

Project Description

**Laboratory of Bioinformatics I
Module 2**

29 March, 2019

Emidio Capriotti

<http://biofold.org/>



**Biomolecules
Folding and
Disease**

Department of Pharmacy
and Biotechnology (FaBiT)
University of Bologna



Main Aim

Building a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain.

Kunitz domains are the active domains of proteins that **inhibit the function of protein degrading enzymes** or, more specifically, domains of Kunitz-type are **protease inhibitors**.

Examples of Kunitz-type protease inhibitors are aprotinin (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI).

Aprotinin

The drug **aprotinin** (Trasylol, previously Bayer and now Nordic Group pharmaceuticals), is the small protein **bovine pancreatic trypsin inhibitor** (BPTI), an **antifibrinolytic** molecule that inhibits trypsin and related proteolytic enzymes. Under the trade name Trasylol, aprotinin was used as a medication administered by injection **to reduce bleeding** during complex surgery, such as heart and liver surgery. Its main effect is the slowing down of fibrinolysis, the process that leads to the breakdown of blood clots. The aim in its use was to decrease the need for blood transfusions during surgery, as well as end-organ damage due to hypotension (low blood pressure) as a result of marked blood loss.

BPTI is the classic member of the protein family of Kunitz-type serine protease inhibitors. Its physiological functions include the **protective inhibition of the major digestive enzyme trypsin** when small amounts are produced by cleavage of the trypsinogen precursor during storage in the pancreas.

Aprotinin Structure

Aprotinin is a **monomeric** (single-chain) globular polypeptide derived from bovine lung tissue. It has a molecular weight of 6512 and consists of a chain 58 residues long that folds into a **stable, compact tertiary structure of the 'small SS-rich' type, containing 3 disulfides, a twisted β -hairpin and a C-terminal α -helix.**

There are 10 positively-charged lysine (K) and arginine (R) side chains and only 4 negative aspartate (D) and glutamates (E), making the protein strongly basic

The high stability of the molecule is due to the **3 disulfide bonds linking the 6 cysteine members of the chain (Cys5-Cys55, Cys14-Cys38 and Cys30-Cys51).**

The long, basic lysine 15 side chain on the exposed loop binds very tightly in the specificity pocket at the active site of trypsin and inhibits its enzymatic action. BPTI is synthesized as a longer, precursor sequence, which folds up and then is cleaved into the mature sequence given above.

Start from the Structure

In the Protein Data Bank the crystal of 3TGI a complexed of the BPTI




[Summary](#) [3D View](#) [Sequence](#) [Annotations](#) [Seq. Similarity](#) [3D Similarity](#) [Literature](#) [Biol. & Chem.](#) [Methods](#) [Links](#)

WILD-TYPE RAT ANIONIC TRYPSIN COMPLEXED WITH BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI)


DOI:10.2210/pdb3tgi/pdb


Primary Citation


Comparison of anionic and cationic trypsinogens: the anionic activation domain is more flexible in solution and differs in its mode of BPTI binding in the crystal structure.


Pasternak, A. , Ringe, D. , Hedstrom, L. 

Journal: (1999) Protein Sci. **8**: 253-258

PubMed: 10210204 

PubMedCentral: PMC2144100 




DOI: 10.1110/ps.8.1.253 

Search Related Articles in PubMed 


PubMed Abstract:


Unlike bovine cationic trypsin, rat anionic trypsin retains activity at high pH. This alkaline stability has been attributed to stabilization of the salt bridge between the N-terminal Ile16 and Asp194 by the surface negative charge (Soman K, Yang A-S, Honig B, Fletterick R., 1989, Biochemistry 28:9918-9926). The formation of this salt bridge controls the conformation of the activation domain in trypsin. In this work we probe the structure of rat trypsinogen to determine the effects of the surface negative charge on the activation domain in the absence of the Ile16-Asp194 salt bridge. We determined the crystal structures of the rat trypsin-BPTI complex and the rat trypsinogen-BPTI complex at 1.8 and 2.2 Å, respectively. The BPTI complex of rat trypsinogen resembles that of rat trypsin. Surprisingly, the side chain of Ile16 is found in a similar position in both the rat trypsin and trypsinogen complexes, although it is not the N-terminal residue and cannot form the salt bridge in trypsinogen. The resulting position of the activation peptide alters the conformation of the adjacent autolysis loop (residues 142-153). While bovine trypsinogen and trypsin have similar CD spectra, the CD spectrum of rat trypsinogen has only 60% of the intensity of rat trypsin. This lower intensity most likely results from increased flexibility around two conserved tryptophans, which are adjacent to the activation domain. The NMR spectrum of rat trypsinogen contains high field methyl signals as observed in bovine trypsinogen. It is concluded that the activation domain of rat trypsinogen is more flexible than that of bovine trypsinogen, but does not extend further into the protein core.

