YORK UNIVERSITÉ UNIVERSITY
redefine THE POSSIBLE.

Bayesian Network Modeling for Evolutionary Genetic Structures

Lisa Jing Yan
lisayan@cse.yorku.ca

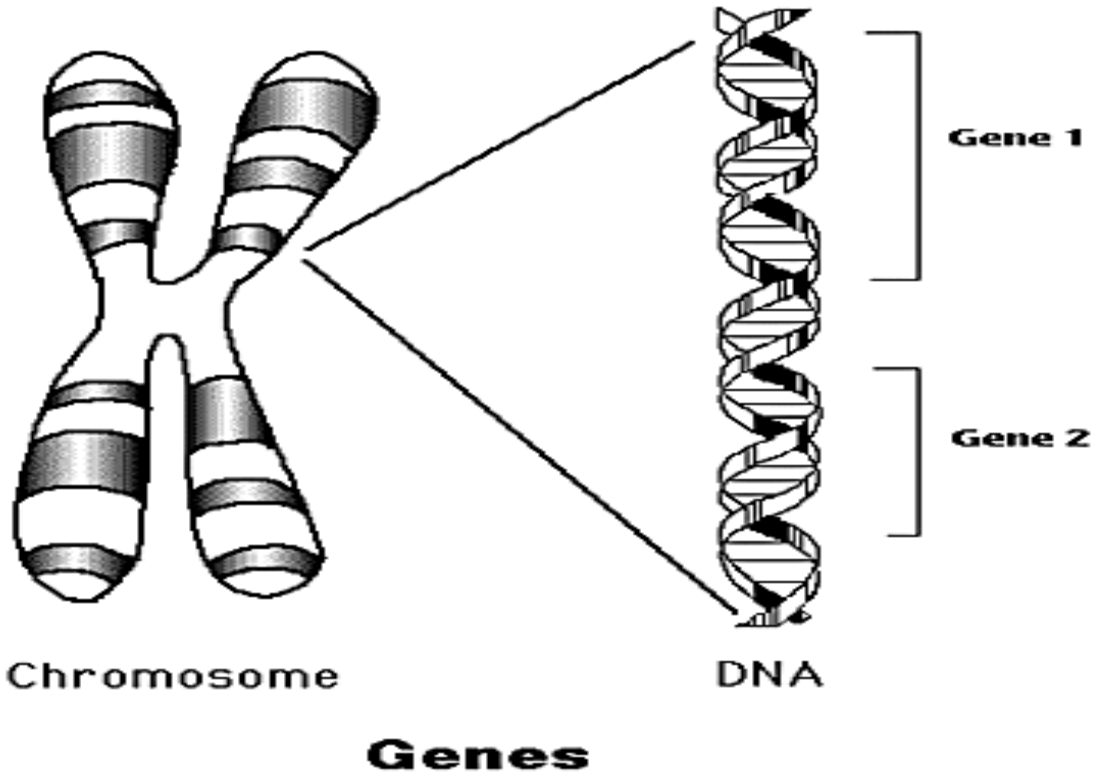
Table of Contents

Part 1: Research Problem: Gene Selection

Gene Selection

- Evolution forces genes to naturally change and adapt so organism can survive; called ‘natural selection’ in evolutionary biology
- Takes millions of generations to measure and assess; change is gradual and genetically adapts organism to its environment

Picture of a real gene



Gene Selection

- Artificial genes (AGenes) can be created and quickly ‘evolved’ in an Artificial Life (AL) environment
- Mimics evolution and determines rapidly which factors promote survival ‘fitness’

- How??

Genes are codes

- Genetic information is a code representing characteristics of an organism or entity

- Can be complex, with secret or hidden relationships between genes and gene combinations

Research Problem

- Bayesian Network (BN)

  –------ (research bridge)-------

  Genetic Algorithm (GA)

- Our Work:

  Evolutionary process can be analyzed using BN methods

Significance of Research

- Reveal importance to use BN to analyze incomplete or complex data
- Propose new utility and flexibility of GA-based AL, to provide dataset we need, not just optimized algorithm
- Suggest important new applications for business world, and for biology

Part 2: Research Methods

Research Design

We propose to combine AL and BN:

- Artificial Life means create a hypothetical simulation of the real world and behavior of real organisms, which could provide interesting data for BN analysis
- BN graphically models data about the ‘best’ artificial genes which emerge from evolutionary simulation. This allows prediction about which genes are optimal to achieve certain desired goals

Research Steps

Step 1: BN as an analytical tool (E-algorithm)

Step 2: Develop GA based AL Model-ALGAE (Artificial Life Genetic Algorithm Expression): provide AGene Database

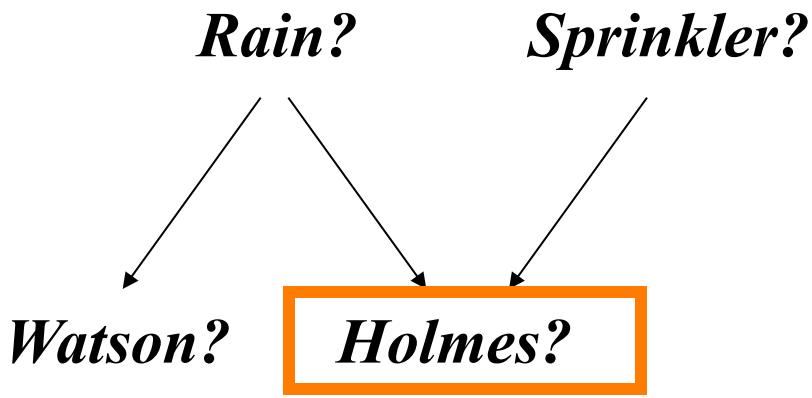
Step 3: Develop BANANA (BAyesian Networks ANAlysis):

Seek for hidden relationships among AGenes

Step 1: BN Learning

Why BN?

