Data Mining

- Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.

- Common data mining tasks:
  - Classification
  - Estimation
  - Prediction
  - Affinity Grouping
  - Clustering
  - Description

Knowledge Discovery

Knowledge is a pattern that exceeds certain threshold of interestingness.

Factors that contribute to interestingness:
- coverage
- confidence
- statistical significance
- simplicity
- unexpectedness
- actionability

Classification

- Classifying observations into different categories given characteristics

Estimation

- Rules that explain how to estimate a value given characteristics
Prediction
• Rules that explain how to predict a future value or classification, given characteristics

Affinity Grouping
• Grouping by relations (not by characteristics)

Clustering
• Segmenting a diverse population into more similar groups
• In clustering, there are no pre-defined classes and no examples. Records are grouped together by some similarity measure.

Knowledge Discovery

Scientific Models

Artificial Intelligence in Biosciences
- Neural Networks (NN)
- Genetic Algorithms (GA)
- Formal Grammars (FG)
Neural Networks

- interconnected assembly of simple processing elements (units or nodes)
- nodes functionality is similar to that of the animal neuron
- processing ability is stored in the inter-unit connection strengths (weights)
- weights are obtained by a process of adaptation to, or learning from, a set of training patterns

\[
Y = \begin{cases} \text{1 if } \sum w_i x_i > \Theta \\ 0 \text{ otherwise} \end{cases}
\]

Learning process: \( \Delta w_i = (T_p - Y_p)x_i \)

Artificial Intelligence in Biosciences

Neural Networks (NN)
Genetic Algorithms (GA)
Formal Grammars (FG)
Genetic Algorithms

Search or optimization methods using simulated evolution. Population of potential solutions is subjected to natural selection, crossover, and mutation.

- Choose initial population
- Evaluate each individual's fitness
- Repeat
  - Select individuals to reproduce
  - Mate pairs at random
  - Apply crossover operator
  - Apply mutation operator
  - Evaluate each individual's fitness
  - Until terminating condition

Crossover

Parent A
Parent B
Child AB
Child BA

Mutation

Parents
10 00 01 00 10
10 00 00 01 11

Children
10 00 10 01 11
10 00 01 00 10

GA simulation of folding

Artificial Intelligence in Biosciences

Neural Networks (NN)
Genetic Algorithms (GA)
Formal Grammars (FG)

Membrane binding domain of Blood Coagulation Factor VIII (J Moult)
Grammars and Language

**grammar**  
1. the study of the way the sentences of a language are constructed

   4. *Generative Gram*. a device, as a body of rules, whose output is all of the sentences that are permissible in a given language, while excluding all those that are not permissible.

   Random House Unabridged Dictionary

Language Components

Semantics (meaning)

Syntax (structure, form)

Semantics

Derived from syntax

Semantic content derived from vocabulary within a context

Vocabulary element has its own meanings

   dictionary lookup

   meanings depending on context

   Time flies like an arrow

   Fruit flies like a banana

Formal Grammars

**formal grammar**  
a means for specifying the syntactic structure of natural language by a set of transformation functions

Chomsky hierarchy (for string grammars)

   type 0: phrase structure

   type 1: context sensitive

   type 2: context free

   type 3: regular

(SCFG)

Hidden Markov models

Chomsky, *Syntactic Structures* (1957)

Markov Model (or Markov Chain)

Probability for each character based only on several preceding characters in the sequence

   # of preceding characters = order of the Markov Model

   Probability of a sequence


Hidden Markov Models

   \[ \begin{array}{cccccc}
   A & T & C & T & A & G \\
   \end{array} \]

   Observed sequences

   A 0.7 A 0.1 C 0.8 A 0.4 A 0.8 C 0.3

   T 0.3 T 0.9 G 0.2 T 0.6 T 0.2 G 0.7

   Probabilistic model - true state is unknown
Hidden Markov Models

States -- well defined conditions
Edges -- transitions between the states

ATGAC
ATTAC
ACGAC
ACTAC

Each transition assigned a probability.

Probability of the sequence:
- single path with the highest probability -- Viterbi path
- sum of the probabilities over all paths -- Baum-Welch method

Hidden Markov Model for Exon and Stop Codon (VEIL Algorithm)

Hidden Markov Model in Structural Analysis

A hidden Markov model consists of Markov states connected by directed transitions. Each state emits an output symbol, representing sequence or structure. There are four categories of emission symbols in our model: b, d, r, and c, corresponding to amino acid residues, three-state secondary structure, backbone angles (discretized into regions of phi/psi space) and structural context (e.g. hairpin versus diverging turn, middle versus end-strand), respectively.

Adopted from C.Bystroff et al, 2000
Artificial Intelligence in Biosciences

Other machine learning algorithms:
- Support vector machines
- Decision trees
- Random forests

Support Vector Machines (SVM) Algorithm

Decision surface is a hyperplane (line in 2D, plane in 3D, etc.) in feature space

Define what an optimal hyperplane is (in way that can be identified in a computationally efficient way): maximize margin

Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications

Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space

Applications of ML methods

Mapping in topological space