3TGI

 Display Files ▾
 Download Files ▾
 Download Citation ▾

Biological Assembly



 **3D View** [More Images...](#)

Stoichiometry: **Hetero 2-mer - AB**

Biological assembly 1 assigned by authors and generated by PISA (software)

Downloadable viewers:

[Simple Viewer](#) [Protein Workshop](#)

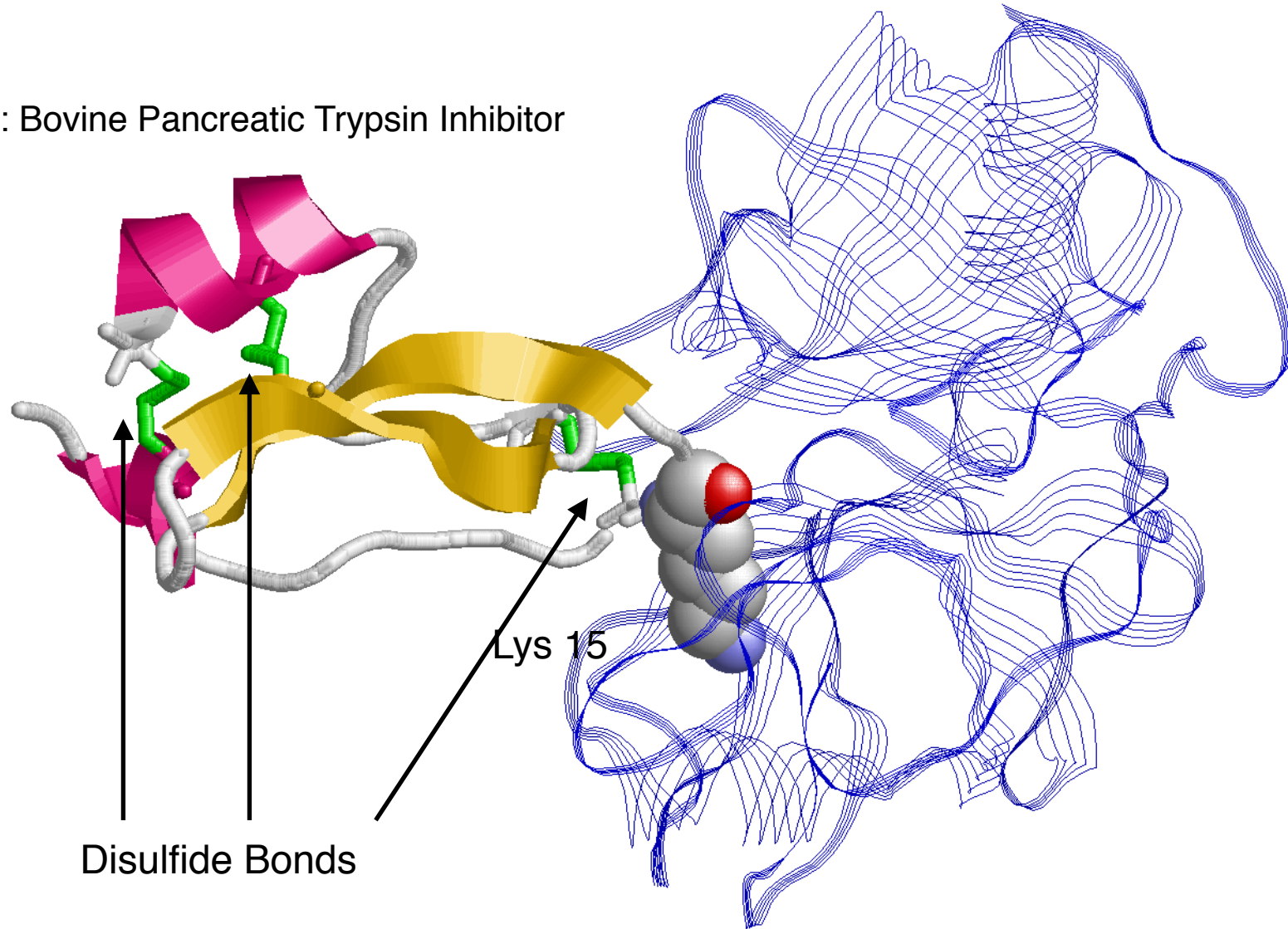
[Kiosk Viewer](#)

