

Project Description

Laboratory of Bioinformatics I
Module 2

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

[Structure Summary](#) [3D View](#) [Annotations](#) [Experiment](#) [Sequence](#) [Genome](#) [Versions](#)

Biological Assembly 1 ?



3D View: [Structure](#) | [Electron Density](#) | [Ligand Interaction](#)

Global Symmetry: Asymmetric - C1 ?

Global Stoichiometry: Hetero 2-mer - A1B1 ?

[Find Similar Assemblies](#)

3TGI

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

DOI: [10.2210/pdb3TGI/pdb](https://doi.org/10.2210/pdb3TGI/pdb)

Classification: **COMPLEX (SERINE PROTEASE/INHIBITOR)**

Organism(s): [Rattus norvegicus](#), [Bos taurus](#)

Mutation(s): No ?

Deposited: 1998-07-15 **Released:** 1998-12-23

Deposition Author(s): [Pasternak, A.](#), [Ringe, D.](#), [Hedstrom, L.](#)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 1.80 Å






R-Value Free: 0.210

R-Value Work: 0.177

R-Value Observed: 0.177

wwPDB Validation ?

[3D Report](#) [Full Report](#)

Metric	Percentile Ranks	Value
Rfree		0.193
Clashscore		4
Ramachandran outliers		0.4%
Sidechain outliers		4.3%
RSRZ outliers		0.7%

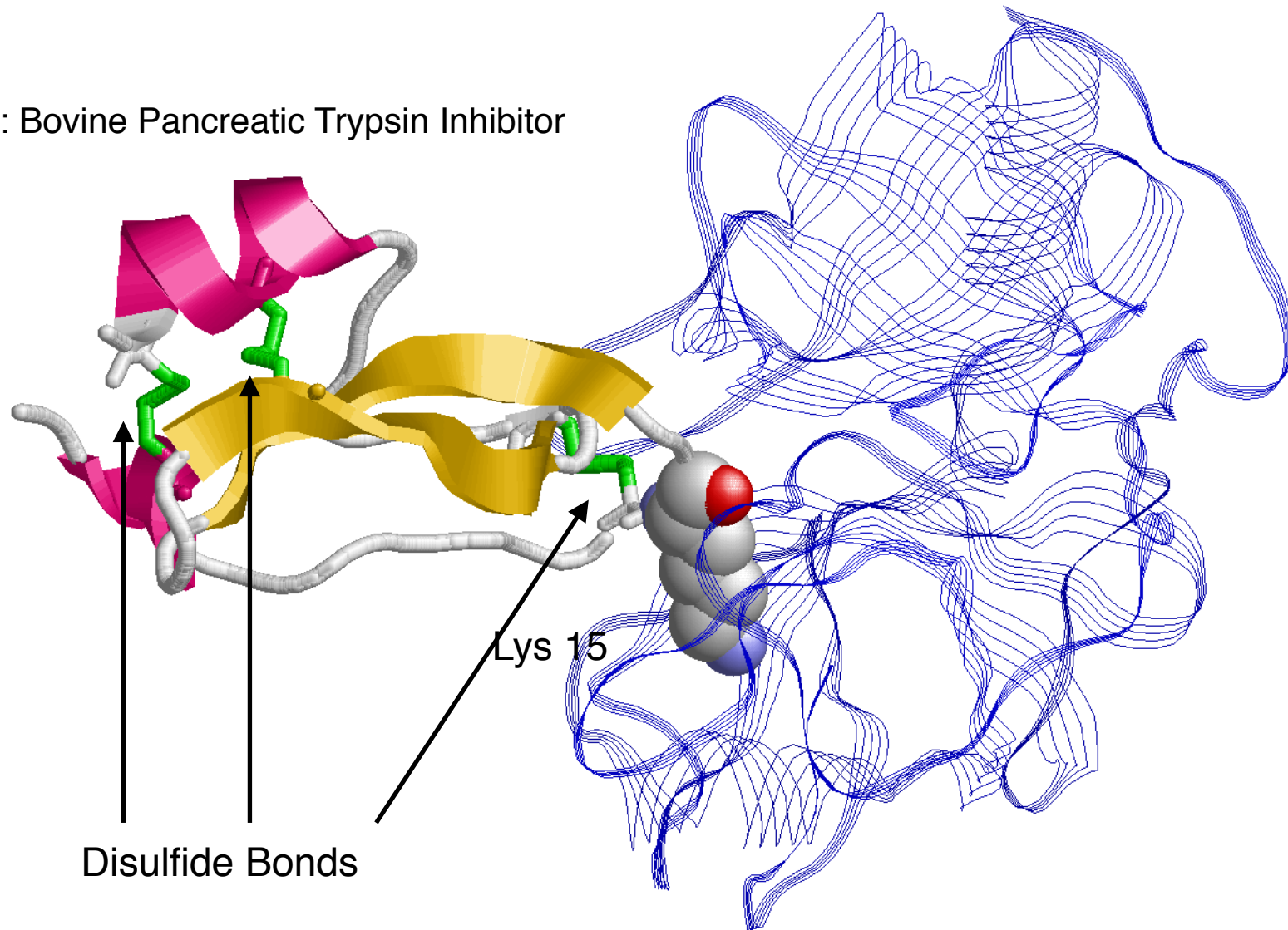
Worse Better
■ Percentile relative to all X-ray structures
□ Percentile relative to X-ray structures of similar resolution

