

A Model for Protein Secondary Structure Prediction Meta - Classifiers

Anita Wasilewska

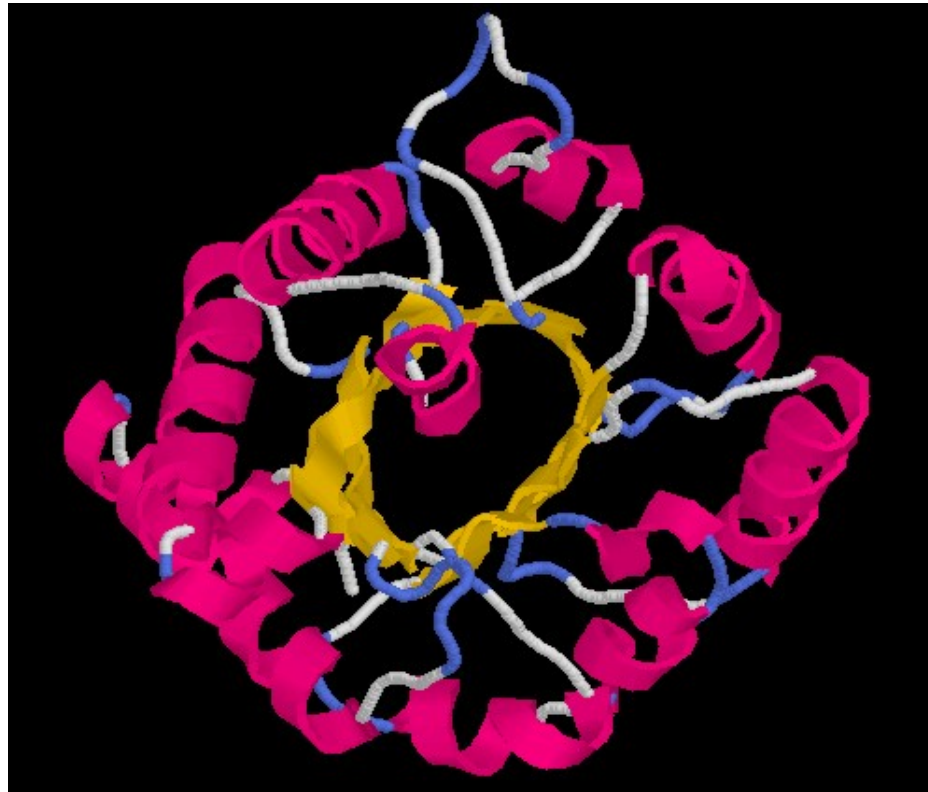
Computer Science Department

Stony Brook University, Stony Brook, NY, USA

Overview

- **Introduction to proteins**
- **Four levels of protein structure: symbolic model**
- **Proteomic databases**
- **PSSP datasets**
- **Protein Secondary Structure Prediction**
 - **The Window:** role and symbolic model
 - Use of evolutionary information
- **PSSP Metaclassifiers**
- **Future Research**

Introduction to proteins



Proteins

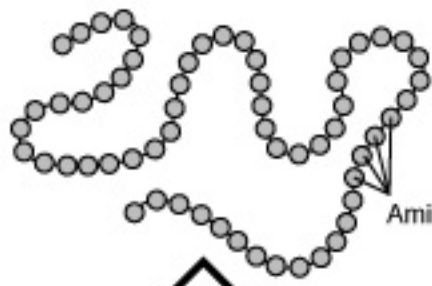
- **Protein**: from the Greek word **PROTEUO** which means "to be first (in rank or influence)"
- **Why are proteins important to us:**
 - Proteins make up about 15% of the mass of the average person
 - Enzyme** – acts as a biological catalyst
 - Storage and transport – **Haemoglobin**
 - Antibodies**
 - Hormones – Insulin

Proteins

- **Why are proteins important to us (c.d.):**
 - **Ligaments and arteries** (mainly former by elastin protein)
 - **Muscle** – Proteins in the muscle respond to nerve impulses by changing the packing of their molecules
 - **Hair, nails and skins:** protein α -keratin as main component
 - And more

