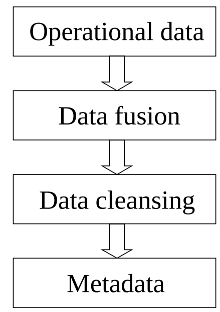
BINF 630: Introduction to Bioinformatics

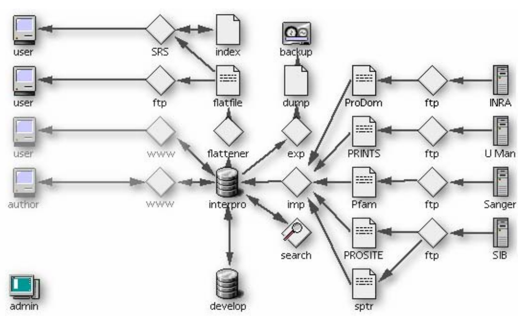
Iosif Vaisman

Email: ivaisman@gmu.edu

Data Warehouse



InterPro Database



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

- Directed and Undirected KD
- Directed KD
  - Purpose: Explain value of some field in terms of all the others
  - Method: We select the target field based on some hypothesis about the data. We ask the algorithm to tell us how to predict or classify it
  - Similar to hypothesis testing (e.g., in regression modeling) in statistics

Knowledge Discovery

- Undirected KD
  - Purpose: Find patterns in the data that may be interesting
  - Method: clustering, affinity grouping
  - Closest to ideas of machine learning in artificial intelligence
- Comparison
  - UKD helps us to recognize relationships & DKD helps us to explain them

Classification

- Classifying observations into different categories given characteristics

Prediction

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

Estimation

- Rules that explain how to estimate a value 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.

Scientific Models

Physical models -- Mathematical models

Mechanistic models

Mechanism

Predictive power
Elegance
Consistency

Stochastic models

Black box

Predictive power

Artificial Intelligence in Biosciences

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

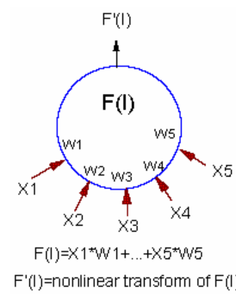
Artificial Intelligence in Biosciences

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

Neural Networks

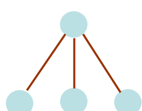
- interconnected assembly of simple processing elements (units or nodes)
- nodes functionality is similar to that of the animal neuron
- processing ability is stored in the inter-unit connection strengths (weights)
- weights are obtained by a process of adaptation to, or *learning* from, a set of training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

Neural Networks



Neural Networks

Perceptron



Output layer

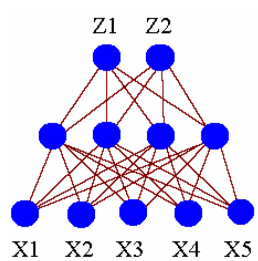
Input layer

$$Y = \begin{cases} 1 \text{ if } \Sigma\, w_i i_i > \Theta \\ 0 \text{ otherwise} \end{cases}$$

Learning process: $\Delta w_i = (T_p - Y_p) i_{pi}$

Neural Networks

Hierarchical neural network

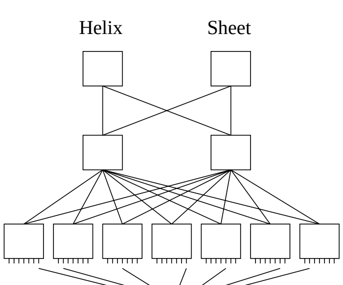


Output layer

Hidden layer

Input layer

Neural Networks



Output layer (2 units)

Hidden layer (2 units)

Input layer (7x21 units)

MKFGNFLLTYQP [PELSQTE] VMKRLVNLGKASEGC...