[MyPDB Personal Annotations](#) [Hide](#)

Structure Analysis

Chain E: Rat Anionic Trypsin

Chain I: Bovine Pancreatic Trypsin Inhibitor



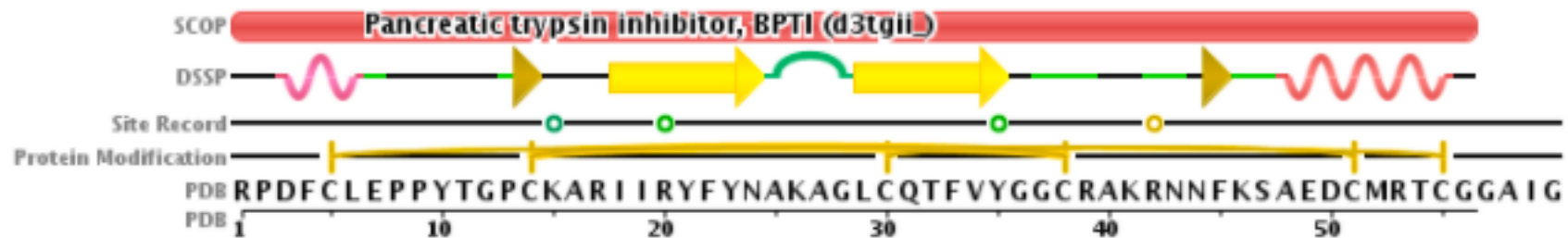
The Protein Fold

The **structure is a disulfide rich alpha+beta fold**. Bovine pancreatic trypsin inhibitor is an extensively studied model structure.

The majority are restricted to metazoa with a single exception: *Amsacta moorei entomopoxvirus*, a species of poxvirus.

They are short (about 50 to 60 amino acid residues) alpha/beta proteins with few secondary structures. The fold is constrained by three disulfide bonds.

Sequence Chain View



Annotation

In UniProt we found the information about the function and important sites

UniProtKB - P00974 (BPT1_BOVIN) Basket

Display

Entry

Publications

Feature viewer

Feature table

None

Function

Names & Taxonomy

Subcellular location

Pathology & Biotech

PTM / Processing

Expression

BLAST

Align

Format

Add to basket

History

Feedback

Help video

Other tutorials and videos

Protein

Gene

Organism

Status

Pancreatic trypsin inhibitor

N/A

Bos taurus (Bovine)

Reviewed

 - Annotation score: ●●●●● - Experimental evidence at protein levelⁱ

Functionⁱ

Inhibits trypsin, kallikrein, chymotrypsin, and plasmin.

Sites

Feature key	Position(s)	Description	Actions	Graphical view	Length
Site ⁱ	50 – 51	Reactive bond for trypsin			2

PTM / Processingⁱ

Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Signal peptide ⁱ	1 – 21	<div>Sequence analysis</div>	<div>Add</div> <div>BLAST</div>		21
Propeptide ⁱ (PRO_0000016852)	22 – 35		<div>Add</div> <div>BLAST</div>		14
Chain ⁱ (PRO_0000016853)	36 – 93	Pancreatic trypsin inhibitor	<div>Add</div> <div>BLAST</div>		58
Propeptide ⁱ (PRO_0000016854)	94 – 100				7

