuracy and MCC. By this Kunitz HMM profile, we can do good classification automatically, and because of its good performance, it can be used to annotate the unknown proteins which haven't been clusterized yet. Once Kunitz proteins HMM profile is generated, we can produce the hmm-logo by Skylign online website (a tool for creating logos representing both sequence alignments and profile hidden Markov models) in order to represent the model graphically(Figure 4) [13].

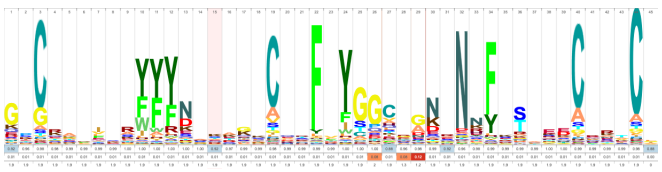

Fig. 4. HMMlogo of Kunitz protein which produced by skylign based on hmm model

Different families can be compared visually based on HMM Logos picture[14]. The model logo is shown in Figure 2. The height of letters is determined by the deviation of the position's letter emission frequencies from the background frequencies domain, and each amino acids amino acid which contains more conserved is shown with a longer height. In this picture produced, it can be seen that the conservation of the cysteines is very well which their duty is to make the stability of the protein.

4.1 Optimization and performance

To calculate the performance of the training set model, we use the cross-validation between two sets of set-all1.res and set-all2.res. In order to do cross-validation, a first different e-value threshold has been applied on the model for the set of set-all1.res and set-all2.res, and two measure the ACC, MCC, FPR, TPR, and the confusion matrice parameters (Figure 5, the result has been shown in table and aslo the supplementary materials S5, S6, S7).

Optimum E-value of set_all1 and set_all2							
Group Name	Applied E-value	ACC	MCC	CM	FPR	TPR	wrong prediction
set_all1.res	1.00E-09	0.999978481	0.983505717	[[278640,3],[3,179]]	0.016483516	0.983516484	P0DV04 P0DV06 P17726 Q11101 O62247 D3GGZ8
set_all2.res	1.00E-06	0.999989241	0.991897361	[[278641,0],[3,183]]	0.016129032	0.983870968	P0DV05 P0DV03 P36235

Fig. 5. Optimum e-value of each set with calculating ACC, MCC, CM, FPR, TPR

Group Name	Applied E-value	ACC	MCC	CM	FPR	TPR	wrong prediction
set_all1.res	1.00E-09	0.999978481	0.983505717	[[278640,3],[3,179]]	0.016483516	0.98351648	P0DV04 P0DV06 P17726 Q11101 O62247 D3GGZ8
set_all2.res	1.00E-06	0.999989241	0.991897361	[[278641,0],[3,183]]	0.016129032	0.98387097	P0DV05 P0DV03 P36235

Fig. 6. cross-validation based on optimum evalue of each sets. Wrong prediction of models is shown in the table with its e-value.

Once we sort the threshold based on the best MCC and we found the obtained optimum value of set-all1.res and the set-all2.res, we do cross-validation to find the wrong predictions and performance of the model. Here the optimum threshold of both sets with the wrong prediction (Figure 6, the result is shown in the table) is represented in the table and, all of the results for different e-value are shown in detail in the supplementary materials. (Table S5, S6, S7, linux command 2.16-2.21, python script 3, supplementary materials). As can be seen in the table, the optimal e-value threshold for set-all1.res is 1e-06 which shows the good values of ACC 0.9999 and MCC near one, 0.992 with FPR 0.01, which offers a good performance of the model. to do cross-validation and finding the false prediction we should set this threshold on set-all2.res and calculate the statistical parameters. the wrong predicted proteins in set-all1.res are

- False Negatives(FP): D3GGZ8,O62247 and Q11101 (Figure 7, Figure 8)that the first two proteins correspond to the same protein, with the name of Kunitz-type protein bli-5. D3GGZ8 is annotated by Pfam as Kunitz-type (PF00014) but not recognized by the HMM profile that it is poorly annotated (inferred from homology) but O62247 is

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

expressed by Caenorhabditis elegans and has a very good annotation score, experimental evidence at protein level . This two proteins are orthologues that appears to have serine protease activity, but is uncertain if this activity since it seems that they lack in some features of the Kunitz-type proteins. So may be this protein face to changes and these changes lead to loss of function. This may be one of the possible reasons why the HMM profile did misclassification for them. Moreover, the Q11101 doesn't have a high score of annotation and its function has been inferred by homology and this sequence share a low sequence identity with our set of representative proteins used to build the HMM and this can also be due to their evolutionary distance.

