

BINF 731

Protein Structure Analysis

Iosif Vaisman

2007

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)

Decision Tree (DT)

Support Vector Machines (SVM)

Artificial Intelligence in Biosciences

Neural Networks (NN)

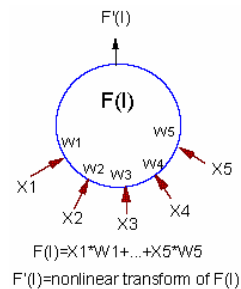
Genetic Algorithms (GA)

Formal Grammars (FG)

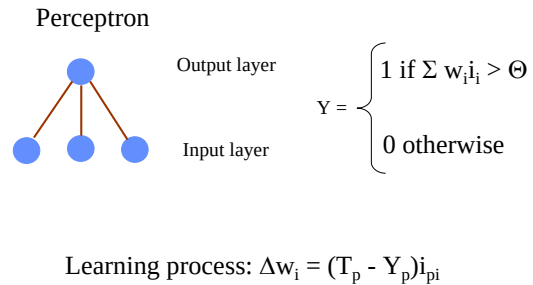
Neural Networks

- 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

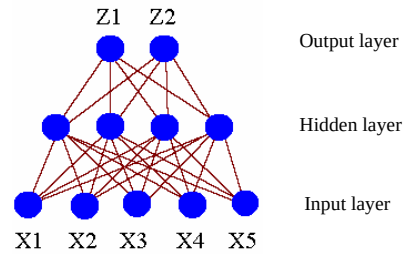


Neural Networks

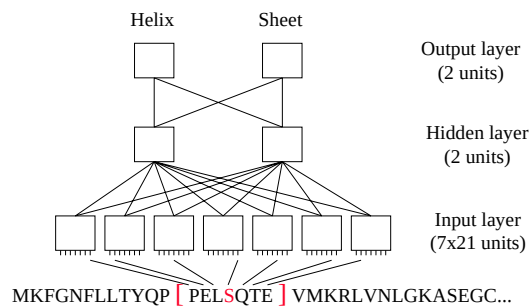


Neural Networks

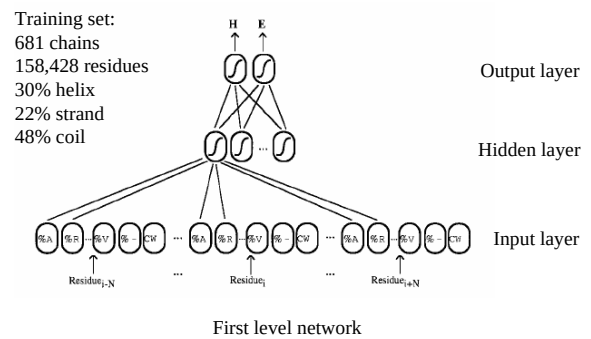
Hierarchical neural network



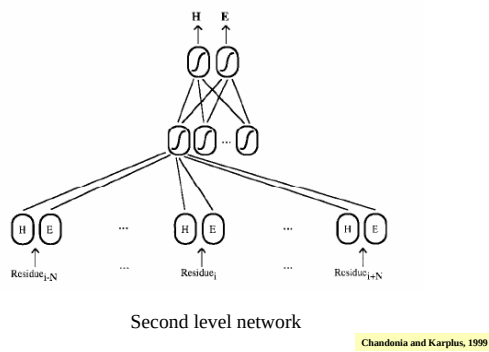
Neural Networks



Neural Networks



Neural Networks



Artificial Intelligence in Biosciences

Neural Networks (NN)

Genetic Algorithms (GA)

Formal Grammars (FG)

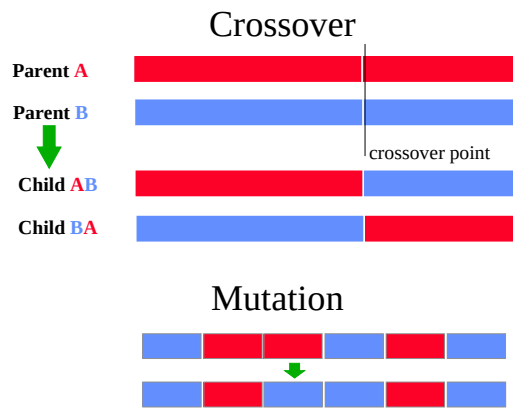
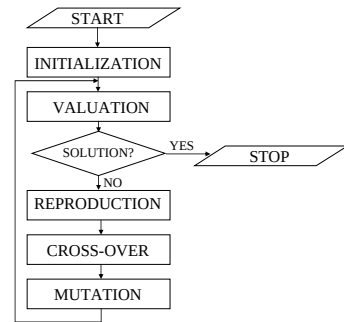
Genetic Algorithms

Search or optimization methods using simulated evolution.

