BINF 630: Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

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)

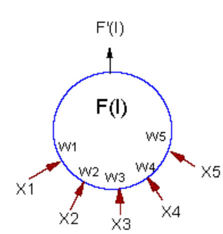
Artificial Intelligence in Biosciences

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

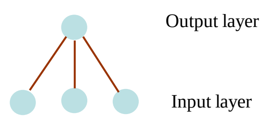
Neural Networks



F(I)=X1*W1+...+X5*W5
F'(I)=nonlinear transform of F(I)

Neural Networks

Perceptron

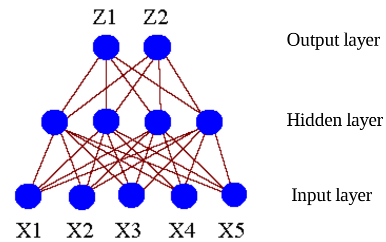


$$Y = \begin{cases} 1 \text{ if } \Sigma\, w_i i_i > \Theta \\ 0 \text{ otherwise} \end{cases}$$

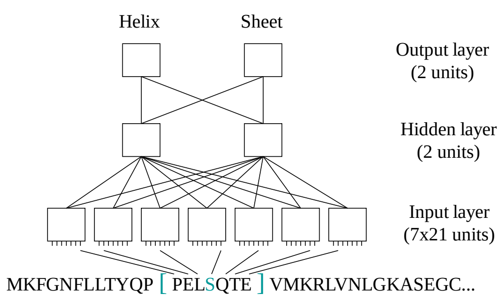
Learning process: $\Delta w_i = (T_p - Y_p) i_{pi}$

Neural Networks

Hierarchical neural network



Neural Networks



Artificial Intelligence in Biosciences

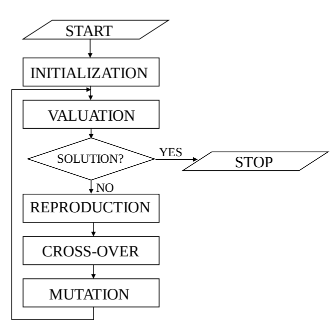
Neural Networks (NN)
Genetic Algorithms (GA)
Formal Grammars (FG)

Genetic Algorithms

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

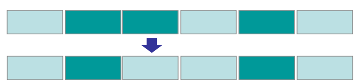
```
choose initial population
evaluate each individual's fitness
repeat
        select individuals to reproduce
        mate pairs at random
        apply crossover operator
        apply mutation operator
        evaluate each individual's fitness
until terminating condition
```

Genetic Algorithms

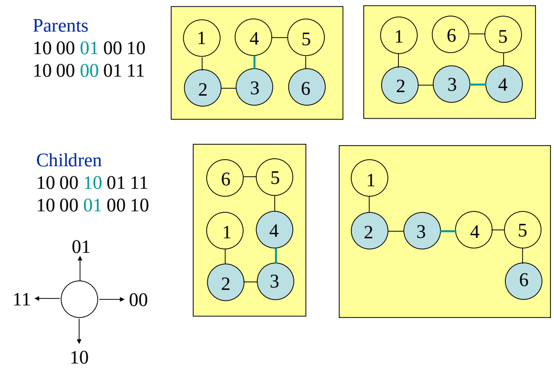


Crossover

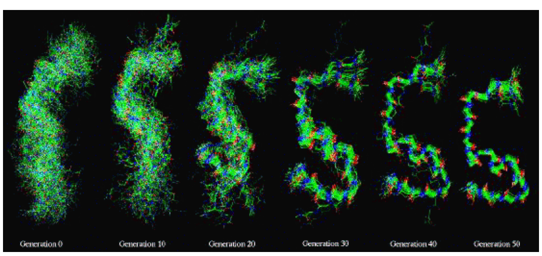
Mutation



Genetic Algorithms Applications



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

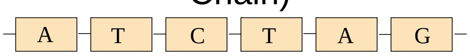
a means for specifying the syntactic structure of natural language by a set of transformation functions

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)



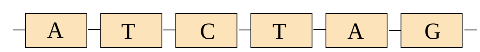
Probability for each character based only on several preceding characters in the sequence

\# 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

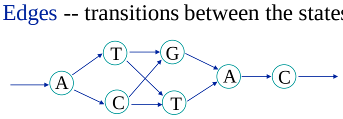


Observed frequencies	A 0.7 T 0.3	A 0.1 T 0.9	C 0.8 G 0.2	A 0.4 T 0.6	A 0.8 T 0.2	C 0.3 G 0.7