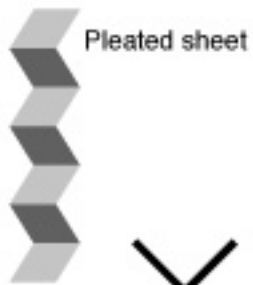
Proteins Research Benefits

- **Medicine** – design of drugs which inhibit specific enzyme targets for therapeutic purposes (engineering of insulin)
- **Agriculture** – Treat diseases of plants and to modify growth and development of crops
- **Industry** – Synthesis of enzymes to carry out industrial processes on a mass scale



Primary protein structure
is sequence of a chain of amino acids

Amino Acids

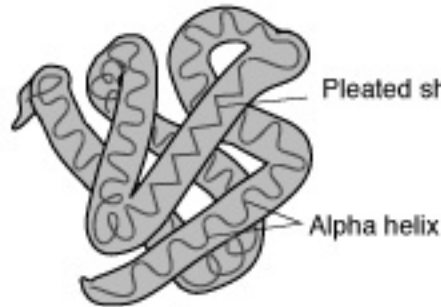


Pleated sheet



Alpha helix

Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds



Pleated sheet

Alpha helix

Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.



Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Four levels of protein structure

Four Levels of Protein Structure:

AMINOACIDS

- **AMINOACIDS:** There are 20 aminoacids:
 - Alanine (A), Cysteine (C), Aspartic Acid (D),
 - Glutamic Acid (E), Phenylalanine (F), Glicine (G),
 - Histidine (H), Isoleucine (I),Lycine (K), Leucine (L),
 - Methionine (M), AsparagiNe (N), Proline (P),
 - Glutamine (Q), ARginine (R), Serine (S),
 - Threonine (T), Valine (V),Tryptophan (W),
 - TYrosine (Y)
- **AMINOACIDS SYMBOLS:**
A,C,D,E,F,G,H,I,J,K,L,M,N,P,Q,R,S.T,V,W,Y

Primary Structure

Symbolic Definition

- $A = \{A,C,D,E,F,G,H,I,J,K,L,M,N,P,Q,R,S.T,V,W,Y\}$ – set of symbols denoting all aminoacids
- A^* - set of all finite sequences formed out of elements of A .
- Elements of A^* are denoted by $x, y, z \dots$.i.e. we write $x \in A^*, y \in A^*, z \in A^*, \dots$ etc
- Any $x \in A^*$ is called a **protein sequence** or **protein sub-unit primary structure**
- Any $x_1, x_2, \dots x_n \in (A^*)^*$, is called a **protein primary structure**.

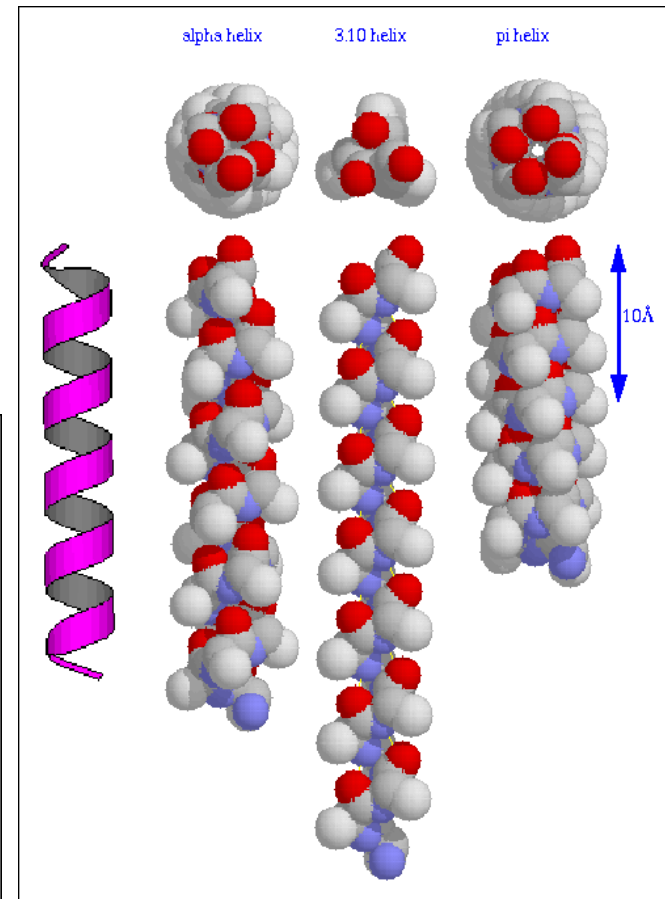
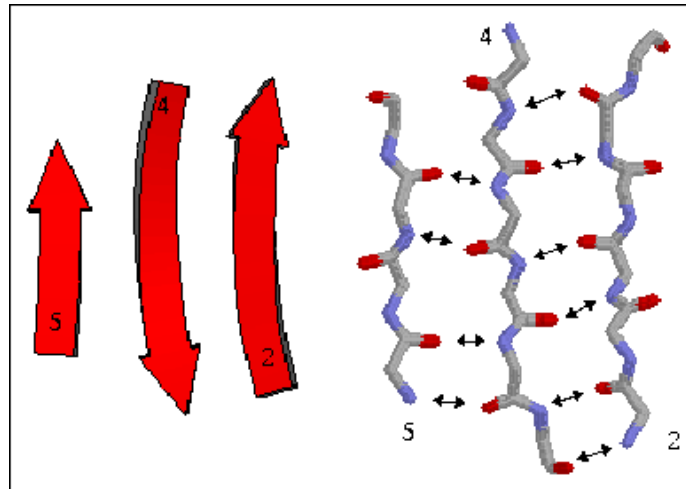
Protein Secondary Structure SS