- False Positives(FN): one part of the false positives related to protein family of PI-stichotoxin such as P0DV04, P0DV06, P0DV05 and P0DV03 that they have been clustered as bpti kunitz domain in prosite with PS00280 , PS50279 ID and they haven't classified by Pfam as Pfam as Kunitz-type even if they contain the domain, so for this reason they haven't been detected by our model. the other part of the false positives result relate to two other proteins with PDB ID P36235 (Disagregin protein) and P17726 (Tick anticoagulant peptide protein) which they haven't annotated by Pfam but it has clustered as bpti kunitz domain in INTERPRO with the codes of IPR036880. It is noteworthy that we say because all of these six proteins hasn't clustered by Pfam as Kunitz-type family so even if they contain the domain, consequently, they are recognized as false positives in our result.

CM set_all1	ACTUAL VALUE	
PREDICTED VALUE	278640	3
	3	179

Fig. 7. Confusion Matrix of set-all1.res. The results show that there is 3 False Negative and 3 False Positives in this set.

CM set_all2	ACTUAL VALUE	
PREDICTED VALUE	278641	0
	3	183

Fig. 8. Confusion Matrix of set-all2.res. The results show that there is 3 False Positives in this set.

5 Conclusion

This report simply shows that Hidden Markov Model profiles perform well to classify bpti kunitz domains of proteins and the wrong predicted related to inconsistencies in the databases annotation of the specific domain under examination.

References

[1]Bardet, G. (1920) Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire. doi:10.1016/0014-5793(94)00941-4. PMID 7925983. S2CID 2280234.

[2]Rawlings ND, Barrett AJ, Tolle DP (2004). "Evolutionary families of peptidase inhibitors". *Biochem. J. 378 (Pt 3)* doi:10.1042/BJ20031825. PMC 1224039. PMID 14705960.

[3]St Charles R, Padmanabhan K, Arni RV, Padmanabhan KP, Tulinsky A (2000). "Structure of tick anticoagulant peptide at 1.6 A resolution complexed with bovine pancreatic trypsin inhibitor". *Protein Sci.* doi:10.1110/ps.9.2.265. PMC 2144540. PMID 10716178.

[Wlodawe.,1991]Wlodawer A, Housset D, Kim KS, Fuchs J, Woodward C (1991). "Crystal structure of a Y35G mutant of bovine pancreatic trypsin inhibitor".*J. Mol. Biol.* doi:10.1016/0022-2836(91)90115-M. PMID 1714504.

[4]Salier JP (1990). "Inter-alpha-trypsin inhibitor: emergence of a family within the Kunitz-type protease inhibitor superfamily". *Trends Biochem* 15 (11): 435–439. doi:10.1016/0968-0004(90)90282-G. PMID 1703675.

[5]Lehmann, A (2008). "Ecallantide (DX-88), a plasma kallikrein inhibitor for the treatment of hereditary angioedema and the prevention of blood loss in on-pump cardiothoracic surgery".*xpert Opinion on Biological Therapy* 8 (8): 1187–99. doi:10.1517/14712598.8.8.1187. PMID 18613770. S2CID 72623604.

[6]Takahashi K, Ikeo K, Gojobori T (1992). "Evolutionary origin of a Kunitz-type trypsin inhibitor domain inserted in the amyloid beta precursor protein of Alzheimer's disease".*J. Mol. Evol.* 34(6): 536–543. doi:10.1007/BF00160466. PMID 1593645. S2CID 26698630.

[7]S Eddy – (1992)- HMMER user's guide , eddylab.org

[8]A Bateman, DH Haft (2002). HMM-based databases in InterPro - *Briefings in bioinformatics* doi. 10.1093/bib/3.3.236

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

[9]B Schuster-Böckler, J Schultz, S Rahmann (2004) HMM Logos for visualization of protein families *BMC bioinformatics* doi.10.1186/1471-2105-5-7

[10]TJ Wheeler, J Clements, RD Finn, (2014) Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models *BMC bioinformatics* doi. 0.1186/1471-2105-15-7

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture