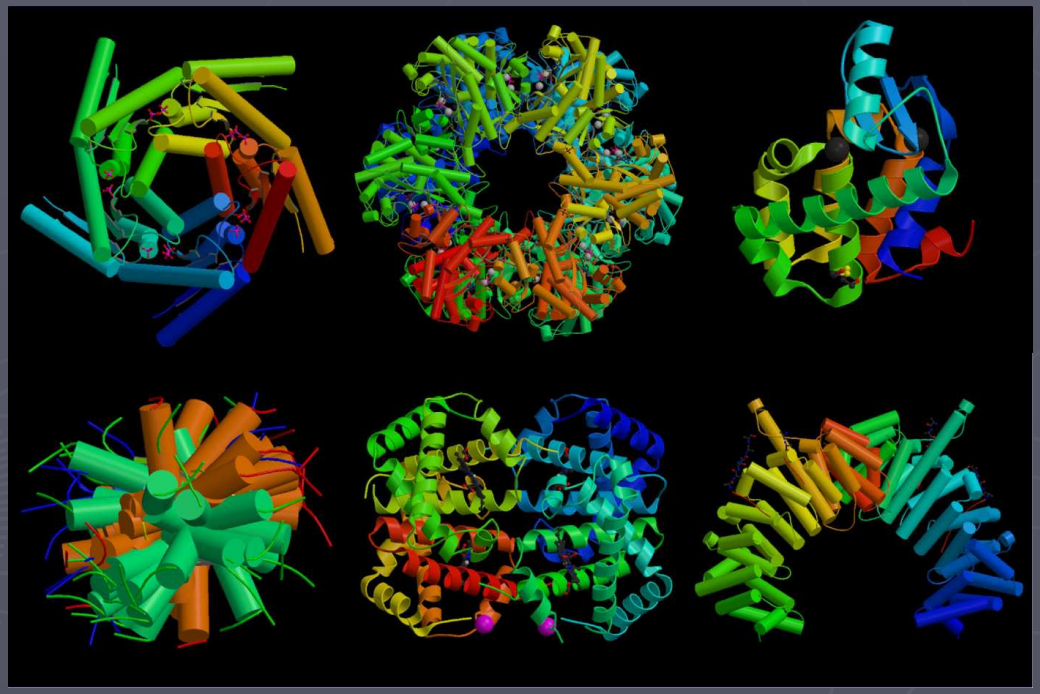
Biological Data Mining
(Predicting Post-synaptic Activity in Proteins)

Rashmi Raj

(rashmi@cs.stanford.edu)