- **Secondary structure** is the term protein chemist give to the arrangement of the peptide backbone in space. It is produced by hydrogen bondings between aminoacids
- **The assignment of the SS categories** to the experimentally determined three-dimensional (3D) structure of proteins is a non-trivial process and is typically **performed by widely used DSSP program**
- **PROTEIN SECONDARY STRUCTURE** consists of : **protein sequence and its hydrogen bonding patterns called SS categories**

Secondary Structure

8 different categories (DSSP):

- **H**: α - helix
- **G**: 3_{10} - helix
- **I**: π - helix (extremely rare)
- **E**: β - strand
- **B**: β - bridge
- **T**: β - turn
- **S**: bend
- **L**: the rest



Protein Secondary Structure

- Databases for **protein sequences** are expanding rapidly due to the genome sequencing projects and the gap between the number of determined protein structures (**PSS – protein secondary structures**) and the number of known protein sequences in public
- Protein data banks (**PDB**) are growing bigger and bigger
- **PSSP (Protein Secondary Structure Prediction)** research is trying to breach this gap.

Three secondary structure states

- Prediction methods are normally trained and assessed for only **3 states** (**residues**):
- **H (helix)**, **E (strands)** and **L (coil)**
- There are many published **8-to-3 states reduction** methods
- **Standard reduction methods are defined by programs DSSP** (Dictionary of SS of Proteins), **STRIDE**, and **DEFINE**
- **Improvement of predictive accuracy of different SSP** (Secondary Structure Prediction) programs depends on the choice of the reduction method

Three SS states: Reduction methods

- **Method 1**, used by **DSSP** program:
- **H**(helix) = { G (3_{10} – helix), H (α - helix)}
- **E** (strands) = {E (β -strand), B (β -bridge)} , **L** (coil) – all the rest
 - Shortly: **E,B** => **E**; **G,H** => **H**; **Rest** => **C**
 - We are using this method that is the **most difficult to predict** and is used in **CASP** (International contests for PSSP programs)
- **Method 2**, used by **STRIDE** program:
- **H**(helix) as in Method 1
- **E** (strands) = {E (β -strand), B (β -bridge)},
- **L** (coil) – all the rest

Three SS states: Reduction methods

- **Method 3**, used by **DEFINE** program:
- **H**(helix) as in Method 1
- **E** (strands) = {E (β -strand)}, **L** (coil) – all the rest

- **Some other methods:**

Method B: $E \Rightarrow E$; $H \Rightarrow H$; $\text{Rest} \Rightarrow L$;
 EE and $HHHH \Rightarrow L$

Method C: $GGGHHHH \Rightarrow HHHHHHH$;
 $B, GGG \Rightarrow L$; $H \Rightarrow H$; $E \Rightarrow E$

Example of typical PSS Data

- **Protein Data** is gathered in a form of **protein sub-units (sequences)** and assigned to them **sequences of SS categories H, E, L** as observed **empirically** in their 3-dimensional structures. **The SS categories are assigned by DSSP program**

- **Example:**

Sequence

KELVLALYDYQEKSPREVTTHKKGDILTLLNSTNKDWWKYEYNDRQGFVP

Observed SS

HHHHHLLLLEEEHHHLLLEEEEEELLHHHHHHHHLLLEEEEEELLHHH

Protein Secondary Structure (PSS): Symbolic Definition

- Given $A = \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, Y\}$
– set of symbols denoting aminoacids and a protein sequence (sub-unit) $x \in A^*$
- Let $S = \{H, E, L\}$ be the set of symbols of 3 states (residues): H (helix), E (strands) and L (coil) and S^* be the set of all finite sequences of elements of S.
- We denote elements of S^* by e, o , with indices if necessary i.e. we write $e \in S^*$, $e_1, e_2 \in S^*$, etc...