Probablistic model - true state is unknown

Hidden Markov Models

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



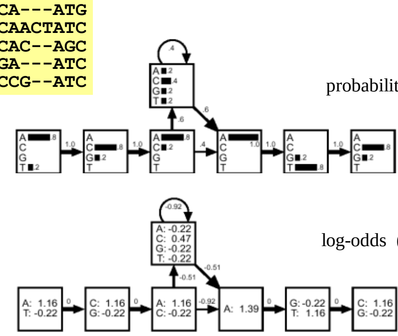
ATGAC
ATTAC
ACGAC
ACTAC

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

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

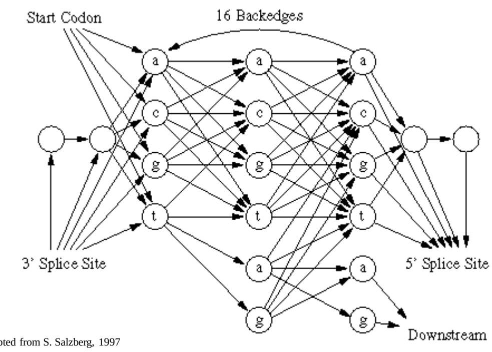


probabilities

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

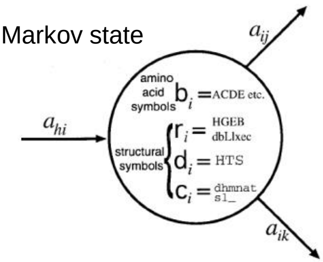
Adopted from Anders Krogh, 1998

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



Adopted from S. Salzberg, 1997

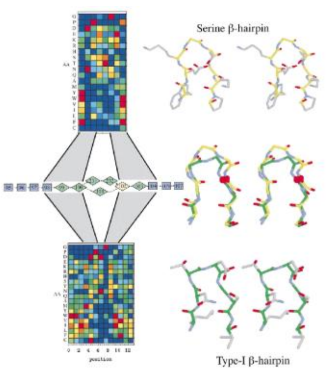
Hidden Markov Model in Structural Analysis



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*

Comparison of AI methods

Comparison of machine learning approaches according to a number of model characteristics

Characteristic	GLM	CART	ANN	EA
Data Requirements				
Accommodate "mixed" data types	Low	High	Low	Moderate
Accommodate missing values of predictors	Low	High	Low	Low
Insensitive to monotonic transformations of predictors	Low	High	Moderate	Moderate
Robust to outliers in predictors	Low	Moderate	Moderate	Moderate
Insensitive to irrelevant predictors	Low	High	Moderate	Moderate
Modeling Process				
Automation (i.e., low degree of user involvement)	High	Moderate	Moderate	Low
Transparency of the modeling process	High	Moderate	Low	Low
Ability to model nonlinear relationships	Low	Moderate	High	High
Accommodate interactions among predictors	Low	Moderate	High	High
Model Output				
Explanatory insight and variable interpretability	High	Moderate	Moderate	Low
Predictive power	Low	Moderate	High	High
Software Availability and Ease-of-Use	High	Moderate	Low	Low

Classification and regression trees (CARTs), artificial neural networks (ANNs), and evolutionary algorithms (EAs) are compared to the family of generalized linear models (GLMs) that are traditionally used in ecology. Comparisons are generalized to include both classification and prediction problems. Values are based on Hastie et al. (2001), peer-reviewed literature, and the personal experiences of the authors.

Olden et al., 2008

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

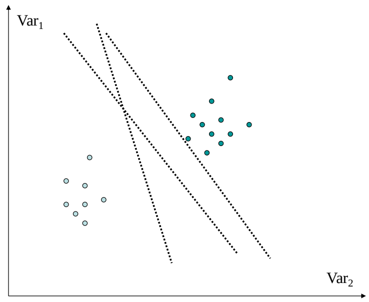
Define what an optimal hyperplane is (in way that can be identified in a computationally efficient way): maximize margin

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

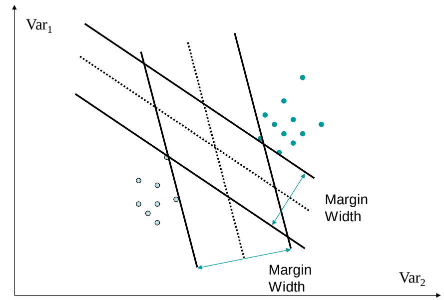
Aliferis & Tsamardinos

Support Vector Machines (SVM)



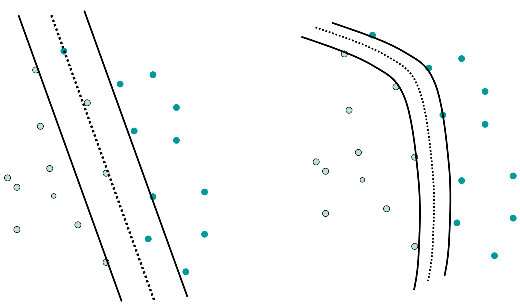
Aliferis & Tsamardinos

Support Vector Machines (SVM)



Aliferis & Tsamardinos

Support Vector Machines (SVM)

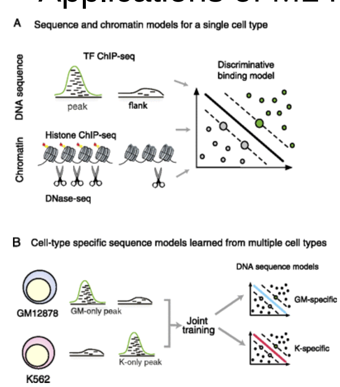


Linear SVM

Non-linear SVM

Aliferis & Tsamardinos

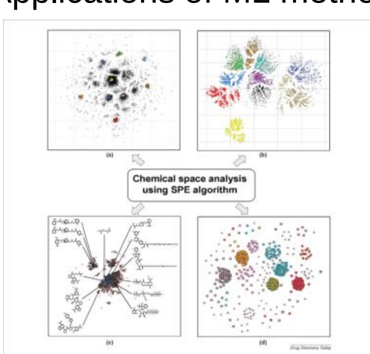
Applications of ML methods



Discrimination between regulatory ChIP-seq peaks and flanking regions within a single cell type using a support vector machine

A Arvey et al., 2012

Applications of ML methods



Mapping in topological space