Iosif Vaisman

Email: ivaisman@gmu.edu

Example of a Genbank entry

LOCUS       VIBHALUXA    3141 bp DNA             BCT       15-FEB-1996
DEFINITION  V.harveyi luciferase alpha and beta subunit (luxA and luxB) genes, complete cds.
ACCESSION   M10961 M13494
NID         g155174
KEYWORDS    luciferase.
SOURCE      Vibrio harveyi DNA.
ORGANISM  Vibrio harveyi
Eubacteria; Proteobacteria; gamma subdivision; Vibrionaceae; Vibrio.
REFERENCE   1  (bases 1 to 1838)
AUTHORS   Cohn,D.H., Mileham,A.J., Simon,M.I., Nealson,K.H., Rausch,S.K.,
Bonam,D. and Baldwin,T.O.
TITLE     Nucleotide sequence of the luxA gene of Vibrio harveyi and the complete amino acid sequence of the alpha subunit of bacterial luciferase
JOURNAL   J. Biol. Chem. 260 (10), 6139-6146 (1985)
MEDLINE   85207595
REFERENCE   2  (bases 1745 to 3141)
AUTHORS   Johnston,T.C., Thompson,R.B. and Baldwin,T.O.
TITLE     Nucleotide sequence of the luxB gene of Vibrio harveyi and the complete amino acid sequence of the beta subunit of bacterial luciferase
JOURNAL   J. Biol. Chem. 261 (11), 4805-4811 (1986)
MEDLINE   86168191

Example of a Genbank entry

FEATURES             Location/Qualifiers
gene            707..1774 /gene="luxA"
CDS             707..1774 /gene="luxA" /codon_start=1 /product="luciferase alpha subunit" /db_xref="PID:g155175" /transl_table=11
FEATURES             Location/Qualifiers
BASE COUNT      883 a    665 c    741 g    852 t
ORIGIN      1 bp upstream of EcoRI site.
1 gaattcacca tgacgacggg caaaaatagt ttgtgcactg tttatcactg gctgcagacc
61 aagggcacac aaaacattgg cttgattgcg gcaagtctct cagctcgtgt cgcctatgaa
121 gttatctctg atctggagct gtcttttctg attactgcgg ttggtgtggt gaacttgcgt
181 gacacactag aaaaagcgct tggttttgat tacctcagtt tgcctatcga tgagctacca
....

Database Organization

Database Management System (DBMS)

Four major components of DBMS:
Data * Hardware * Software * Users

Database
an a collection of related structured information about entities
file
a collection of records
record
a set of fields
field
a single characteristic of an entity
character
a symbol used in data field
Data Model
- A named logical unit (record type, data item)
- Relationships among logical units

Relationships among logical units
- one to one
- one to many
- many to one

DNA vs. Protein searches

DNA sequence

Protein sequence

DNA database

Protein database

DNA sequence - DNA database
- larger databases
- more random hits
- simpler scoring functions
- missing hits (similar proteins encoded by different DNAs)

Database administration
- Redundancy eliminated
- Inconsistency avoided
- Data shared
- Standards enforced
- Security applied
- Integrity maintained
- Requirements balanced

Data Warehouse

Operational data

Data fusion

Data cleansing

Metadata

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

Estimation
• Rules that explain how to estimate a value given characteristics

Prediction
• Rules that explain how to predict a future value or classification, given 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.

Affinity Grouping
• Grouping by relations (not by characteristics)
Scientific Models

Mechanistic models
- Mechanism
  - Predictive power
  - Elegance
  - Consistency

Stochastic models
- Black box
  - Predictive power

Physical models -- Mathematical models

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

- 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
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

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
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


Hidden Markov Models

A G A T C T A G

Observed frequencies

A 0.7 T 0.3 C 0.8 G 0.2 A 0.4 T 0.6 C 0.3

Probabilistic model - true state is unknown

Hidden Markov Models

ATGAC ATTAC ACGAC ACTAC

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

ACA---ATG
TCAGCTACTC
ACAC--AGC
AGA---ATC
ACCG--ATC

probabilities

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

Adopted from Anders Krogh, 1998

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

Adopted from S. Salzberg, 1997
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.

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