Protein Secondary Structure (PSS): Symbolic Definition

- Any **PARTIAL, ONE –TO –ONE FUNCTION**

$$f : A^* \rightarrow S^* \quad \text{i.e.} \quad f \subseteq A^* \times S^*$$

is called a protein secondary structure (PSS) sub-unit identification function

- An element $(x, e) \in f$ is called **protein secondary structure** (of the protein sub- unit x)
- The element **e** (of $(x, e) \in f$) is called secondary structure sequence, or for short a secondary structure of x.

Protein Secondary Structure (PSS)

Symbolic Definition

- Following the standard way we write in **PSS research** the pairs sequence-structure, we represent the pair (x,e) i.e. the protein secondary structure in a vertical form: $\left(\begin{array}{c} \boxed{x} \\ \boxed{e} \end{array} \right)$ and hence the PSS identification function **f** is viewed as the set of all secondary structures it identifies i.e.

- $$\mathbf{f} = \left\{ \left(\begin{array}{c} \boxed{x} \\ \boxed{e} \end{array} \right) : x \in A^* \cap x \in \text{Dom} \mathbf{f} \cap e = \mathbf{f}(x) \right\}$$

PSS Identification Function

Examples

- Any Data Set (DS) used in PSS Prediction defines its own identification function f_{DS} empirically and by DSSP program and we identify DS with f_{DS} and write

$$DS = \left\{ \left(\begin{array}{c} x \\ e \end{array} \right) : f_{DS}(x) = e \right\}$$

- For example:** if **DS** is such that a protein sequence **ARNVSTVVLA** has the observed **SS** sequence **HHHEEECCCH**

we put :

$$f_{DS}(\text{ARNVSTVVLA}) = \text{HHHEEECCCHH} \text{ and write}$$

$$\left(\begin{array}{c} \text{ARNVSTVVLA} \\ \text{HHHEEECCCH} \end{array} \right) \in DS$$

Tertiary Structure

- The tertiary structure of a protein is the arrangement in space of all its atoms
- The overall **3D shape** of a protein molecule is a compromise, where the structure has the best balance of attractive and repulsive forces between different regions of the molecule
- For a given protein we can **experimentally determine its tertiary structure by X-rays or NMR**
- Given the **tertiary structure** of a protein we can know its **secondary structure** with the **DSSP** program

Protein Sequence Tertiary Structure Symbolic Definition

- Let $s \in (A^*)^*$ denote a protein sequence,
- $\mathbf{P} = (s, e) \in \mathbf{F}$ be secondary structure of s ,
the element

$$\varphi_x = (\mathbf{P}, t_s)$$

is a tertiary structure of s

where t_s is the sequence's tertiary folding function

Quaternary Structure

- Many globular proteins are made up of several polypeptide chains called **sub-units**, stuck to each other by a variety of attractive forces but rarely by covalent bonds. Protein chemists describe this as **quaternary structure**.



Protein Quaternary Structure

- Quaternary structure is a pair

$$(Q, f_Q)$$

- where Q is a multiset of tertiary structures,

(for example $Q = [\alpha, \alpha, \beta, \beta]$ in a haemoglobin) and

- f_Q is the quaternary folding function

Protein Symbolic Definition

PROTEIN **P** = { protein **P** primary structure, protein **P** sub-units, their secondary structures, their tertiary structure, their quaternary structure}

PROTEIN **P** = { $x_1 \dots x_n$, $(x_1, e_1)(x_2, e_2) \dots (x_n, e_n)$,
 $\alpha_{x_1} = ((x_1, e_1), t_{x_1}) \dots \alpha_{x_n} = ((x_n, e_n), t_{x_n})$,
 $([\alpha_{x_1}, \dots, \alpha_{x_n}], f_{\alpha_{x_1}, \dots, \alpha_{x_n}}) \}$

where x_i is protein **P** i th sub-unit, t_{x_i} is x_i 's tertiary folding function and $f_{\alpha_{x_1}, \dots, \alpha_{x_n}}$ is protein **P** quaternary folding function