This is version 1.2 of the entry. See complete [history](#).

Structure Analysis

Chain E: Rat Anionic Trypsin

Chain I: Bovine Pancreatic Trypsin Inhibitor



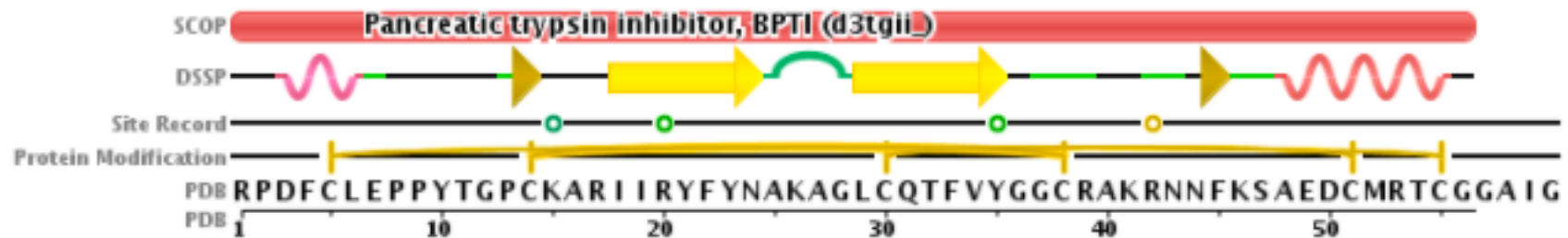
The Protein Fold

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 | Pancreatic trypsin inhibitor

Gene | N/A

Organism | *Bos taurus* (Bovine)

Status | 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	Sequence analysis	Add BLAST		21
Propeptide ⁱ (PRO_0000016852)	22 – 35		Add BLAST		14
Chain ⁱ (PRO_0000016853)	36 – 93	Pancreatic trypsin inhibitor	Add BLAST		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	PROSITE-ProRule annotation 1 Publication			
Disulfide bond ⁱ	65 ↔ 86	PROSITE-ProRule annotation 1 Publication			

PFAM

The Kunitz BPTI family is described in PFAM database



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



Family: *Kunitz_BPTI* (PF00014)

1247 architectures 19493 sequences 0 interactions 468 species 319 structures

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Structures

Jump to...

enter ID/acc



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 is the Wikipedia entry entitled "[Kunitz domain](#)". [More...](#)

Kunitz domain [Edit Wikipedia article](#)

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](#). They are relatively small with a length of about 50 to 60 [amino acids](#) and a molecular weight of 6 [kDa](#). 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).

Standalone Kunitz domains are used as a framework for the development of new [pharmaceutical drugs](#).^[2]

Contents^[hide]

- [1 Structure](#)
- [2 Drug development](#)
- [3 Examples](#)
- [4 References](#)

Structure

The structure is a [disulfide](#) rich alpha+beta fold. 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.^[3] TAP molecules are highly [dipolar](#),^[4] and are arranged to form a twisted two-stranded antiparallel [beta sheet](#) followed by an [alpha helix](#).^[3]

The majority of the sequences having this domain belong to the [MEROPS](#) inhibitor family I2, clan IB; the Kunitz/bovine pancreatic trypsin inhibitor family, they inhibit proteases of the S1 family^[5] and are restricted to the [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. The type example for this family is BPTI^[6] (or basic protease inhibitor), but the family includes numerous other members, ^{[7][8][9][10]} such as snake venom basic protease; mammalian [inter-alpha-trypsin inhibitors](#); [trypstatin](#), 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-termini](#) of the [alpha-1](#) and [alpha-3 chains](#) of type VI and type VII [collagens](#); tissue factor pathway inhibitor precursor; and [Kunitz STI protease inhibitor](#) contained in [legume](#) seeds.