Artificial Intelligence in Biosciences

Neural Networks (NN)

Genetic Algorithms (GA)

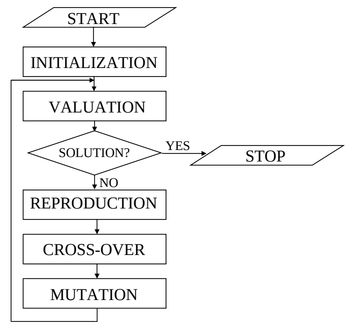
Formal Grammars (FG)

Genetic Algorithms

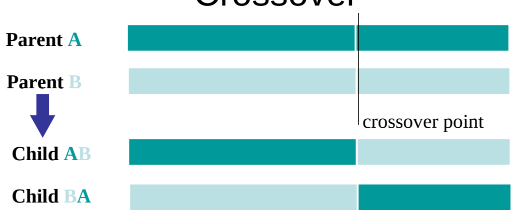
Search or optimization methods using simulated evolution.
Population of potential solutions is subjected to
natural selection, crossover, and mutation

```
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
```

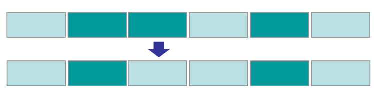
Genetic Algorithms



Crossover



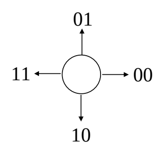
Mutation

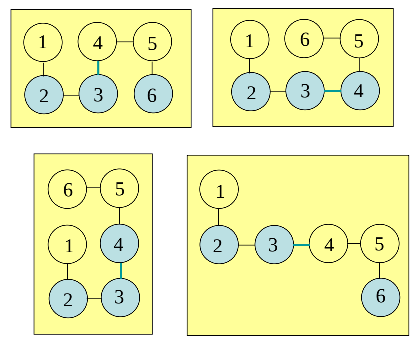


Genetic Algorithms Applications

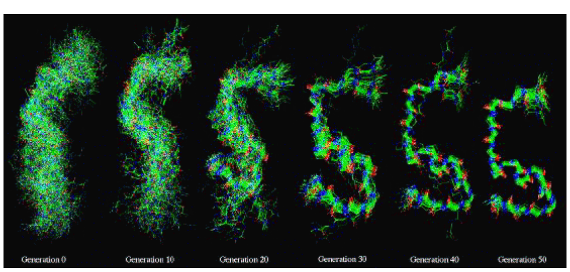
Parents
10 00 01 00 10
10 00 00 01 11

Children
10 00 10 01 11
10 00 01 00 10





GA simulation of folding



Membrane binding domain of Blood Coagulation Factor VIII (J.Moult)

Artificial Intelligence in Biosciences

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

Grammars and Language

gram•mar *n.*

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)

Language Syntax

Alphabet
- Primitive elements
- Letters, phonemes

Vocabulary
- Elements composed from the alphabet
- Words, phrases, sentences,…

Grammar
- Legal composition of vocabulary
- Rules, operators

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 (SCFG)
- type 3: regular (Hidden Markov models)

Chomsky, *Syntactic Structures* (1957)

Markov Model (or Markov Chain)

A — T — C — T — A — G

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

P(s) = P[A] P[A,T] P[A,T,C] P[T,C,T] P[C,T,A] P[T,A,G]

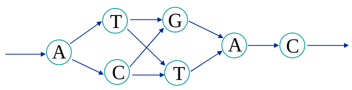
Hidden Markov Models

	A	T	C	T	A	G
Observed frequencies	A 0.7 T 0.3	A 0.1 T 0.9	C 0.8 G 0.2	A 0.4 T 0.6	A 0.8 T 0.2	C 0.3 G 0.7

Probablistic model - true state is unknown

Hidden Markov Models

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

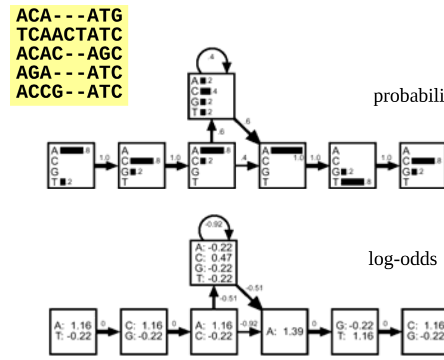


ATGAC
ATTAC
ACGAC
ACTAC

Each transition asigned 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 Models