Protein: Symbolic Definition

- In **PSSP** research we deal with protein sub-units (sequences) x_i , not with the whole sequence of sub-units x_1, x_2, \dots, x_n
- We write P_{x_k} when we refer only to the sub-unit x_k of the protein **P**
- We write $P_{(x_i, e_i)}$ when we refer to the sub-unit x_i of the protein **P** and its secondary structure
- We write $P_{\alpha_{x_i}}$ when we refer to the sub-unit x_i of the protein **P** and its secondary structure and its tertiary structure

Protein sub-units

Given a protein $\mathbf{P} = \{x_1 \dots x_n, (x_1, e_1)(x_2, e_2) \dots (x_n, e_n),$

$\alpha_{x_1} = ((x_1, e_1), t_{x_1}) \dots \alpha_{x_n} = ((x_n, e_n), t_{x_n}), ([\alpha_{x_1}, \dots, \alpha_{x_n}],$
 $f_{\alpha_{x_1}, \dots, \alpha_{x_n}}) \}$

- $\mathbf{P}_{x_i} = \{ x_i \}$
- $\mathbf{P}(x_i, e_i) = \{ x_i, (x_i, e_i) \}$
- $\mathbf{P}_{\alpha_{x_i}} = \{ x_i, (x_i, e_i), \alpha_{x_i} = ((x_i, e_i), t_{x_i}) \}$

Example: Haemoglobin

- Haemoglobin = $\{ x, y, (x, e_x), (y, e_y), \alpha = ((x, e_x), t_x), \beta = ((y, e_y), t_y), ([\alpha, \alpha, \beta, \beta], f_{\alpha, \beta}) \}$

Where $x, y \in A^*$, are called haemoglobin sub-units

Proteomic Databases

- The most important proteomic databases are:

Swiss-Prot + TrEMBL

PIR-PSD

PIR-NREF

PDB

Swiss-Prot + TrEMBL

Web site: <http://us.expasy.org/sprot/>

- **Swiss-Prot** is a protein sequence database with high level of annotations, a minimal level of redundancy and high level of integration with other databases. **124464 entries (and growing)**
- **TrEMBL** is a computer-annotated supplement of Swiss-Prot that contains all sequence entries not yet integrated in Swiss-Prot. **828210 entries (and growing)**

PIR-PSD

- Web site: <http://pir.georgetown.edu/>
- **PIR-PSD**: Protein Information Resource
- Protein Sequence Database
- Founded in 1960 by Margaret Dayhoff
- Comprehensive and **annotated** protein sequence database in the public domain. **283308 entries** (and growing)

PIR-NREF

- Web site: <http://pir.georgetown.edu/>
- **PIR-NREF: PIR Non-Redundant Reference Protein Database**
- It contains all sequences in **PIR-PSD**, **SwissProt**, **TrEMBL**, RefSeq, GenPept, and PDB.
- **1,186,271 entries**
- **The most used** for finding protein profiles with **PSI-BLAST** program.
- **A mandatory in Protein Secondary Structure Prediction (PSSP) research**

PDB: Protein Data Bank

- Web site: <http://www.rcsb.org/pdb/>
- PDB contains 3-D biological macromolecular structure data
- 22-April-2003 => 20747 Structures
- How do we use PDB?
- All **PSSP** datasets start with some **PDB** sequences with known secondary structures. Then, with **DSSP** program we get the **secondary structure** and its **reduction** to **three categories** and use it as **learning data** for our algorithms

Protein Secondary Structure Prediction

- Techniques for the **prediction of protein secondary structure** provide information that is useful both in
- ab initio **structure prediction** and
- as an **additional constraint** for **fold-recognition** algorithms.
- Knowledge of **secondary structure** alone can help the design of site-directed or deletion mutants that will **not destroy** the **native protein structure**.
- For all these applications it is essential that the **secondary structure prediction** be **accurate**, or at least that, the reliability for each residue can be assessed.