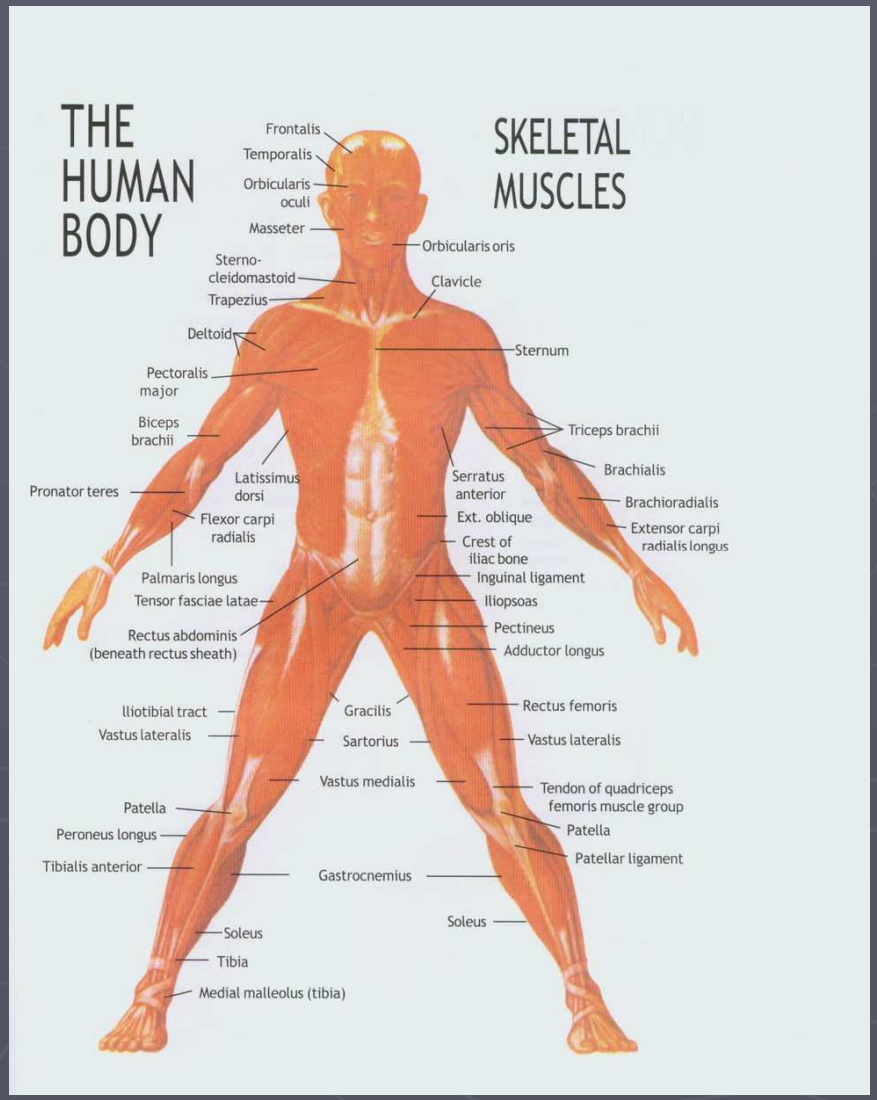
Protein

Enzymatic Proteins	Transport Proteins	Regulatory Proteins

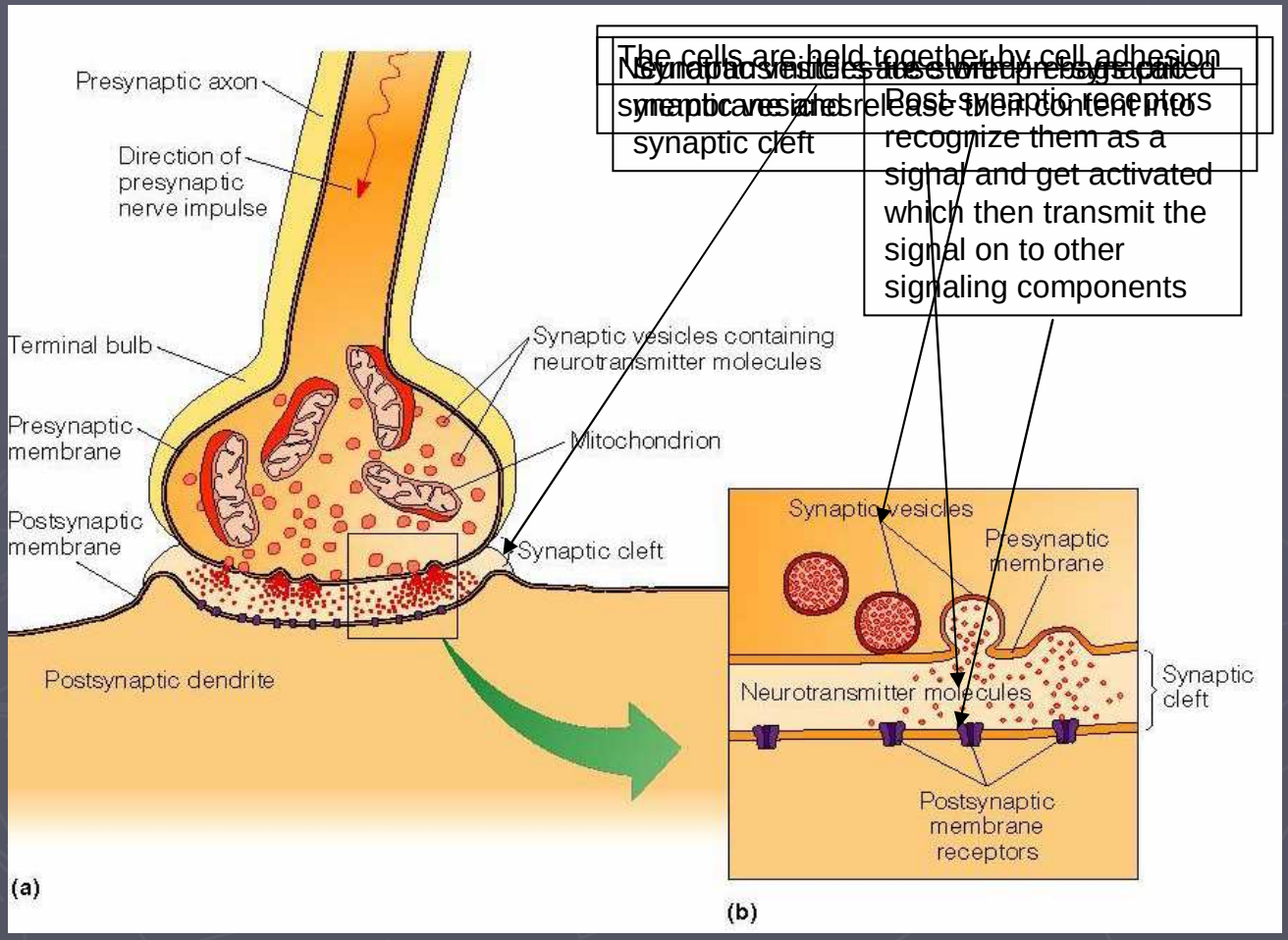


Storage Proteins	Hormonal Proteins	Receptor Proteins

Synaptic Activity



Pre-Synaptic And Post-Synaptic Activity



Source - http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-13/1319.jpg

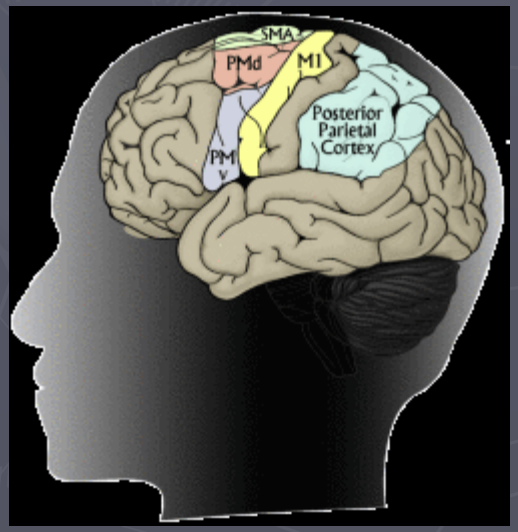
Problem

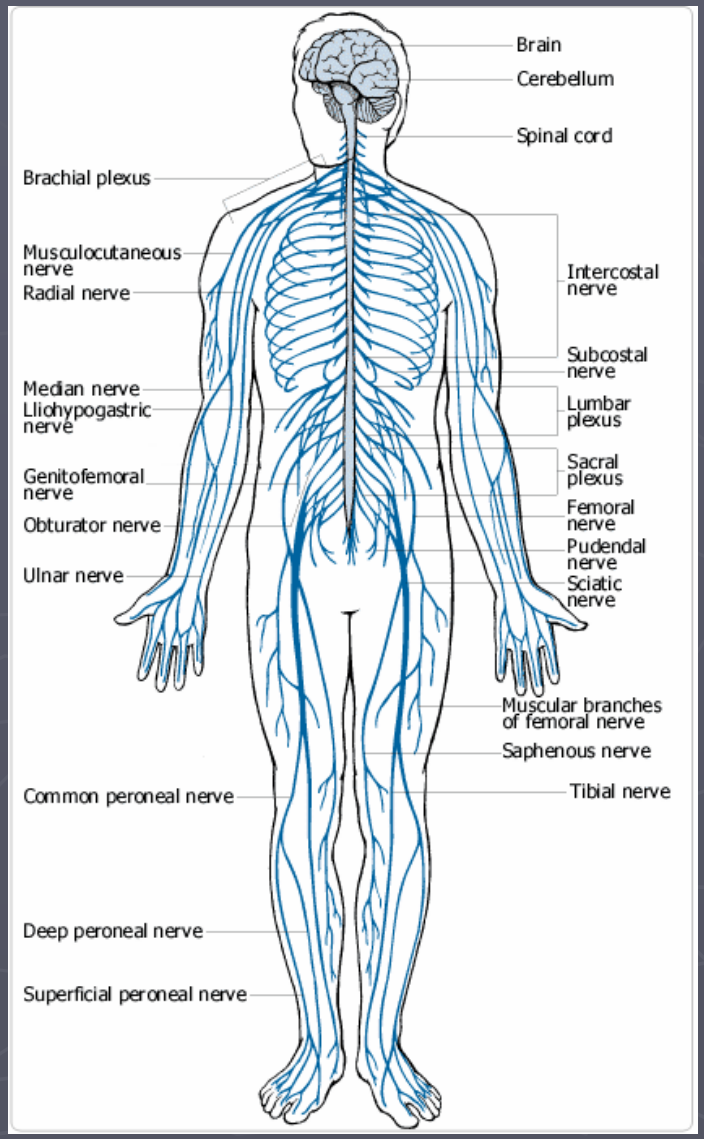
Predicting post-synaptic activity in proteins

Why Post-Synaptic Protein?

Neurological Disorder

- Alzheimer Disease
- Wilson Disease
- Etc.





Solution

BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii19–ii25
doi:10.1093/bioinformatics/bti1102

Databases

Predicting post-synaptic activity in proteins with data mining

Gisele L. Pappa[1], Anthony J. Baines[2] and Alex A. Freitas[1,*]

[1]Computing Laboratory, University of Kent, Canterbury CT2 7NF, UK and [2]Department of Biosciences, University of Kent, Canterbury CT2 7NJ, UK

ABSTRACT

Summary: The bioinformatics problem being addressed in this paper is to predict whether or not a protein has post-synaptic activity. This problem is of great intrinsic interest because proteins with post-synaptic activities are connected with functioning of the nervous system. Indeed, many proteins having post-synaptic activity have been functionally characterized by