Amino acid modifications

Feature key	Position(s)	Description	Actions	Graphical view	Length
Disulfide bond ⁱ	40 ↔ 90				
Disulfide bond ⁱ	49 ↔ 73	<div>PROSITE-ProRule annotation</div> <div>1 Publication</div>			
Disulfide bond ⁱ	65 ↔ 86	<div>PROSITE-ProRule annotation</div> <div>1 Publication</div>			

PFAM

The Kunitz BPTI family is described in PFAM database



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)



Family: *Kunitz_BPTI* (PF00014)

836 architectures 11819 sequences 7 interactions 275 species 259 structures

Summary

Domain organisation

[Clan](#)

[Alignments](#)

[HMM logo](#)

[Trees](#)

[Curation & model](#)

[Species](#)

[Interactions](#)

[Structures](#)

Jump to...

Summary: Kunitz/Bovine pancreatic trypsin inhibitor domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

[Wikipedia: Kunitz domain](#) [Pfam](#) [InterPro](#)

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

Kunitz/Bovine pancreatic trypsin inhibitor domain [Provide feedback](#)

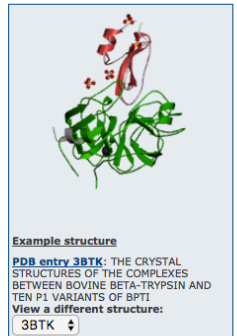
Indicative of a protease inhibitor, usually a serine protease inhibitor. Structure is a disulfide rich alpha+beta fold. BPTI (bovine pancreatic trypsin inhibitor) is an extensively studied model structure. Certain family members are similar to the tick anticoagulant peptide (TAP, [P17726](#)). This is a highly selective inhibitor of factor Xa in the blood coagulation pathways [1]. TAP molecules are highly dipolar [2] and are arranged to form a twisted two- stranded antiparallel beta-sheet followed by an alpha helix [1].

Literature references

1. Antuch W, Guntert P, Billeter M, Hawthorne T, Grossenbacher H, Wuthrich K; , FEBS Lett 1994;352:251-257.: NMR solution structure of the recombinant tick anticoagulant protein (rTAP), a factor Xa inhibitor from the tick *Ornithodoros moubata*. [PUBMED:7925983](#) [EPMC:7925983](#)
2. St Charles R, Padmanabhan K, Arni RV, Padmanabhan KP, Tulinsky A; , Protein Sci 2000;9:265-272.: Structure of tick anticoagulant peptide at 1.6 A resolution complexed with bovine pancreatic trypsin inhibitor. [PUBMED:10716178](#) [EPMC:10716178](#)

External database links

HOMSTRAD: [kunitz](#)
PRINTS: [PR00759](#)
PROSITE: [PDOC00252](#)
SCOP: [5pti](#)



Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Domain Organization

The Kunitz domain is present in many proteins with different architectures

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

There are 1580 sequences with the following architecture: Kunitz_BPTI

[W6UEG6_ECHGR](#) [Echinococcus granulosus (Hydatid tapeworm)] Kunitz-type proteinase inhibitor 5 II {ECO:0000313|EMBL:EUB59840.1} (239 residues)



[Show](#) all sequences with this architecture.

There are 404 sequences with the following architecture: Kunitz_BPTI x 2

[A0A183BW84_GLOPA](#) [Globodera pallida (Potato cyst nematode)] Uncharacterized protein {ECO:0000313|WPParaSite:GPLIN_000487300} (146 residues)



[Show](#) all sequences with this architecture.

There are 209 sequences with the following architecture: Kunitz_BPTI x 3

[L9KVVW5_TUPCH](#) [Tupaia chinensis (Chinese tree shrew)] Tissue factor pathway inhibitor {ECO:0000313|EMBL:ELW66614.1} (293 residues)



[Show](#) all sequences with this architecture.

There are 101 sequences with the following architecture: MANEC, Kunitz_BPTI, Ldl_recept_a, Kunitz_BPTI

[K7G5T1_PELSI](#) [Pelodiscus sinensis (Chinese softshell turtle) (Trionyx sinensis)] Uncharacterized protein {ECO:0000313|Ensembl:ENSPSIP00000015642} (518 residues)



[Show](#) all sequences with this architecture.