Protein Secondary Structure Prediction

- If a protein sequence **shows clear similarity** to a protein of **known** three dimensional structure, then the **most accurate method of predicting the secondary structure** is to **align** the sequences by **standard** dynamic programming algorithms, as the **homology modelling** is much more **accurate** than **secondary structure prediction** for high levels of sequence identity.
- **Secondary structure prediction** methods are of **most use** when **sequence similarity** to a **protein of known structure** is undetectable.
- It is important that **there is no detectable sequence similarity** between sequences **used to train** and **test secondary structure prediction** methods.

PSSP Datasets

- Historic **RS126** dataset. Contains **126 sub-units** with known secondary structure selected by **Rost** and **Sander**. Today is not used anymore
- **CB513** dataset. Contains **513 sub-units** with known secondary structure selected by **Cuff** and **Barton** in 1999. **Very much used in PSSP research**
- **HS17771** dataset. Created by **Hobohm** and **Scharf**. In March-2002 it contained **1771 sub-units**
- This family of datasets are non redundant **PDB** (Protein Data Bank) subsets. Sub-units in the dataset never have an identity bigger than 25%.
- Lots of authors has their own and “*secret*” datasets

PSSP Algorithms

- There are three generations in PSSP algorithms
 - **First Generation**: based on **statistical** information of single aminoacids
 - **Second Generation**: based on **windows** (segments) of aminoacids. Typically a window contains 11-21 aminoacids
 - **Third Generation**: based on the use of **windows** on **evolutionary information**

PSSP: First Generation

- First generation PSSP systems are based on **statistical information on a single aminoacid**
- The most relevant algorithms:
 - Chow-Fasman, 1974**
 - GOR, 1978**
- Both algorithms claimed **74-78%** of predictive accuracy, but tested with better constructed datasets were proved to have **the predictive accuracy ~50%** (Nishikawa, 1983)

PSSP: Second Generation

- Based on the information contained in **a window of aminoacids** (11-21 aa.)
- The most systems use algorithms based on:
 - Statistical information
 - Physico-chemical properties
 - Sequence patterns
 - **Multi-layered neural networks**
 - Graph-theory
 - Multivariate statistics
 - **Expert rules**
 - Nearest-neighbour algorithms
 - **Bayesian networks**

PSSP: Second Generation

- **Main problems:**

Prediction accuracy <70%

Prediction accuracy for β -strand 28-48%

Predicted chains are usually too short

what leads do the difficult use

of predictions

PSSP: Third Generation

- **PHD**: First algorithm in this generation (**1994**)
- Evolutionary information **improves the prediction accuracy to 72%**
- **Use of evolutionary information:**
 1. **Scan** a database with known sequences with alignment methods for finding **similar sequences**
 2. **Filter** the previous list with a threshold to identify the **most significant** sequences
 3. **Build aminoacid exchange profiles** based on the probable **homologs** (most significant sequences)
 4. The **profiles** are used in the prediction

PSSP: Third Generation

- Many of the **second** generation algorithms have been **updated to third** generation
- **The most important algorithms of today**
 - Predator**: Nearest-neighbour
 - PSI-Pred**: Neural networks
 - SSPro**: Neural networks
 - SAM-T02**: Homologs (Hidden Markov Models)
 - PHD**: Neural networks
- Due to the improvement of protein information in databases i.e. better evolutionary information, today's predictive accuracy is **~80%**
- **It is believed that maximum reachable accuracy is 88%**

PSSP Data Preparation

- **Public Protein Data Sets** used in **PSSP** research contain **protein secondary structure sequences**. In order to use **classification algorithms** we must **transform secondary structure sequences** into **classification data tables**.
- **Records** in the **classification** data tables are called, in **PSSP** literature (learning) **instances**.
- **The mechanism used in this transformation process is called window**.
- **A window algorithm** has a secondary structure as **input** and returns a **classification table: a set of instances** for the classification algorithm.

Window

- Consider a secondary structure (x, e) .
- $(x, e) = (x_1x_2 \dots x_n, e_1e_2 \dots e_n)$
- **Window** of the length k chooses a **subsequence of length k** of $x_1x_2 \dots x_n$, and **an element e_i** from $e_1e_2 \dots e_n$, corresponding to a special position in the window, usually the middle
- **Window moves** along the sequences $x = x_1x_2 \dots x_n$ and $e = e_1e_2 \dots e_n$ simultaneously, starting at the **beginning** moving **to the right one letter** at the time at each step of the process.