biochemical, immunological and proteomic exercises. They represent a wide variety of proteins with functions in extracellular signal reception and propagation through intracellular apparatuses, cell adhesion molecules and scaffolding proteins that link them in a web. The challenge is to automatically discover features of the primary sequences of proteins that typically occur in proteins with post-synaptic activity but rarely (or never) occur in proteins without post-synaptic activity, and vice-versa. In this context, we used data mining to automatically discover classification rules that predict whether or not a protein has post-synaptic activity. The discovered rules were analysed with respect to their predictive accuracy (generalization ability) and with respect to their interestingness to biologists (in the sense of representing novel, unexpected knowledge).

Contact: A.A.Freitas@kent.ac.uk

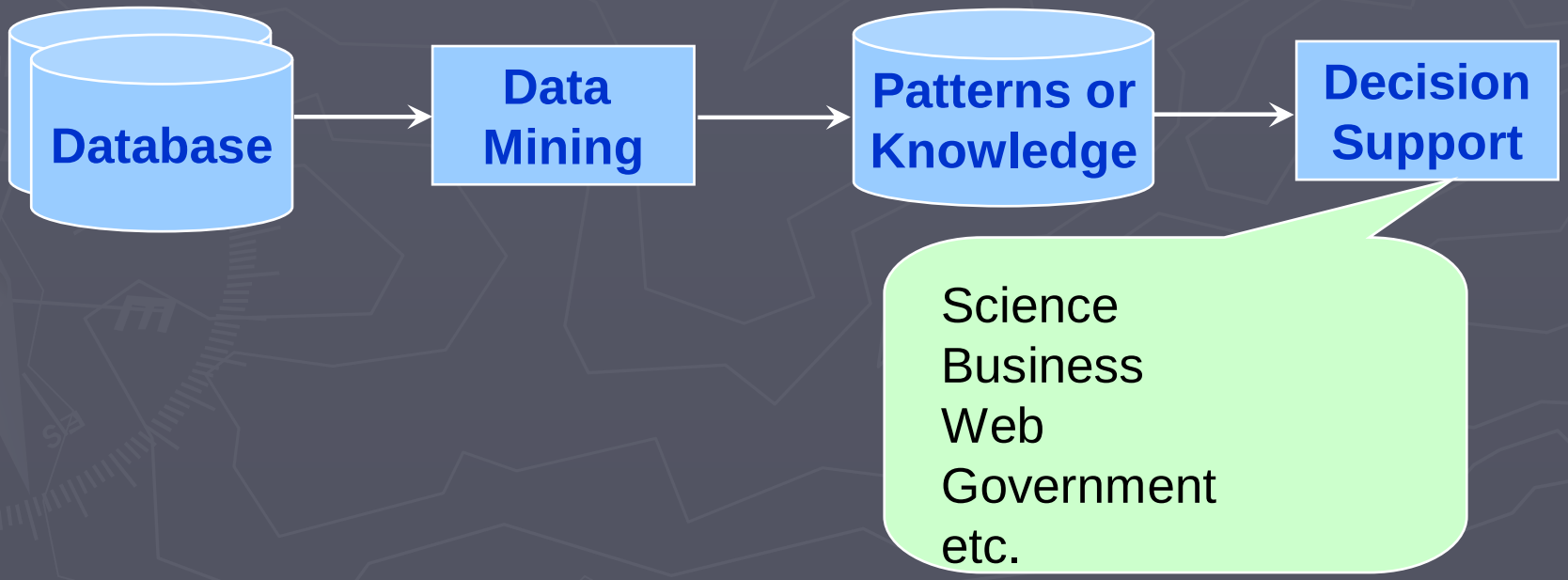
This paper proposes a data mining approach to the problem of predicting whether or not a protein has post-synaptic activity, based on features of the protein's primary sequence. The proposed approach will be described later, in Sections 3 and 4. In this Introduction we only emphasize a major difference between the proposed data mining approach and a more traditional bioinformatics approach for predicting protein function, as follows.