There are 94 sequences with the following architecture: APP_N, APP_Cu_bd, Kunitz_BPTI, APP_E2, APP_amyloid

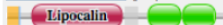
[W5LK57_ASTMX](#) [Astyanax mexicanus (Blind cave fish) (Astyanax fasciatus mexicanus)] Uncharacterized protein {ECO:0000313|Ensembl:ENSAMXP00000020219} (782 residues)



[Show](#) all sequences with this architecture.

There are 92 sequences with the following architecture: Lipocalin, Kunitz_BPTI x 2

[W5M2V3_LEPOC](#) [Lepisosteus oculatus (Spotted gar)] Uncharacterized protein {ECO:0000313|Ensembl:ENSLOCP00000002711} (349 residues)



[Show](#) all sequences with this architecture.

There are 89 sequences with the following architecture: APP_N, APP_Cu_bd, Kunitz_BPTI, APP_E2, Beta-APP, APP_amyloid

[Q6DJ86_XENTR](#) [Xenopus tropicalis (Western clawed frog) (Silurana tropicalis)] Uncharacterized protein {ECO:0000313|Ensembl:ENSXETP000000029842} (750 residues)



[Show](#) all sequences with this architecture.

There are 79 sequences with the following architecture: WAP, Kunitz_BPTI

[U1MCF8_ASCSU](#) [Ascaris suum (Pig roundworm) (Ascaris lumbricoides)] Wap four-disulfide core domain protein 8 {ECO:0000313|EMBL:ERG86819.1} (281 residues)



[Show](#) all sequences with this architecture.

There are 65 sequences with the following architecture: Reeler, Spond_N, TSP_1 x 3, Kunitz_BPTI, TSP_1 x 2

[Q7K3Y9_DROME](#) [Drosophila melanogaster (Fruit fly)] Uncharacterized protein {ECO:0000313|EMBL:AAM68661.1} (873 residues)



[Show](#) all sequences with this architecture.

PFAM Alignments

PFAM stores different alignments with increasing number of sequences.

Alignments

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database ([reference proteomes](#)) using the family HMM. We also generate alignments using four [representative proteomes](#) (RP) sets, the UniProtKB sequence database, the NCBI sequence database, and our metagenomics sequence database. [More...](#)

View options

We make a range of alignments for each Pfam-A family. You can see a description of each [above](#). You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (100)	Full (11819)	Representative proteomes				UniProt (24966)	NCBI (25020)	Meta (22)
			RP15 (3597)	RP35 (5870)	RP55 (8328)	RP75 (9477)			
Jalview	✓	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	—	✗	✗	✗	✗	✗	✗	✗
PP/heatmap	✗ ¹	—	✗	✗	✗	✗	✗	✗	✗

¹Cannot generate PP/Heatmap alignments for seeds; no PP data available

Key: ✓ available, ✗ not generated, — not available.

Format an alignment

	Seed (100)	Full (11819)	Representative proteomes				UniProt (24966)	NCBI (25020)	Meta (22)
			RP15 (3597)	RP35 (5870)	RP55 (8328)	RP75 (9477)			
Alignment:	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Format:	Selex ▾								
Order:	<input checked="" type="radio"/> Tree <input type="radio"/> Alphabetical								
Sequence:	<input checked="" type="radio"/> Inserts lower case <input type="radio"/> All upper case								
Gaps:	Gaps as "." or "-" (mixed) ▾								
Download/view:	<input checked="" type="radio"/> Download <input type="radio"/> View								
Generate									

Download options

We make all of our alignments available in Stockholm format. You can download them here as raw, plain text files or as [gzip](#)-compressed files.

	Seed (100)	Full (11819)	Representative proteomes				UniProt (24966)	NCBI (25020)	Meta (22)
			RP15 (3597)	RP35 (5870)	RP55 (8328)	RP75 (9477)			
Raw Stockholm	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gzipped	✓	✓	✓	✓	✓	✓	✓	✓	✓

You can also [download](#) a FASTA format file containing the **full-length sequences** for all sequences in the full alignment.

PFAM Curation

Information about the PFAM family alignment is reported in the Curation page

Curation and family details

This section shows the detailed information about the Pfam family. You can see the definitions of many of the terms in this section in the [glossary](#) and a fuller explanation of the scoring system that we use in the [scores](#) section of the help pages.

