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 Analysis

Chain E: Rat Anionic Trypsin

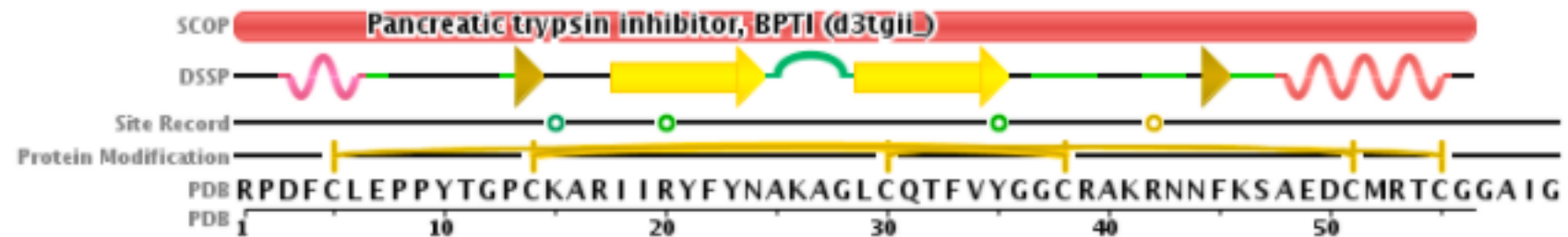Chain I: Bovine Pancreatic Trypsin Inhibitor