In general, the approach most used to predict the function of a new protein—for which we know only its sequence—consists of performing a similarity search in a protein database. In essence, the program finds the most similar protein(s) to the new protein, and if that similarity is higher than a threshold, the function of the most similar protein is transferred to the new protein. Although this approach is very useful in many cases, it also has some limitations, as follows.

First, it is well-known that two proteins might have very similar sequences and perform different functions, or have very different sequences and perform the same or similar function (Syed and Yona, 2003; Gerlt and Babbitt, 2000). Second, the proteins being compared may be similar in regions of the sequence that are not determinants of function (Schug *et al.*, 2002). Third, the prediction of function is

What is Data Mining?

Data mining is the process of searching large volumes of data for patterns, correlations and trends



Market Basket Analysis

An example of market basket transactions.

TID -- Items

1.{Bread, Milk}
2.{Bread, Diapers, Beer, Eggs}
3.{Milk, Diapers, Beer, Cola}
4.{Bread, Milk, Diapers, Beer}
5.{Bread, Milk, Diapers, Cola}

WAL★MART

Why is Data Mining Popular?

Data Flood

- Bank, telecom, other business transactions ...
- Scientific Data: astronomy, biology, etc
- Web, text, and e-commerce

Why Data Mining Popular?

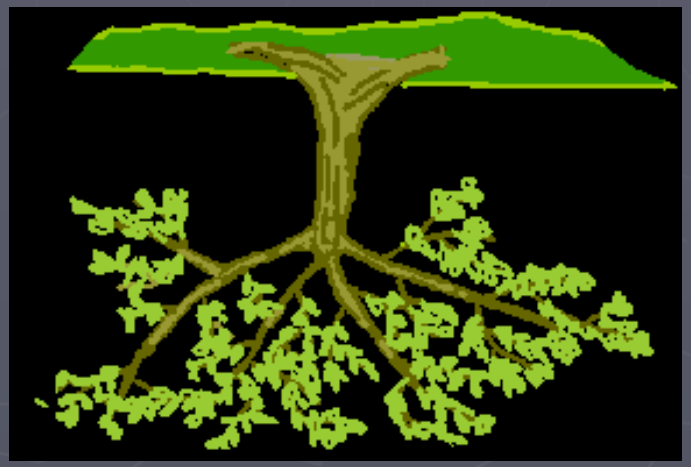
- Limitations of Human Analysis
  - Inadequacy of the human brain when searching for complex multifactor dependencies in data

Tasks Solved by Data Mining

- Learning association rules
- Learning sequential patterns
- Classification
- Numeric prediction
- Clustering
- etc.

Decision Trees

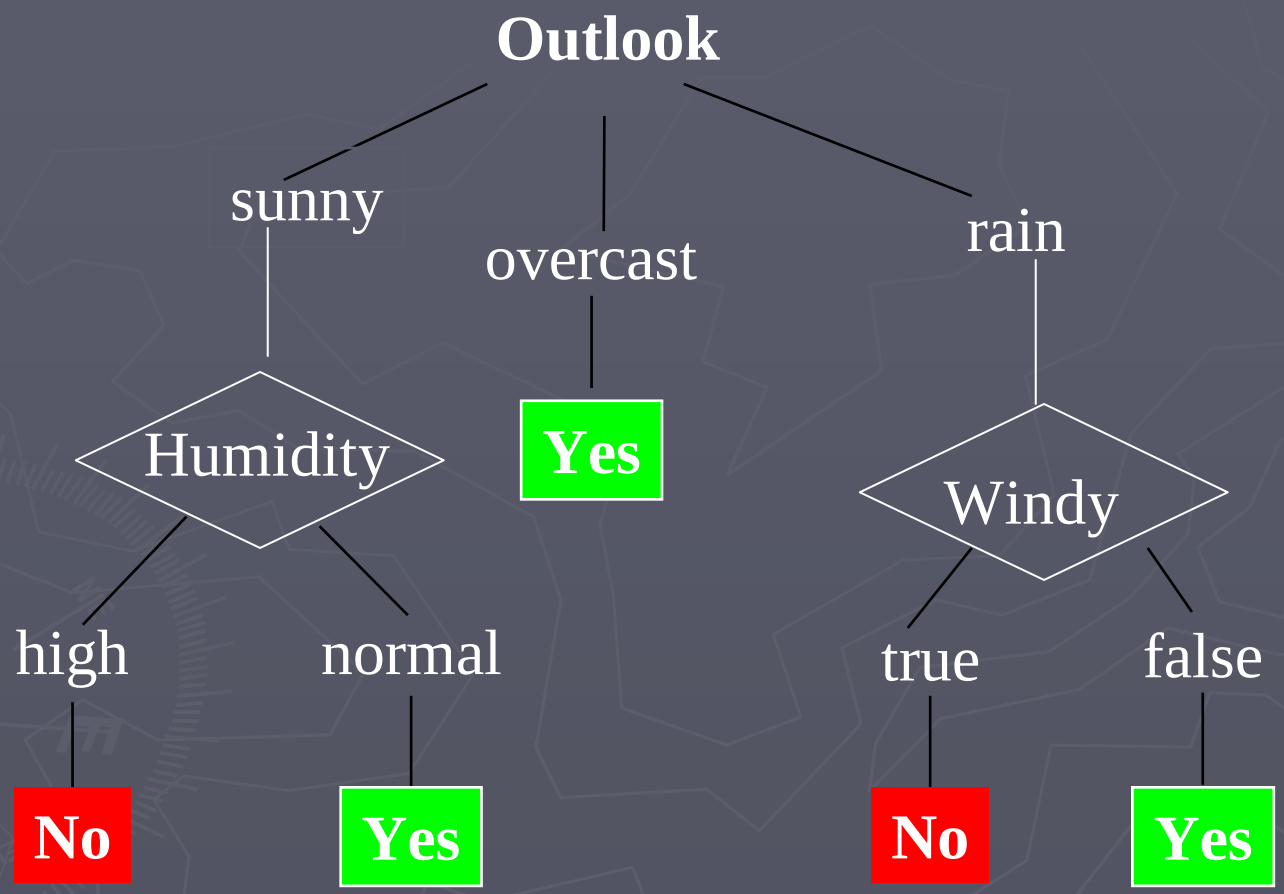
- A decision tree is a predictive model
- It takes as input an object or situation described by a set of properties (predictive properties), and outputs a yes/no decision (class)