Population of potential solutions is subjected to
natural selection, crossover, and mutation

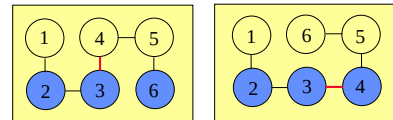
choose initial population
evaluate each individual's fitness
repeat
 select individuals to reproduce
 mate pairs at random
 apply crossover operator
 apply mutation operator
 evaluate each individual's fitness
until terminating condition

Genetic Algorithms

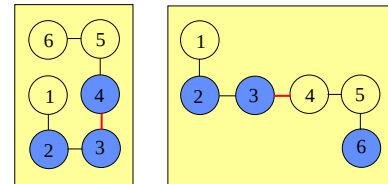
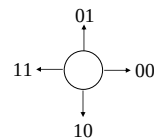


Genetic Algorithms Applications

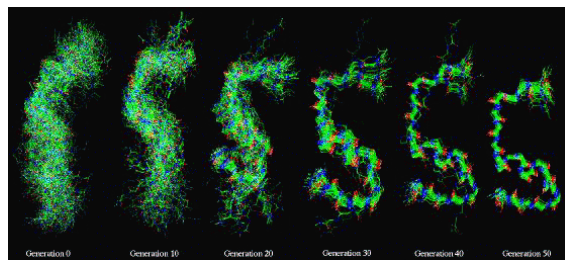
Parents
10 00 01 00 10
10 00 00 01 11



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)



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

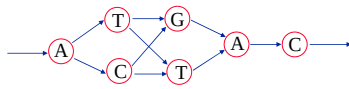


Observed frequencies	A	0.7	A	0.1	C	0.8	A	0.4	A	0.8	C	0.3
	T	0.3	T	0.9	G	0.2	T	0.6	T	0.2	G	0.7

Probabilistic model - true state is unknown

Hidden Markov Models

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



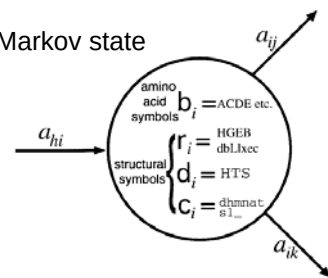
ATGAC
 ATTAC
 ACGAC
 ACTAC

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 in Structural Analysis

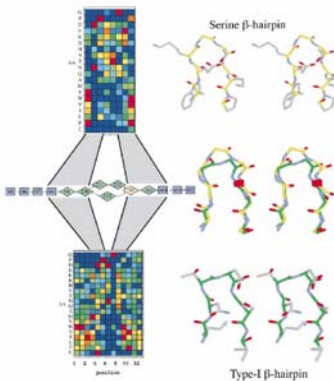
A Markov state



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, 381, 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

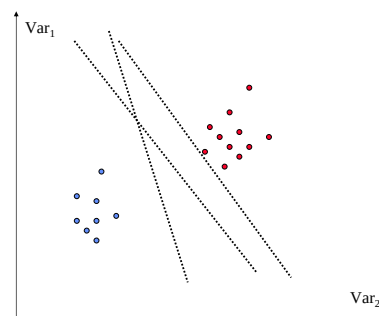
Define what an optimal hyperplane is (in way that can be identified in a computationally efficient way):
maximize margin

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

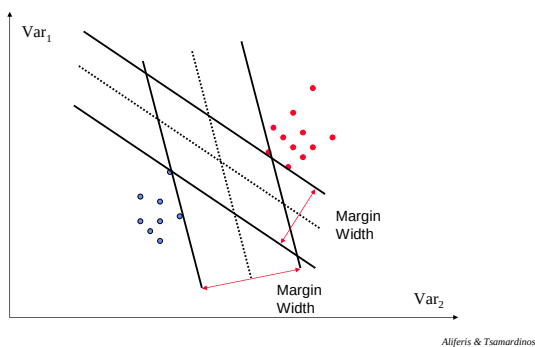
Aliferis & Tsamardinos

Support Vector Machines (SVM)