Window: Sequence to Structure

- Such window is called **sequence to structure window**. We will call it for short **a window**.
- The process **terminates** when the **window** or **its middle** position reaches **the end** of the sequence **x**.
- The pair: (subsequence, element of **e**) is often written in a form **subsequence** \rightarrow **H, E** or **C** is called **an instance**, or **a rule**.

Example: Window

- Consider a secondary structure (x, e) and the window of length 5 with the special position in the middle (bold letters)
- **Fist position** of the window is:

- $x =$

A	R	N	S	T	V	V	S	T	A	A
---	---	----------	----------	----------	---	---	---	---	---	---	------
- $e =$

H	H	H	H	C	C	C	E	E	E
---	---	----------	----------	----------	----------	---	---	---	---

- **Window returns instance:**
- **A R N S T** \rightarrow **H**

Example: Window

- Second position of the window is:

- $x =$ A R N **S** T V V S T A A
- $e =$ H H H **C** C C E E E

- Windows returns instance:

- **R N S T V** \rightarrow **H**

- Next instances are:

- **N S T V V** \rightarrow **C**

- **S T V V S** \rightarrow **C**

- **T V V S T** \rightarrow **C**

Symbolic Notation

- Let **f** be a protein secondary structure (PSS) identification function:
- $f : A^* \rightarrow S^*$ i.e. $f \subseteq A^* \times S^*$
- Let $x = x_1 x_2 \dots x_n$, $e = e_1 e_2 \dots e_n$ and
- $f(x) = e$, we define
- $f(x_1 x_2 \dots x_n) | \{x_i\} = e_i$, i.e.
- **$f(x) | \{x_i\} = e_i$**

Example: Semantics of Instances

Let

• $x = A R N S T V V S T A A \dots$

• $e = H H H C C C E E E$

And assume that the windows returns an **instance (rule)**:

$A R N S T \rightarrow H$

• Semantics of the instance is:

$f(x)|\{N\}=H,$

where **f** is the **identification function** and **N** is preceded by **$A R$** and followed by **$S T$** and the window has the length 5

Classification Data Base (Table)

- **We build the classification table** with **attributes** being the positions $p_1, p_2, p_3, p_4, p_5 \dots p_n$ in the window, where n is length of the window. The corresponding **values of attributes** are elements of the **subsequent** on the given position.
- **Class** attribute is **S** with values in the set $\{H, E, C\}$ **assigned** by the window operation (**instance, rule**).
- The **classification table** for our example (fist few records) is the following.

Classification Table (Example)

- $x = \text{ARNSTVVSTAA} \dots$
- $e = \text{HHHCCC EEE}$

p1	p2	p3	p4	p5	S
A	R	N	S	T	H
R	N	S	T	V	H
N	S	T	V	V	C
S	T	V	V	S	C

Semantics of record $r = r(p1, p2, p3, p4, p5, S)$ is :

$$f(x)|_{\{V_{p3}\}} = V_s$$

where V_a denotes a value of the attribute a .

Missing Values

- **Missing values:** if we want to “cover” assignment of elements of **S** corresponding to **all of the elements** of the sequence x , we have to position the window with its middle position at the first element of x .
- In this case our **classification table** is (for our x and e)
 - $x = \text{ARNSTVVSTAA}$
 - $e = \text{HHHHCCCEEE}$

p1	p2	p3	p4	p5	S
		A	R	N	H
	A	R	N	S	H
A	R	N	S	T	H
R	N	S	T	V	H

Size of classification datasets (tables)

- The **window mechanism** produces very large datasets
- For example **window of size 13** applied to the **CB513** dataset of **513** protein subunits produces about
70,000 records (instances, rules)

Window

- **Window has the following parameters**
- **PARAMETER 1** : $i \in \mathbb{N}^+$ (Natural numbers >0) , the **starting point of the window** as it moves along the sequence $x = x_1 x_2 \dots x_n$. The value $i=1$ means that window starts at x_1 , $i=5$ means that window starts at x_5 , etc.
- **PARAMETER 2**: $k \in \mathbb{N}^+$ denotes the size (length) of the window.
- **For example:**
- the **PHD** system of **Rost and Sander** (1994) uses two window sizes: **13** and **17**
- The **BRNN** (**Bidirectional Recurrent Neural Networks**) of **Pollastri, Rost, Baldi and Przybylski** (2002) uses variable sizes of windows: **7,9,5** with **additional** windows (located on the “wheels”) of sizes **3 , 4 or 2**