Lys 15

Disulfide Bonds
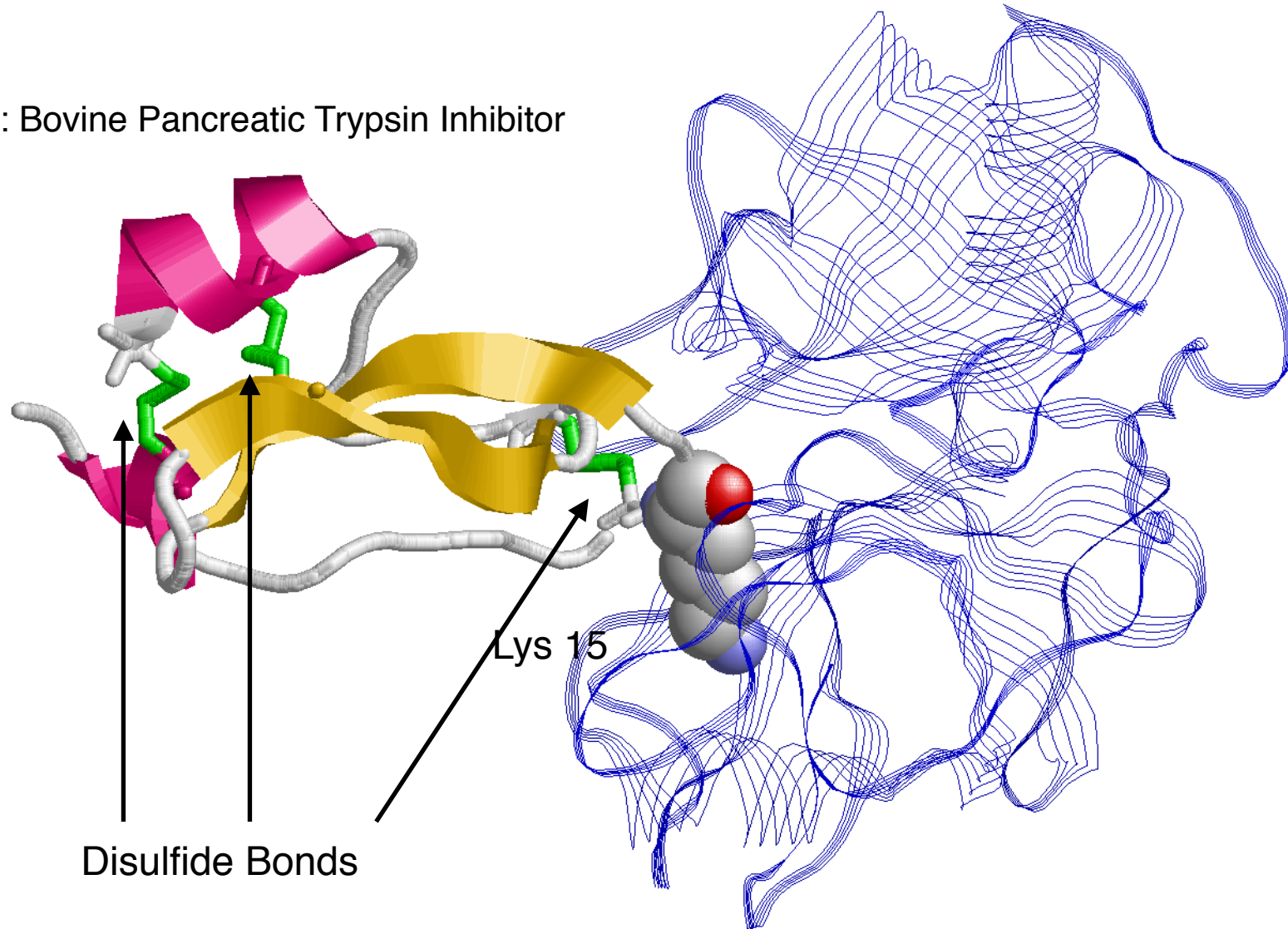
The Protein Fold

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

SCOP: Pancreatic trypsin inhibitor, BPTI (d3tgii_)

PDB: RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGAIG

PDB: 1   10   20   30   40   50

Annotation

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

PFAM

The Kunitz BPTI family is described in PFAM database

Domain Organization

The Kunitz domain in present in many proteins with different architectures

PFAM Alignments

PFAM stores different alignments with increasing number of sequences.

PFAM Curation

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

HMM Logo

Important protein sites can be visualized using HMM Logo
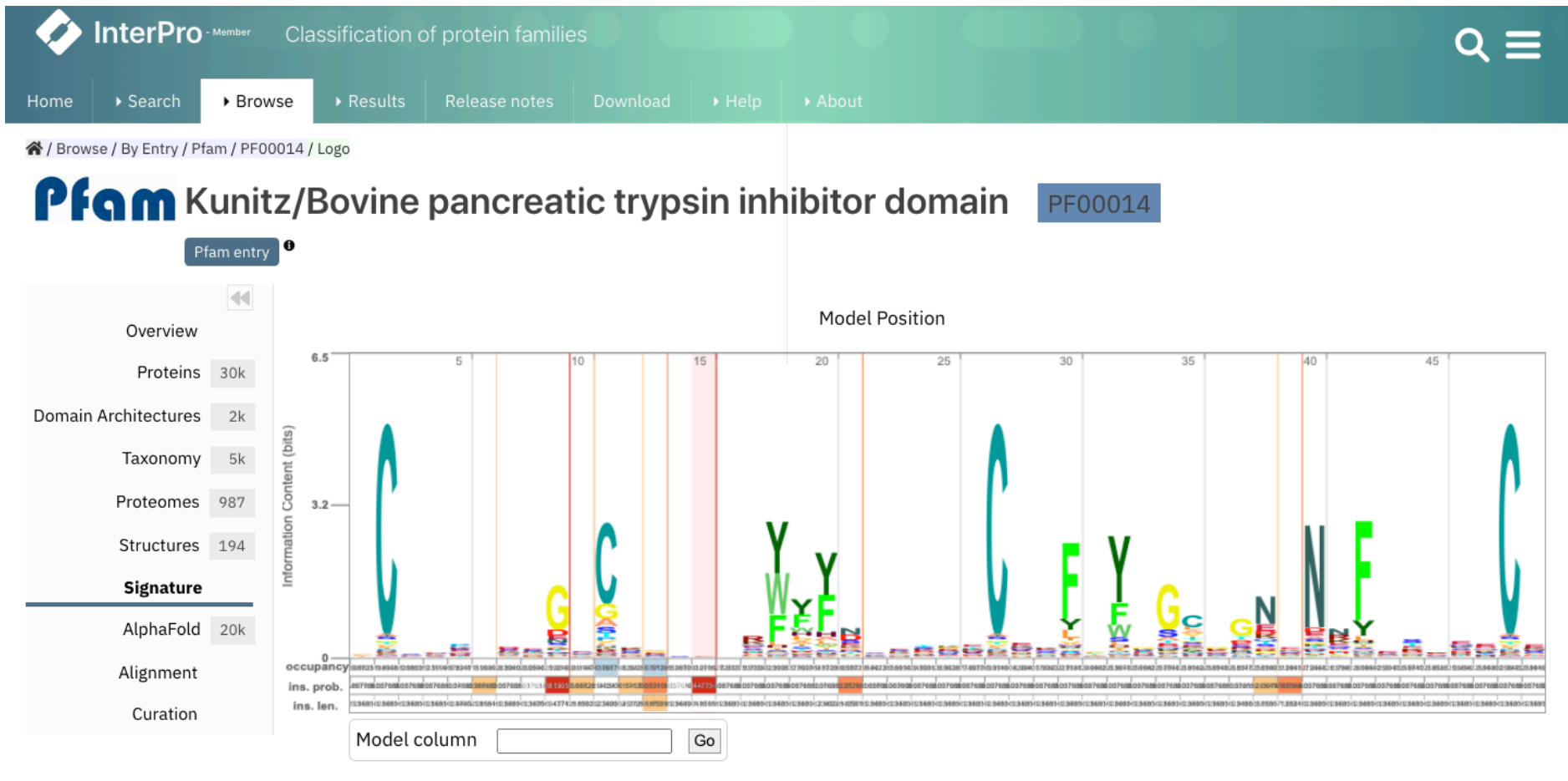
Specific Aims