Decision Tree Example

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Decision Tree Example



Choosing the Splitting Attribute

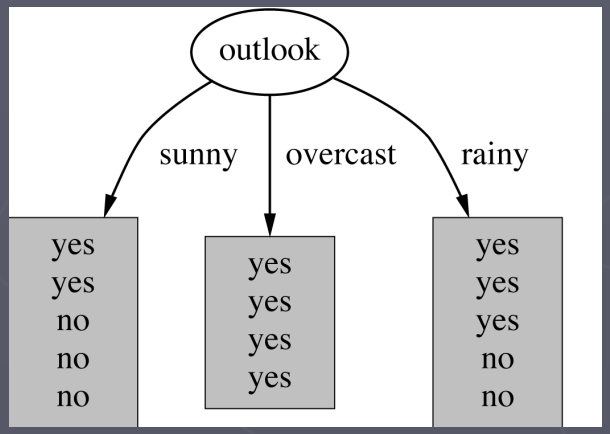
- At each node, available attributes are evaluated on the basis of separating the classes of the training examples. A goodness function is used for this purpose.
- Typical goodness functions:
  - information gain (ID3/C4.5)
  - information gain ratio
  - gini index

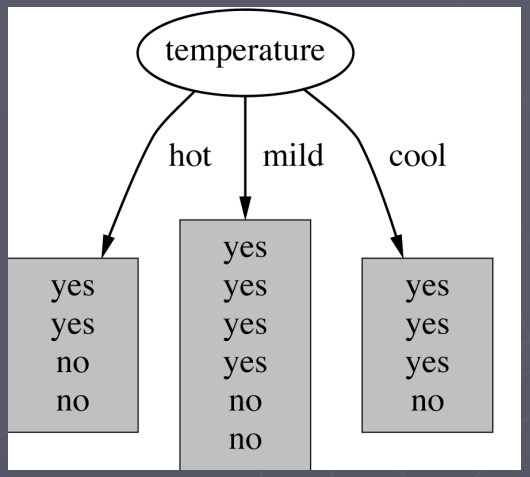
Sunny

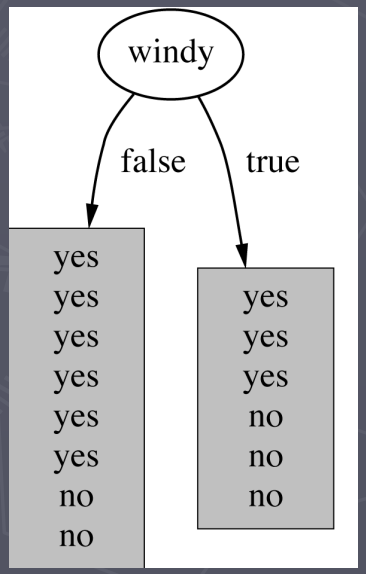
Overcast

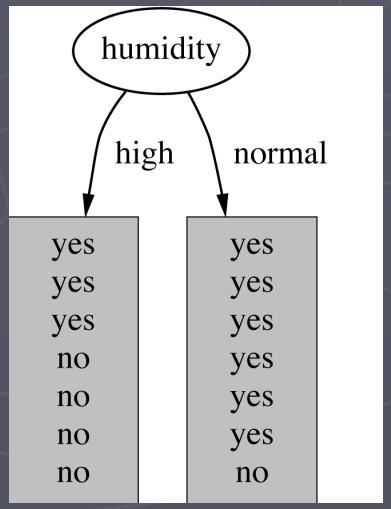
Raining

Which Attributes to Select?









A Criterion for Attribute Selection

- Which is the best attribute?
  - The one which will result in the smallest tree
  - Heuristic: choose the attribute that produces the "purest" nodes

Information Gain

- Popular impurity criterion: information gain
  - Information gain increases with the average purity of the subsets that an attribute produces
- Strategy: choose attribute that results in greatest information gain

Computing Information

- Information is measured in bits
  - Given a probability distribution, the info required to predict an event is the distribution's entropy
- Formula for computing the entropy:

$$\text{entropy}(p_1, p_2, \ldots, p_n) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n$$

Computing Information Gain

- Information gain:

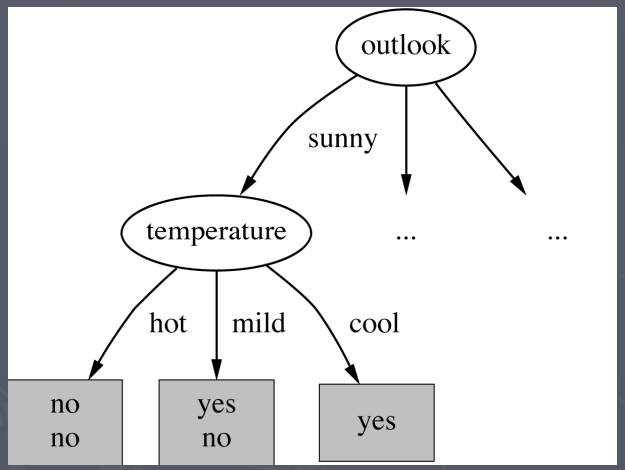
  (information before split) – (information after split)

