Scientific Models

<table>
<thead>
<tr>
<th>Physical models</th>
<th>Mathematical models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanistic models</td>
<td></td>
</tr>
<tr>
<td>Stochastic models</td>
<td></td>
</tr>
</tbody>
</table>

**Mechanism**
- Predictive power
- Elegance
- Consistency

**Black box**
- Predictive power

---

**BINF 630: Bioinformatics Methods**

**Iosif Vaisman**

Email: ivaisman@gmu.edu

---

**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

Perceptron

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

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

<table>
<thead>
<tr>
<th>Start</th>
<th>Initialization</th>
<th>Valuation</th>
<th>Solution END</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>REPRODUCTION</td>
<td>CROSS-OVER</td>
<td>MUTATION</td>
</tr>
</tbody>
</table>

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

crossover point

Mutation

GA simulation of folding

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

Genetic Algorithms Applications

Parents
10 00 01 10 10 00 00 11 10 00 01 11

Children
10 00 10 11 10 00 01 10

Artificial Intelligence in Biosciences

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

Grammars and Language

*grammar* 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
Primitives 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


Hidden Markov Models

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

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 Models

Adapted from Anders Krogh, 1998
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.

A hidden Markov model in structural analysis

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

Comparison of AI methods

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>GML</th>
<th>CART</th>
<th>ANN</th>
<th>EA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Requirements</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Accurate model</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Inferability</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Reintegration into predictions</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Infeasibility in non-linear problems</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Transparency of the modeling process</td>
<td>High</td>
<td>Moderate</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Ability to model non-linear relationships</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Non-linearity and inability to integrate predictions</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

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)

Applications of ML methods

Applications of ML methods

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

Mapping in topological space