The specific aims are:

1. Build your own model for the Kunitz domain, starting from available <u>structural</u> information.

2. Use the model for annotating Kunitz domains in SwissProt.

Write a <u>detailed</u> 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.
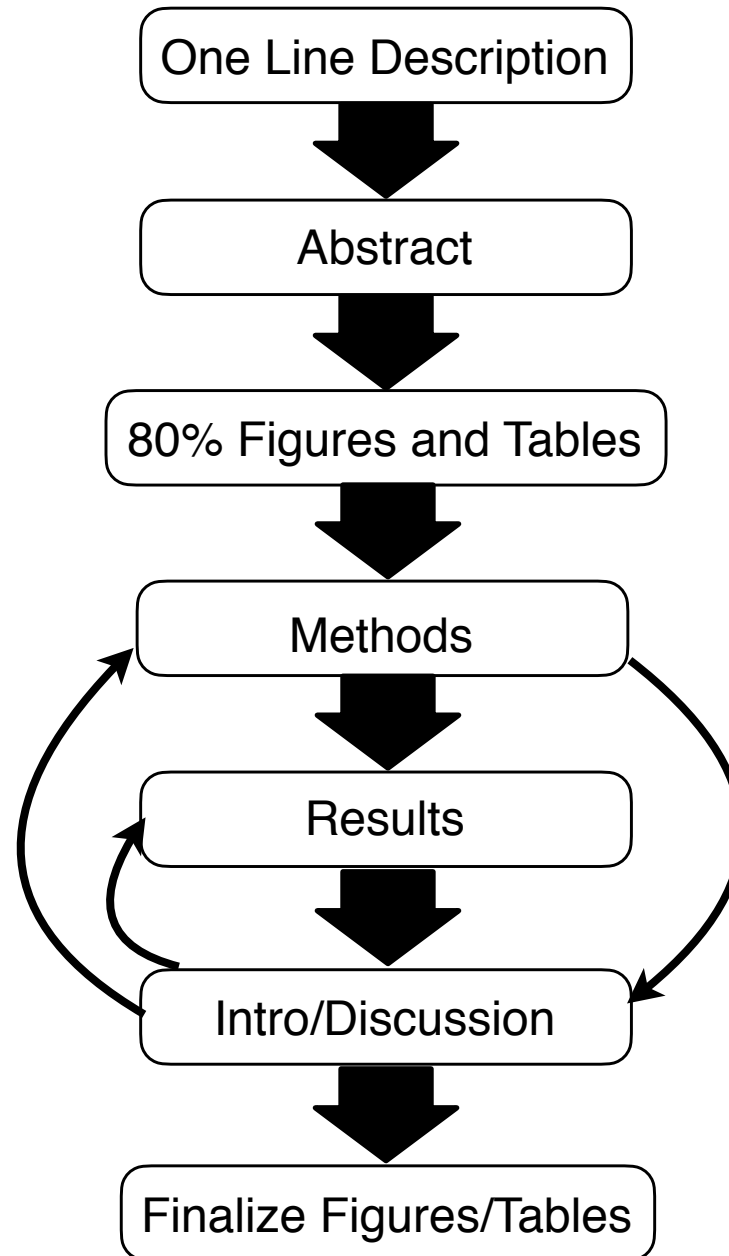
Reference and Data

References

See the template for the appropriate format

Supplementary Materials

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

Flow Chart

One Line Description

↓

Abstract

↓

80% Figures and Tables

↓

Methods

↓

Results

↓

Intro/Discussion

↓

Finalize Figures/Tables

Project Submission

- The presentation and the approval of the project paper is necessary but not sufficient condition to pass the exam.

- 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.