Curation ⓘ

Seed source:	Prosite
Previous IDs:	none
Type:	Domain
Author:	Fenech M
Number in seed:	100
Number in full:	11819
Average length of the domain:	53.30 aa
Average identity of full alignment:	35 %
Average coverage of the sequence by the domain:	20.00 %

HMM information ⓘ

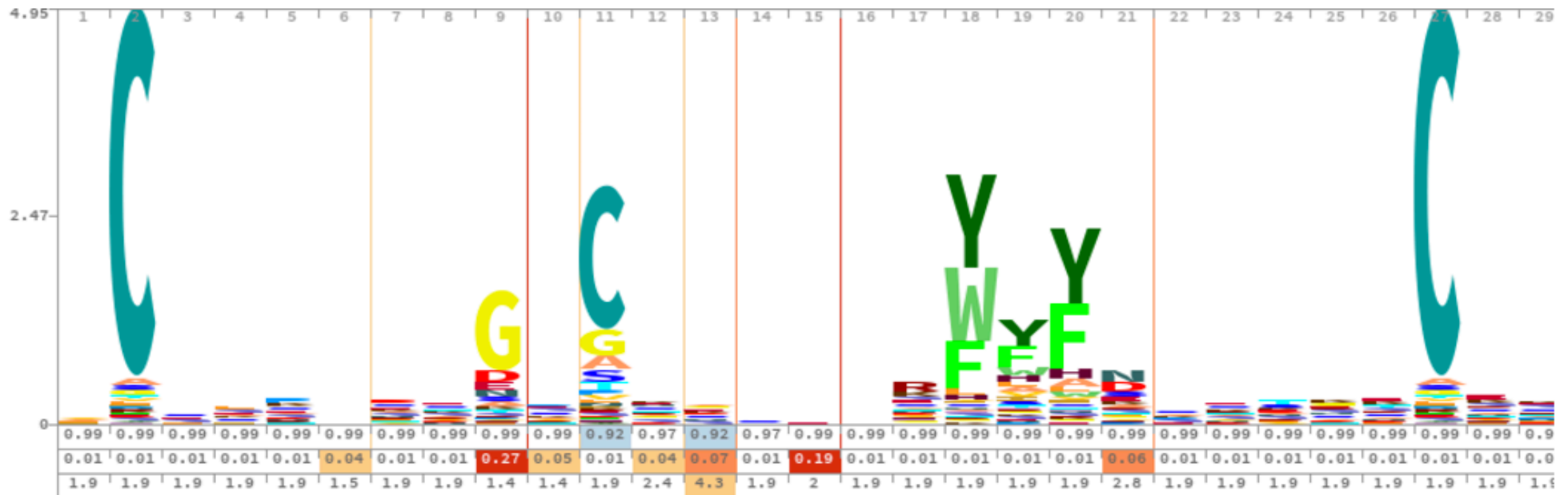
HMM build commands:	<i>build method:</i> hmmbuild -o /dev/null HMM SEED <i>search method:</i> hmmsearch -Z 26740544 -E 1000 --cpu 4 HMM pfamseq		
Model details:	Parameter	Sequence	Domain
	Gathering cut-off	21.0	21.0
	Trusted cut-off	21.0	21.0
	Noise cut-off	20.9	20.9
Model length:	53		
Family (HMM) version:	22		
Download:	download the raw HMM for this family		

HMM Logo

Important protein sites can be visualized using HMM Logo

HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them [here](#). [More...](#)



Specific Aims

The specific aims are:

1. Build your own model for the Kunitz domain, starting from available structural information.
2. Use the model for annotating Kunitz domains in SwissProt.

Write a detailed draft of the project identifying

- the main steps;
- the sources of the data to be analyzed;
- the procedures/programs you would adopt;
- The results to be produced for validating your model

Structure Selection

Retrieve available structures of the Kunitz domain

This is the crucial step: you need to collect an ample set of structures that are endowed with

Source: *PDB*

Method: *different alternative options are possible:*

- a. consider a prototype structure and search in the PDB other similar structures (e.g, by using the PDBe-fold web site)*
- b. retrieve from UniProt the protein endowed with an annotated BPTI/ Kunitz type domain and with a 3D structure covering it.*
- c. Try to directly scan the PDB for structurally-resolved Kunitz domains(e.g., you can use the CATH code 4.10.410.10)*
- d.*

Possible Issues

When **selecting the domains** for building the seed alignment, keep in mind that:

- PDB files can contain **more than one sequence**;
- **A chain can contain different domains** of the same type or of different types;
- Structures of the same sequence can be found in **different PDB files**;
- Sometimes PDB files report the structure of **mutated proteins**;
- **Resolution** can be an issue during structural alignment.