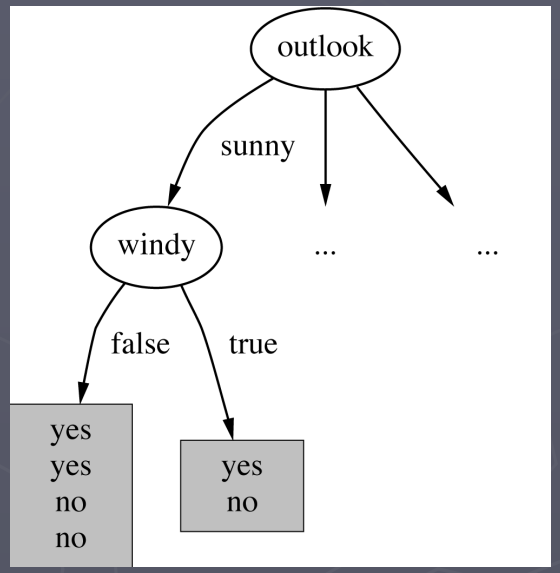
$$\text{gain("Outlook")} = \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) = 0.940 - 0.693$$
$$= 0.247 \text{ bits}$$

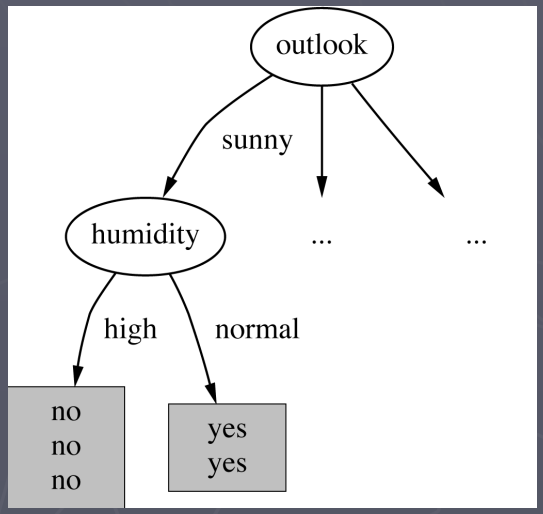
- Information gain for attributes from weather data:

$$\text{gain("Outlook")} = 0.247 \text{ bits}$$
$$\text{gain("Temperature")} = 0.029 \text{ bits}$$
$$\text{gain("Humidity")} = 0.152 \text{ bits}$$
$$\text{gain("Windy")} = 0.048 \text{ bits}$$

Continuing to Split





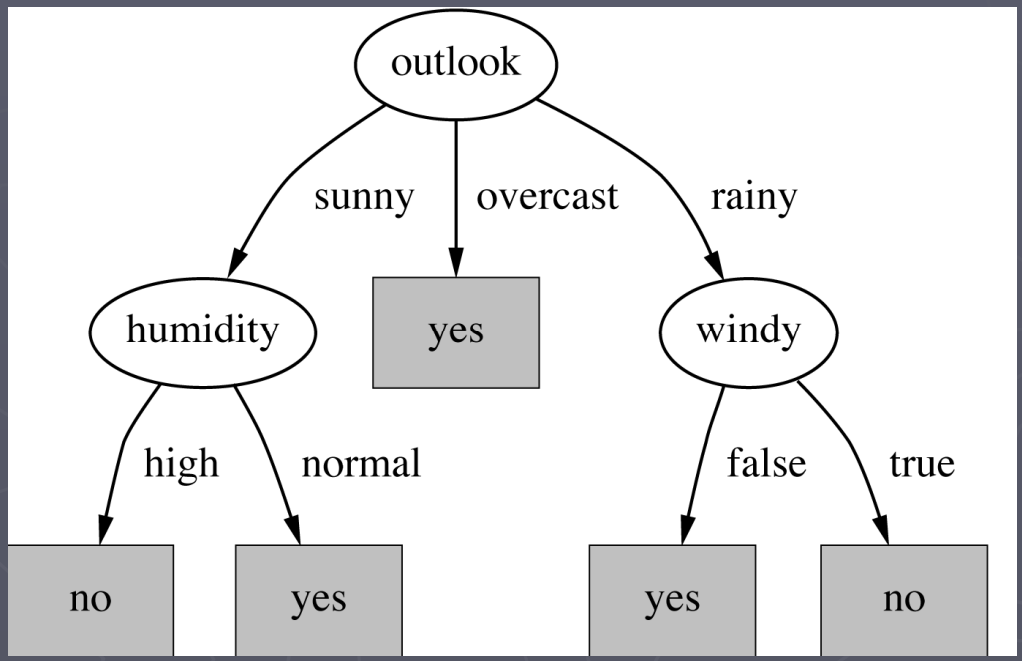


gain("Temperature") = 0.571 bits

gain("Humidity") = 0.971 bits

gain("Windy") = 0.020 bits

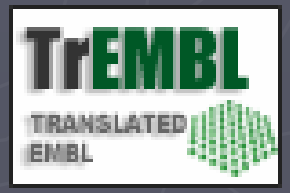
The Final Decision Tree



Splitting stops when data can’t be split any further

Uniprot
[the Universal Protein Resource]

Central repository of protein sequence and function created by joining information contained in Swiss-Prot, TrEMBL and PIR