Kunitz/Bovine pancreatic trypsin inhibitor domain



3D structure of the [C-terminal](#) Kunitz domain from human [collagen alpha-3\(VI\) chain](#).^[1]

Identifiers

Symbol	Kunitz_BPTI
Pfam	PF00014 ↗
InterPro	IPR002223 ↗
PROSITE	PDOC00252 ↗
SCOPe	5pti ↗ / SUPFAM ↗
CDD	cd00109 ↗

Available protein structures: [\[show\]](#)

Domain Organization

Domain organisation

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

There are 2557 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 801 sequences with the following architecture: Kunitz_BPTI x 2

[A0A5E4CT99_MARMO](#) [Marmota monax (Woodchuck)] Uncharacterized protein {ECO:0000313|EMBL:VTJ84379.1} (175 residues)



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

There are 595 sequences with the following architecture: APP_N, APP_Cu_bd, Kunitz_BPTI, APP_E2, Beta-APP, APP_amyloid

[U3IMZ9_ANAPP](#) [Anas platyrhynchos platyrhynchos (Northern mallard)] ABPP {ECO:0000256|ARBA:ARBA00018220} (717 residues)



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

There are 557 sequences with the following architecture: Kunitz_BPTI x 3

[A0A401RXU9_CHIPU](#) [Chiloscyllium punctatum (Brownbanded bambooshark) (Hemiscyllium punctatum)] Tissue factor pathway inhibitor {ECO:0000256|PIRNR:PIRNR001620} (223 residues)



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

There are 391 sequences with the following architecture: APP_N, APP_Cu_bd, Kunitz_BPTI, APP_E2, APP_amyloid

[W5PZA9_SHEEP](#) [Ovis aries (Sheep)] Uncharacterized protein {ECO:0000313|Ensembl:ENSOARP00000015795} (728 residues)



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

There are 252 sequences with the following architecture: Lipocalin, Kunitz_BPTI x 2

[W5M2V3_LEPOC](#) [Lepisosteus oculatus (Spotted gar)] Alpha-1-microglobulin {ECO:0000256|ARBA:ARBA00020539} (349 residues)



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

There are 199 sequences with the following architecture: WAP, I-set, Kunitz_BPTI x 2, NTR

[H3A6B1_LATCH](#) [Latimeria chalumnae (Coelacanth)] Uncharacterized protein {ECO:0000313|Ensembl:ENSLACP00000005182} (575 residues)



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

There are 166 sequences with the following architecture: MANEC, Kunitz_BPTI, Ldl_recept_a, Kunitz_BPTI

[H3CGS7_TETNG](#) [Tetraodon nigroviridis (Spotted green pufferfish) (Chelonodon nigroviridis)] Uncharacterized protein {ECO:0000313|Ensembl:ENSTNIP00000007455} (476 residues)

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](#)¹) using the family HMM. We also generate alignments using four [representative proteomes](#)² (RP) sets and the UniProtKB 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 (99)	Full (19493)	Representative proteomes				UniProt (39468)
			RP15 (6370)	RP35 (10077)	RP55 (17585)	RP75 (22004)	
Jalview	✓	✓	✓	✓	✓	✓	✓
HTML	✓	—	×	×	×	×	×
PP/heatmap	×	—	×	×	×	×	×

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

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

Format an alignment

	Seed (99)	Full (19493)	Representative proteomes				UniProt (39468)
Alignment:	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Format:	Select ▼						
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

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
Sequence Ontology:	SO:0000417
Author:	Fenech M 
Number in seed:	99
Number in full:	19493
Average length of the domain:	53.10 aa
Average identity of full alignment:	36 %
Average coverage of the sequence by the domain:	15.34 %

