Protein Structure Analysis

Iosif Vaisman

2004

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

Perceptron

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

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

Neural Networks

Hierarchical neural network

Training set:
- 681 chains
- 158,428 residues
- 30% helix
- 22% strand
- 48% coil

First level network

Second level network

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

crossover point

Mutation

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

**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 Syntax

**Alphabet**

- Primitive elements
- Letters, phonemes

**Vocabulary**

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

**Grammar**

- Legal composition of vocabulary
- Rules, operators

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 (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


Hidden Markov Models

| A | T | C | T | A | G |

Observed frequencies

<table>
<thead>
<tr>
<th>A</th>
<th>0.7</th>
<th>A</th>
<th>0.1</th>
<th>C</th>
<th>0.8</th>
<th>A</th>
<th>0.4</th>
<th>A</th>
<th>0.8</th>
<th>C</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.3</td>
<td>T</td>
<td>0.9</td>
<td>G</td>
<td>0.2</td>
<td>T</td>
<td>0.6</td>
<td>T</td>
<td>0.2</td>
<td>G</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Probablistic 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

---

Adopted from S. Salzberg, 1997

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.

Adapted 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