Window

- **PARAMETER 3:** $p \in \{1, 2, \dots, k\}$
- where p is a **special position** of the window that returns the **classification attribute values from $S = \{H, E, C\}$** and
- k is the **size (length)** of the window
- **PSSP PROBLEM:**
find optimal **size k** , optimal **special position p** for the best prediction accuracy

Window: Symbolic Definition

- WINDOW ARGUMENTS: window parameters and secondary structure (x,e)
- WINDOW VALUE: (subsequence of x , element of e)
- OPERATION (sequence – to –structure window)
 W is a partial function

$$W: N^+ \times N^+ \times \{1, \dots, k\} \times (A^* \times S^*) \rightarrow A^* \times S$$

$$W(i, k, p, (x,e)) = (x_i x_{i+1} \dots x_{i+k-1}, f(x) \setminus \{x_{i+p}\})$$

where $(x,e) = (x_1 x_2 \dots x_n, e_1 e_2 \dots e_n)$

Sequence Alignment

- We perform **sequence alignment** to know if two sequences are **homologs**
- **Homologs: sequences with the same 3D structure and function**
- Main aspects:
 1. **Alignment classes:** Gapped vs. ungapped, global vs. partial
 2. **Punctuation systems:** Substitution matrices
 3. **Alignment Algorithms:**
 1. **Dynamic programming:** Needleman-Wunsch, Smith-Waterman
 2. **Heuristics: BLAST, FASTA**

Example (ungrapped alignment)

(a)

```
HBA_HUMAN  GSAQVKGHGKKVADALV
             G+ +VK+ HGKKV A+
HBB_HUMAN  GNPKVKAHGKKVLGAF
```

Homologs

(b)

```
HBA_HUMAN  GSAQVKGHGKKVADAL
             ++ ++++H+ KV   +
LGB2_LUPLU NNPELQAHAGKVFKLV
```

Homologs

(c)

```
HBA_HUMAN  GSAQVKGHGKKVADAL
             GS+ + G +   +D L
F11G11.2   GSGYLVGDSLTFVDLL
```

No homologs

Metaclassifier: Many Servers



Metaclassifier: Some servers

NAME	LOCATION	PREDICTION METHOD	Q3 CB513	RESULTS
<i>Predator</i>	Institut Pasteur - Paris	Nearest neighbour	80.0	e-mail
<i>PSIpred</i>	Univ College London	Neural network	79.9	e-mail
<i>Sspro</i>	Univ California Irvine	Neural network	79.1	e-mail
<i>SAM-T02</i>	Univ California, Santa Cruz	Homologs	78.1	e-mail
<i>PHD Exp</i>	Columbia Univ, New York	Neural network	77.6	e-mail
<i>Prof</i>	Univ Wales	Neural network	77.1	e-mail
Jpred	Univ Dundee, Scotland	Consensus	73.4	e-mail
SOPM	Institute of Biology, Lyon	Homologs	66.8	web
GOR	Univ Southampton. UK	Information Theory	55.4	web

Past Results

- The **choice of servers** is essential to the final results obtained by **trained** and **tested meta classifiers**.
- The experiments conducted in *Bayesian Network Multi-classifiers for Protein Secondary Structure Prediction Artificial Intelligence in Medicine, 2004; 31, pp. 117 – 136* (Victor Robles, Pedro Larranaga, Jose M. Pena, Ernestina Menasalvas, Maria S. Perez, Vanessa Herves, Anita Wasilewska)
- involved 4, 5, and 6 "hand selected" servers.
- It gave a **2-3% improvement in accuracy** over the **best single** method and **15-20%** over **the worst**.

Future Research

- Observe that we deal with a large amount of data.
- The **9 datasets** available for **training** and **testing** provide a really (about **1,000,000 records**) large, already well prepared, standardized, and **publicly available** set of data to experiment with.
- The meta classifiers data **do not contain missing values**, and **other than the statistical** methods (Bayes) can be used, **unlike** in the case of **PSSP** classifiers.

Future Research

- It is hence natural to **explore also the non-statistical, descriptive methods.**
- Additionally, these experiments could form a basis for a well founded research into **comparison** of **statistical** and **non-statistical** approaches.
- We refer interested readers to *Bayesian Network Multi-classifiers for Protein Secondary Structure Prediction Artificial Intelligence in Medicine, 2004; 31, pp. 117 – 136* for detailed methods of evaluating obtained results.