Protein Alignment

Perform the structural alignment of the selected domains

Method: Any multiprotein structural alignment method (e.g. PDBe-fold)

On the basis of the structural alignment results you can correct/refine your initial choice of the seed proteins.

Save or convert the alignment among sequences in Stockholm format

Method: JalView or write an ad-hoc program

Generate HMM Model

Train a profile HMM

Method: *HMMER hmmbuild routine*

Verify that the trained HMM is able to recognize the proteins in your dataset (consistency test)

Method: *HMMER hmmsearch routine*

If the performance on the train set is low there is probably some problem in the set of proteins your choose and/or in the alignment you fed to HMM during the training procedure

Method Testing

Retrieve a suitable dataset for validating the HMM prediction

Only manually curated proteins should be considered, avoiding fragments
The dataset should be divided into proteins containing or not containing the BPTI/Kunitz domain (the positive test set should exclude the training data).

Source: UniProt/Swiss-Prot

Method: *The “advanced search” interface in UniProt web site*

Different “Golden standard” for defining the positive class are possible:

- a) the presence of an annotated BPTI/Kunitz domain in the Uniprot entry*
- b) the presence of an annotated PF00014 PFAM domain*
- c) ..*

Search the validation dataset against the trained model

Method: *HMMER hmmsearch routine*

Compute the scoring indexes for evaluating your profile HMM on the validation sets

Method: *Write a program that compares the prediction with the “real” annotations, computes a confusion matrix and the scoring indexes.*

Analyze the Results

Analyze the results and try to understand whether it is possible to improve them

Prediction could be in some cases optimized by changing the E-value threshold or by refining the training alignment.

Discuss the False Positive and the False Negative predictions

Find your domain in all the SwissProt sequences, comment with respect to the available annotations and comment about the distribution of the Kunitz domain

Project Report

Project description in the “Bioinformatics” style paper

http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/submission_online.html

Structured Abstract (see recent issues of journal for examples)

Original papers

Abstracts are structured with a standard layout such that the text is divided into sub-sections under the following five headings: **Motivation**, **Results**, [Availability and Implementation], **Contact** [and Supplementary Information]. In cases where authors feel the headings inappropriate, some flexibility is allowed. The abstracts should be succinct and contain only material relevant to the headings. **A maximum of 150 words is recommended.**

- *Motivation*: This section should specifically state the scientific question within the context of the field of study.
- *Results*: This section should summarize the scientific advance or novel results of the study, and its impact on computational biology.

Project Report Abstract

Project description in the “Bioinformatics” style paper

http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/submission_online.html

Structured Abstract (see recent issues of journal for examples)

Original papers

Abstracts are structured with a standard layout such that the text is divided into sub-sections under the following five headings: **Motivation**, **Results**, [Availability and Implementation], **Contact** [and Supplementary Information]. In cases where authors feel the headings inappropriate, some flexibility is allowed. The abstracts should be succinct and contain only material relevant to the headings. **A maximum of 150 words is recommended.**

- *Motivation*: This section should specifically state the scientific question within the context of the field of study.
- *Results*: This section should summarize the scientific advance or novel results of the study, and its impact on computational biology.

Main Report

Introduction

The section must describe the problem treated in the paper, the available knowledge on it. Only information relevant within the scope of the paper should be reported. Appropriate references must cited.

Materials and Methods

The section must contain the description of the adopted dataset and of the methods that have been used and/or implemented, including the validation procedures and the adopted scoring indexes. Adopted choice must be justified. In principle, **it must contain all the information necessary to integrally reproduce the work.**

Results (and discussion)

The section must present the obtained results, the possible refinements, and the analysis of the strength and the weakness of the method. Discussion (can be a separate section) must report the considerations that can be derived from results, also in relation to the adopted procedures and/or datasets.