HMM information ⓘ

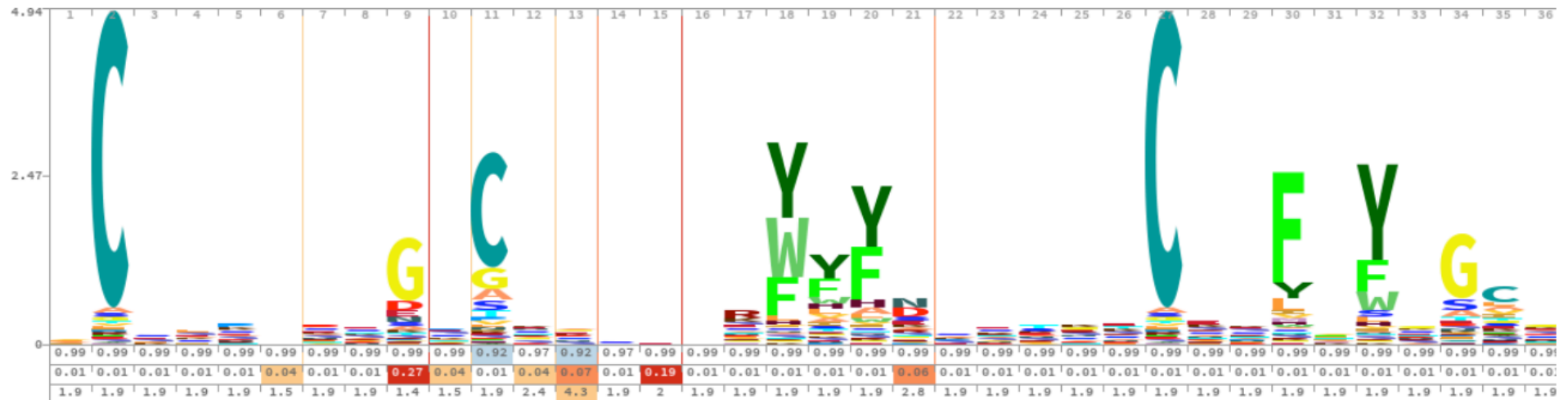
HMM build commands:	<i>build method:</i> hmmbuild -o /dev/null HMM SEED <i>search method:</i> hmmsearch -Z 57096847 -E 1000 --cpu 4 HMM pfamseq		
Model details:	Parameter	Sequence	Domain
	Gathering cut-off	21.0	21.0
	Trusted cut-off	21.1	21.0
	Noise cut-off	20.9	20.9
Model length:	53		
Family (HMM) version:	25		
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 a large 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 chain**;
- **A chain can contain different domains** of the same type or of different types;
- Structures of the same protein can be found in **different PDB files**;
- the PDB collects 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 multiple 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.

If needed convert the alignment 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 “Gold 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.

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

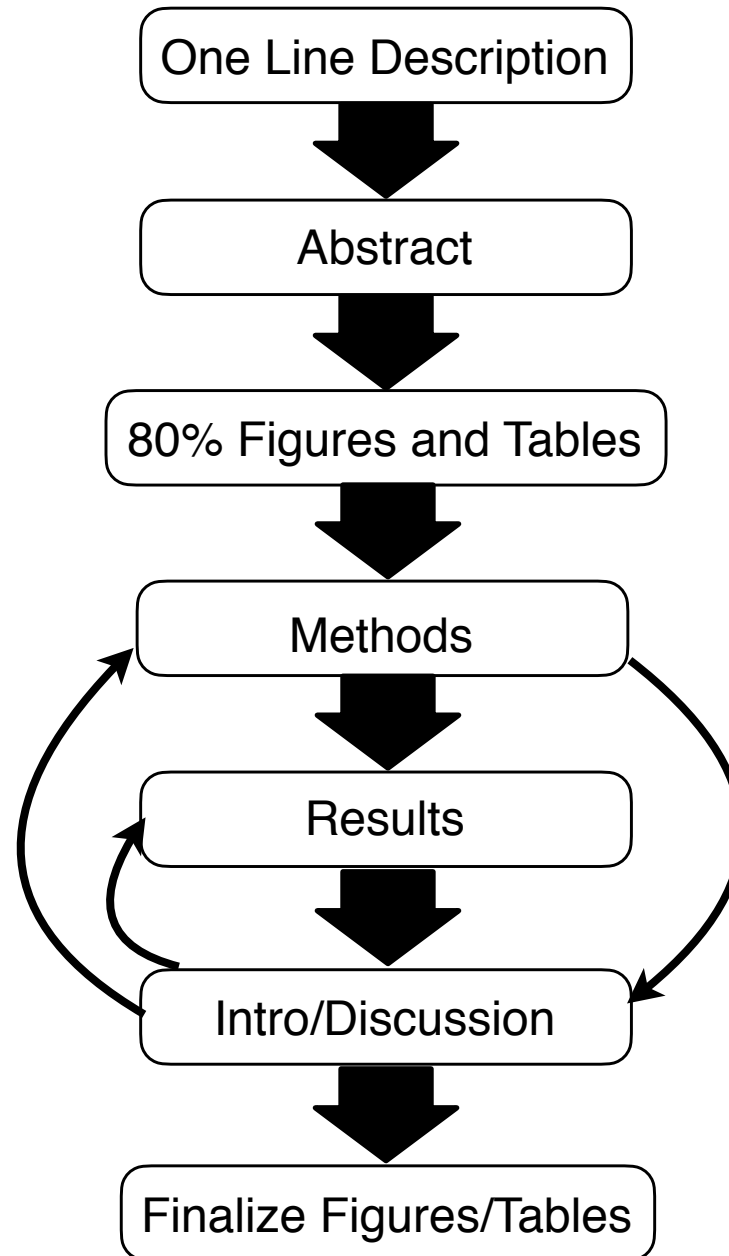
References

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 necessary but not sufficient condition to pass the exam.
- The first version of the paper is due by May 17, 2021 at 23:59 CET.
- The revised version of the paper is due by June 7, 2021 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**.