Prosite

PROSITE
Database of protein families and domains

It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs

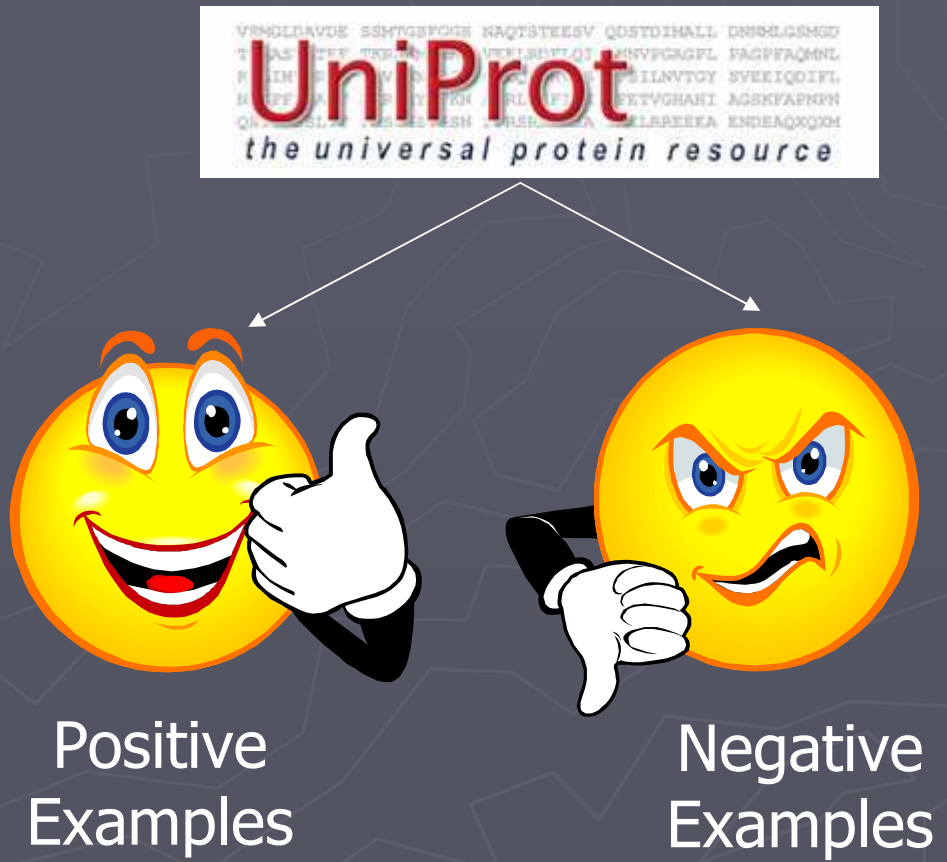
Predicting post-synaptic activity in proteins

- Classes
  - Positive Example = {proteins with post-synaptic activity}
  - Negative Example = {proteins without post-synaptic activity}
- Predictor attribute
  - Prosite patterns

First Phase

select positive and negative Examples:

- carefully select relevant proteins from the UniProt database



Positive Example

- Query-1: Post-synaptic AND !toxin
  - All Species – To maximize the number of examples in the data to be mined
  - !toxin – Several entities in UniProt/SwissProt refer to the toxin alpha-latrotoxin

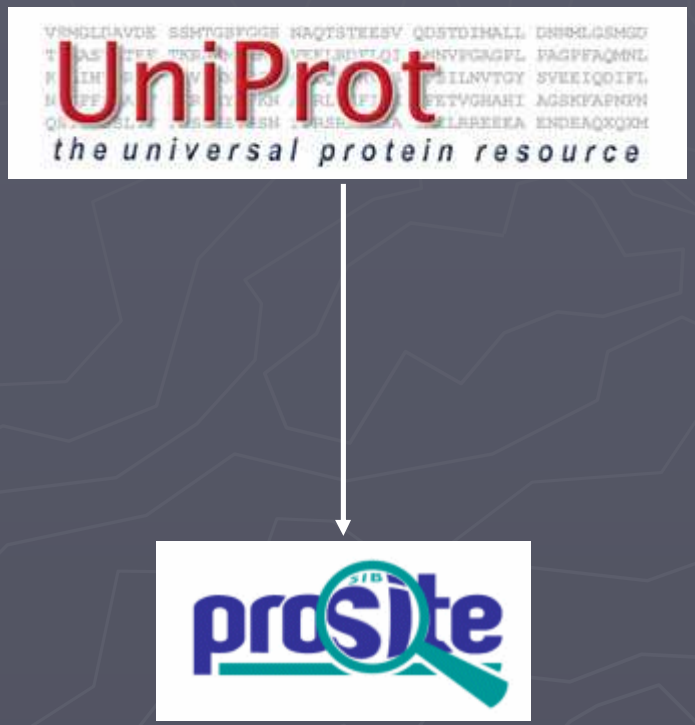
Negative Examples

- Query-2: Heart !(result of query 1)
- Query-3: Cardiac !(result of queries1,2)
- Query-4: Liver !(result of queries 1,2,3)
- Query-5: Hepatic !(result of queries 1,2,3,4)
- Query-6: Kidney !(result of queries 1,2,3,4,5)

Second Phase

generating the predictor attribute:

- use the link from UniProt to Prosite database to create the predictor attributes

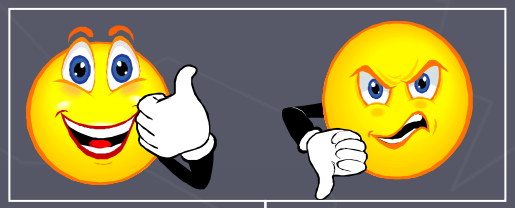


Predictor Attributes

Must have a good predictive power

Facilitate easy interpretation by biologists

Generating The Predictor Attribute



Occurred at least in two proteins

Pattern (NOT Profile)

NO

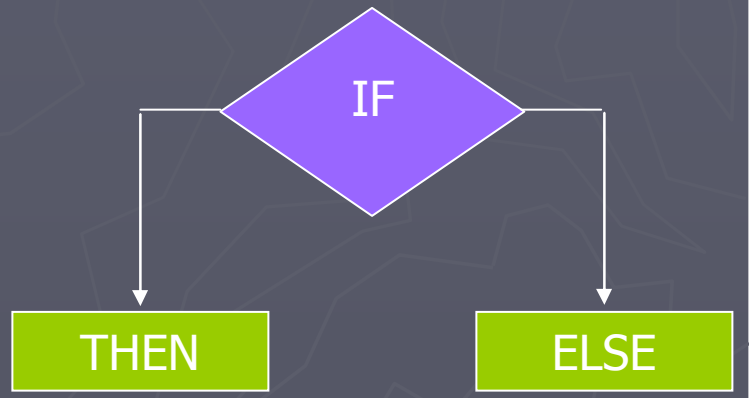
Prosite Entry?