- Why use BN?
- Causal, bottom-up description of ‘wet grass’ model

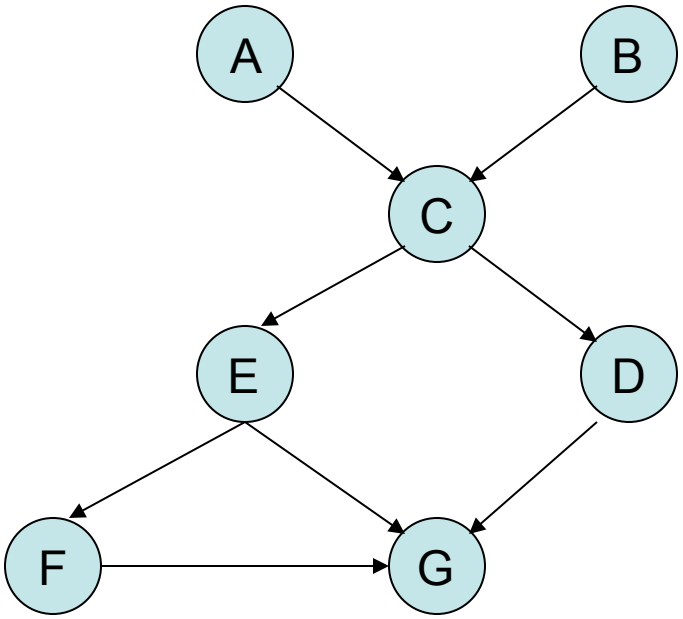


BN Principles

- BN describes dependency
- BN reasoning connects events using probability and causation
- Probability is conditional
- Variables have degrees of dependence or independence

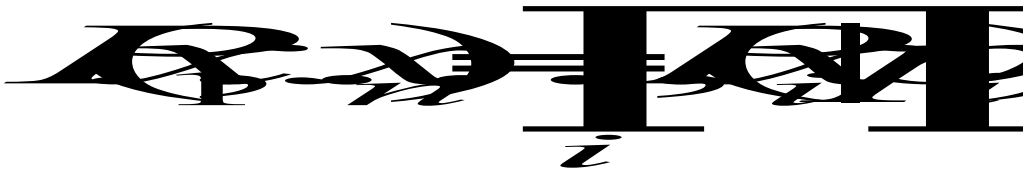
BN Structure

- Variables (nodes) and edges
- Directed Acyclic Graph (DAG)



Joint Probability Distribution (JPD)

In Bayesian networks, for each variable, the conditional probabilities are the set of parents which make them independent of all other parents. After giving this specification, the joint probability distribution can be calculated by the product



How to Model BN?

- BN learning has two main methods:

a) Constraint-based method:

Perform tests of conditional independence (CI) on the data, and search for a network that is consistent with the observed dependencies and independencies .

b) Score-based method:

Define a score that evaluates how well the dependencies in a structure match the data, and search for a structure that maximizes the score.

Pro. & Con. (Constraint Based)

- Con.

  This approach is problematic since conditional independence relations are difficult to achieve with certainty.
- Pro.

  However, Constraint based methods are more intuitive. They follow the definition of a BN more closely, also separate the notion of the independence from the structure construction.

Score Based

- Define a score that evaluates how well the dependencies in a structure match the data, and search for a structure that maximizes the score.
  - the log-likelihood function
  - the *MDL score*.
  - *Bayesian score* (BDE Score)
- They operate on the same principle:

  a scoring function is defined for each network structure, representing how well it fits the data.

Pro. & Con. (Score Based)

- Con.

  Searching in a combinatorial space:

  Not clear how one can find the best-scoring network even with a scoring function. In general, the problem of finding the best-scoring network structure is NP-hard.
- Pro.

  Less sensitive to errors in individual tests:

  Compromises can be made between the extent to which variables are dependent in the data and the cost of adding the edge.