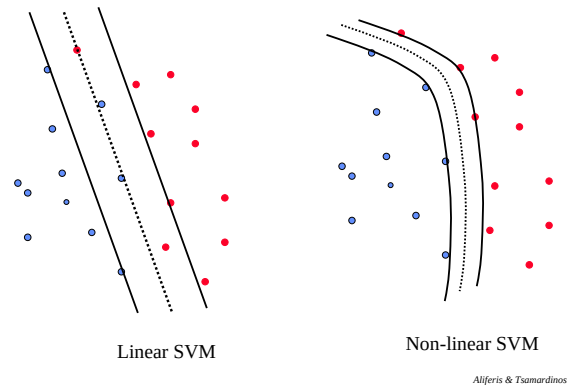


Aliferis & Tsamardinos

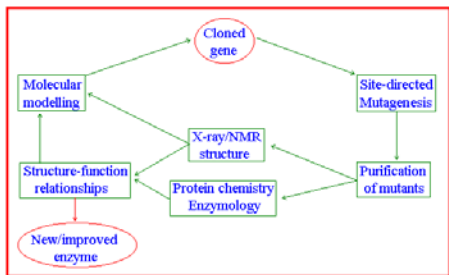
Support Vector Machines (SVM)



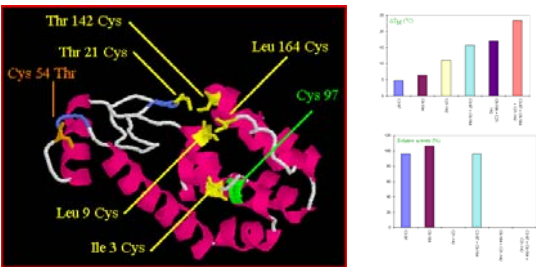
Support Vector Machines (SVM)



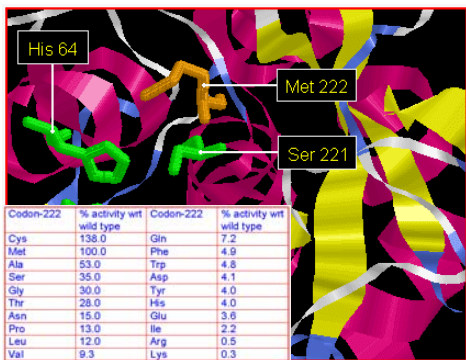
Protein Engineering



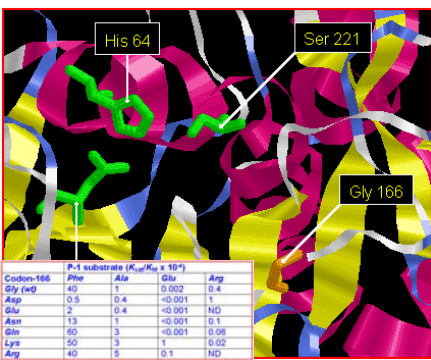
Protein Engineering



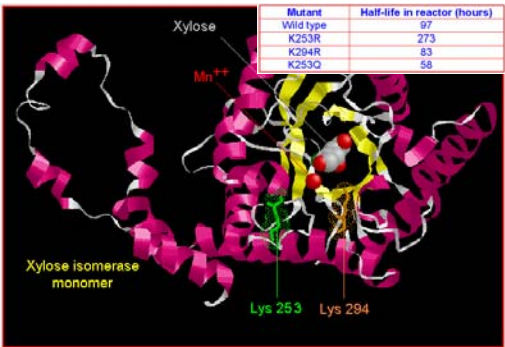
Protein Engineering



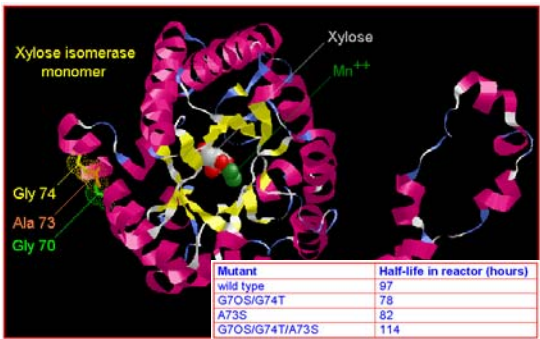
Protein Engineering



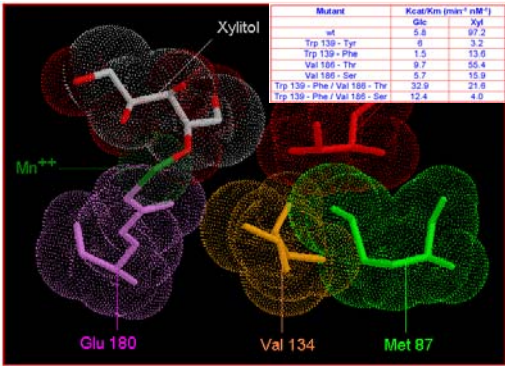
Protein Engineering



Protein Engineering



Protein Engineering



Protein Engineering