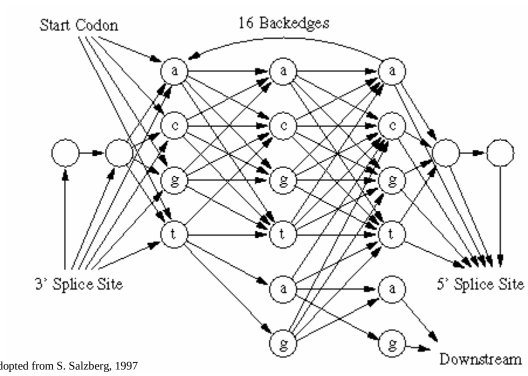


probabilities

log-odds $(\log \frac{P(S)}{0.25^L})$

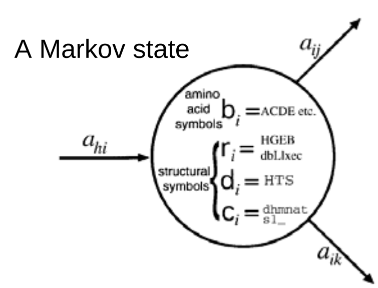
Adopted from Anders Krogh, 1998

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



Adopted from S. Salzberg, 1997

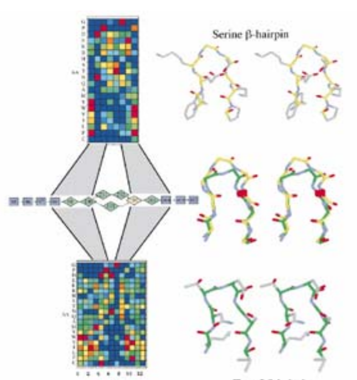
Hidden Markov Model in Structural Analysis



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

Hidden Markov Model in Structural Analysis



HMM topology from merging of two motifs, the extended Type-I hairpin motif and the Serine hairpin.

Adopted from C.Bystroff et al, 2000
JMB, **301**, *173*

Artificial Intelligence in Biosciences

Other machine learning algorithms:

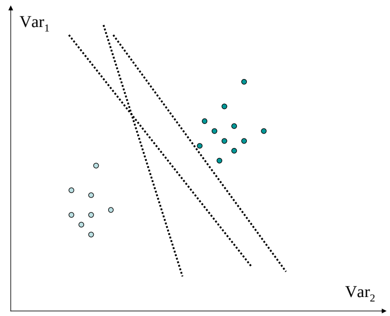
- **Support vector machines**
- **Decision trees**
- **Random forests**

Support Vector Machines (SVM) Algorithm

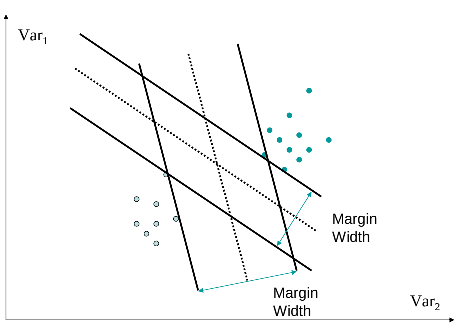
Decision surface is a hyperplane (line in 2D, plane in 3D, etc.) in **feature** space

- 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

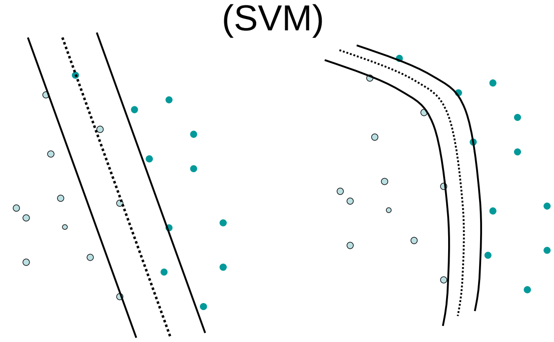
Support Vector Machines (SVM)



Support Vector Machines (SVM)



Support Vector Machines (SVM)



Linear SVM

Non-linear SVM