YES

Remove from the Dataset

Data Mining Algorithm

- C4.5 rules induction algorithm (Quinlan, 1993)
  - Generates easily interpretable classification rules of the form: IF (condition) THEN (class)
  - This kind of rule has the intuitive meaning

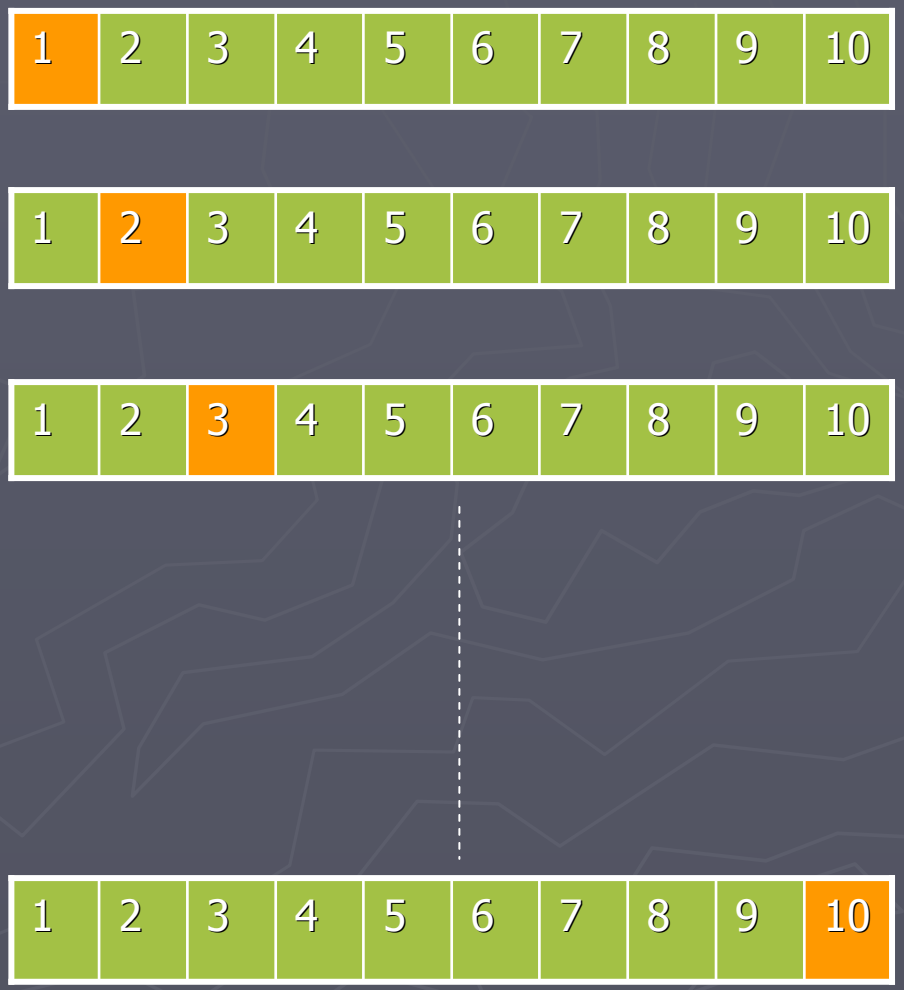


Classification Rules

Id	Classification rule
32	IF (NEUROTR_ION_CHANNEL = yes) THEN (class = yes)
19	IF (CADHERIN_1 = yes) AND (920 < seq_length <= 1025) THEN (class = yes)
29	IF (GUANYLATE_KINASE_1 = yes) AND (78928 < mol_weigth <= 113386) THEN (class = yes)
34	IF (43_KD_POSTSYNAPTIC = yes) THEN (class = yes)
35	IF (NA_DICARBOXYL_SYMP_1 = yes) THEN (class = yes)
8	IF (CARBOXYLESTERASE_B_2 = yes) AND (seq_length > 828) THEN (class = yes)
33	IF (DYNAMIN = yes) THEN (class = yes)
6	IF (LIPASE_SER = yes) AND (seq_length > 699) THEN (class = yes)
10	IF (G_PROTEIN_RECEP_F1_1 = yes) AND (11287 < mol_weigth <= 14398) THEN (class = yes)
14	IF (C1Q = yes) AND (seq_length <= 194) THEN (class = yes)
23	IF (A4_EXTRA = yes) AND (BPTI_KUNITZ_1 = no) THEN (class = yes)
26	IF (PPTA = yes) AND (G_PROTEIN_RECEP_F2_1 = no) AND (seq_length > 895) THEN (class = yes)
17	IF (SER_THR_PHOSPHATASE = yes) AND (seq_length > 318) THEN (class= yes)

Predictive Accuracy

- Well-known 10-fold cross-validation procedures (Witten and Frank 2000)
  - Divide the dataset into 10 partitions with approximately same number of examples (proteins)
  - In i-th iteration, 1=1,2,…,10, the i-th partition as test set and the other 9 partitions as training set



Results

- Sensitivity
  - TPR = TP/(TP+FN)
  - Average values (over the 10 iterations of cross-validation procedure) = 0.85
- Specificity
  - TNR = TN/(TN+FP)
  - Average values (over the 10 iterations of cross-validation procedure) =0.98

Questions?