Conclusions

The section present concisely the achievements of the presented work.

Reference and Data

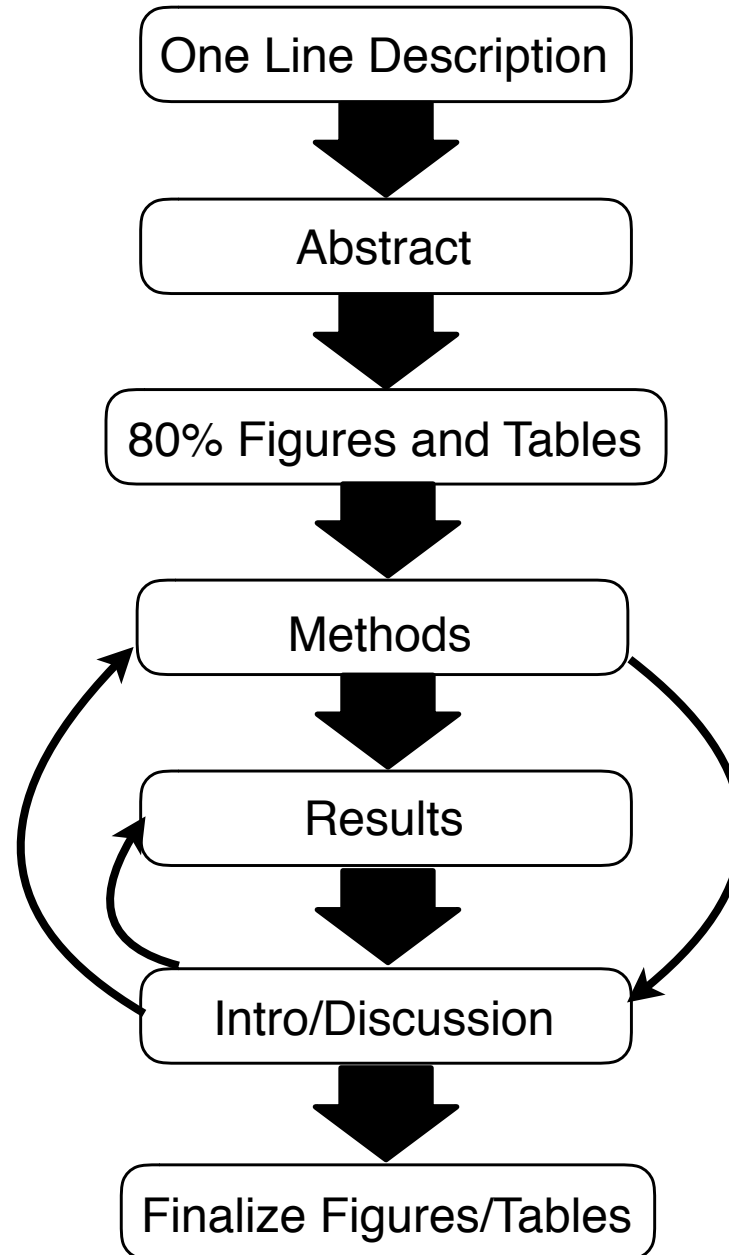
References

See the template for the appropriate format

Supplementary Materials

Supplementary file useful for the presentation of the work can be provided

Flow Chart



Project Submission

- The presentation and the approval of the project paper is a necessary but not sufficient condition to pass the exam.
- The first version of the paper is due by May 6, 2019 at 23:59 CET.
- The revised version of the paper is due by May 27, 2019 at 23:59 CET.
- Submit the paper with subject: project-lb1b - Name Surname to:
emidio.capriotti@unibo.it

Exercise

Build a *blast*-based method to predict the presence of BPTI/Kunitz domain in proteins available in SwissProt using the human proteins as a reference.

- Select all Proteins in SwissProt with BPTI/Kunitz domain.
- Separate human from non human proteins. Use the **non human proteins as a positive** in the testing set.
- Generate a **random set of negative** of the same size of the positive set.
- Remove both positives and negatives from SwissProt and perform the **prediction based on the results of the *blast* search**.