Learning Limitations

Limitations of both methods:

(cf. improvements)

- Too many tests required, thus costly and less efficient
- Complex BN increases structure thus increases time-cost, NP hard

E-algorithm Design

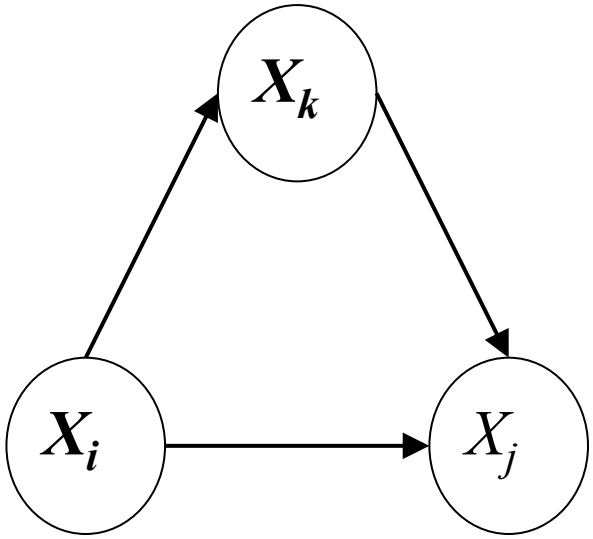
E-algorithm combines CI test and MDL metric search:

- Uses CI initially + Improvement 1
- Combine MDL score and B&B searching + Improvement 2

Improvements are as follows:

E-algorithm Improvements

1. order-0 & order-1 independence tests;



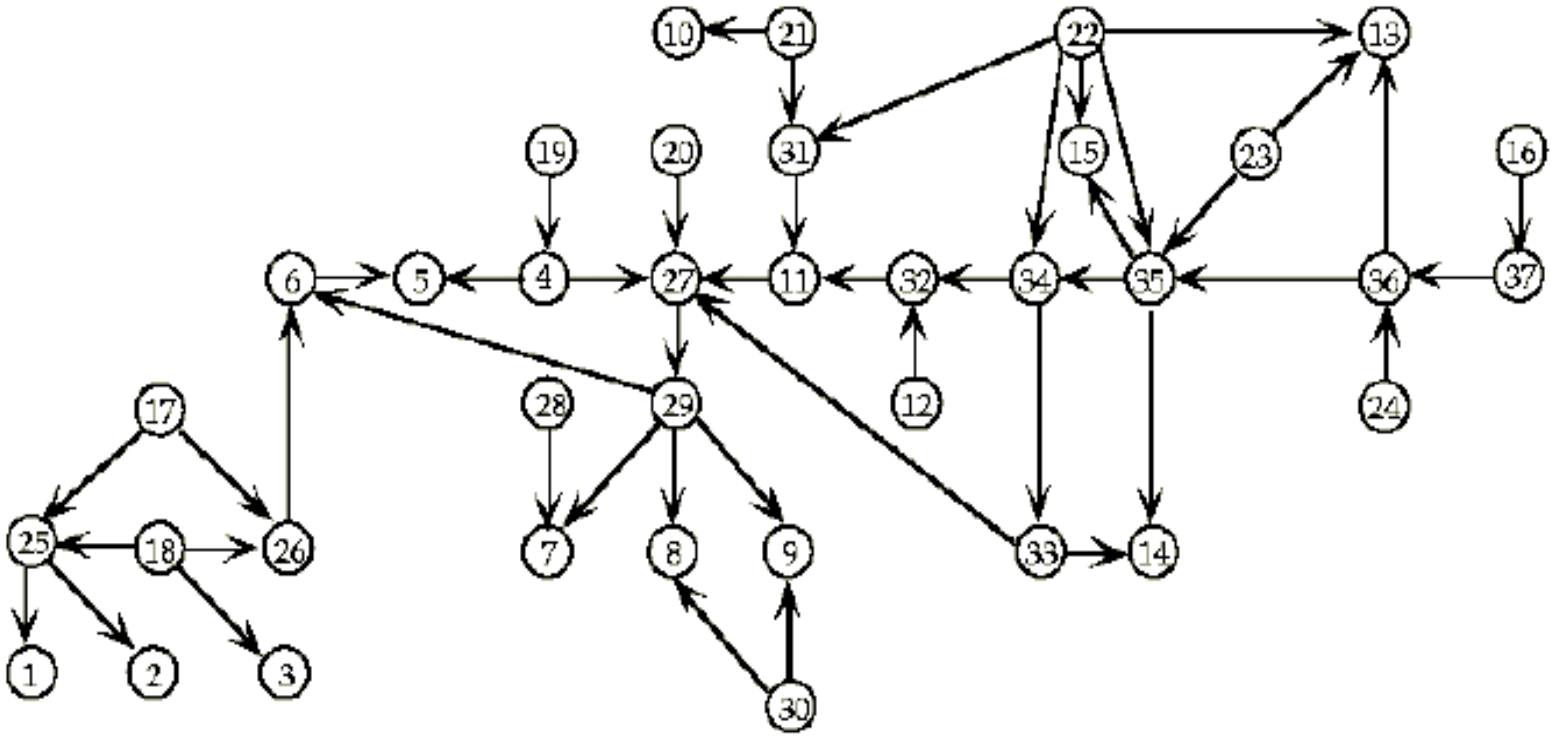
2. sort candidate parent nodes order, as heuristic information (cf. limitations)

Experiment Results

- Benchmark ALARM (A Logical Alarm Reduction Mechanism)
- A medical diagnostic system for patient monitoring:
  - 8 diagnoses
  - 16 findings
  - 13 intermediate factors

ALARM

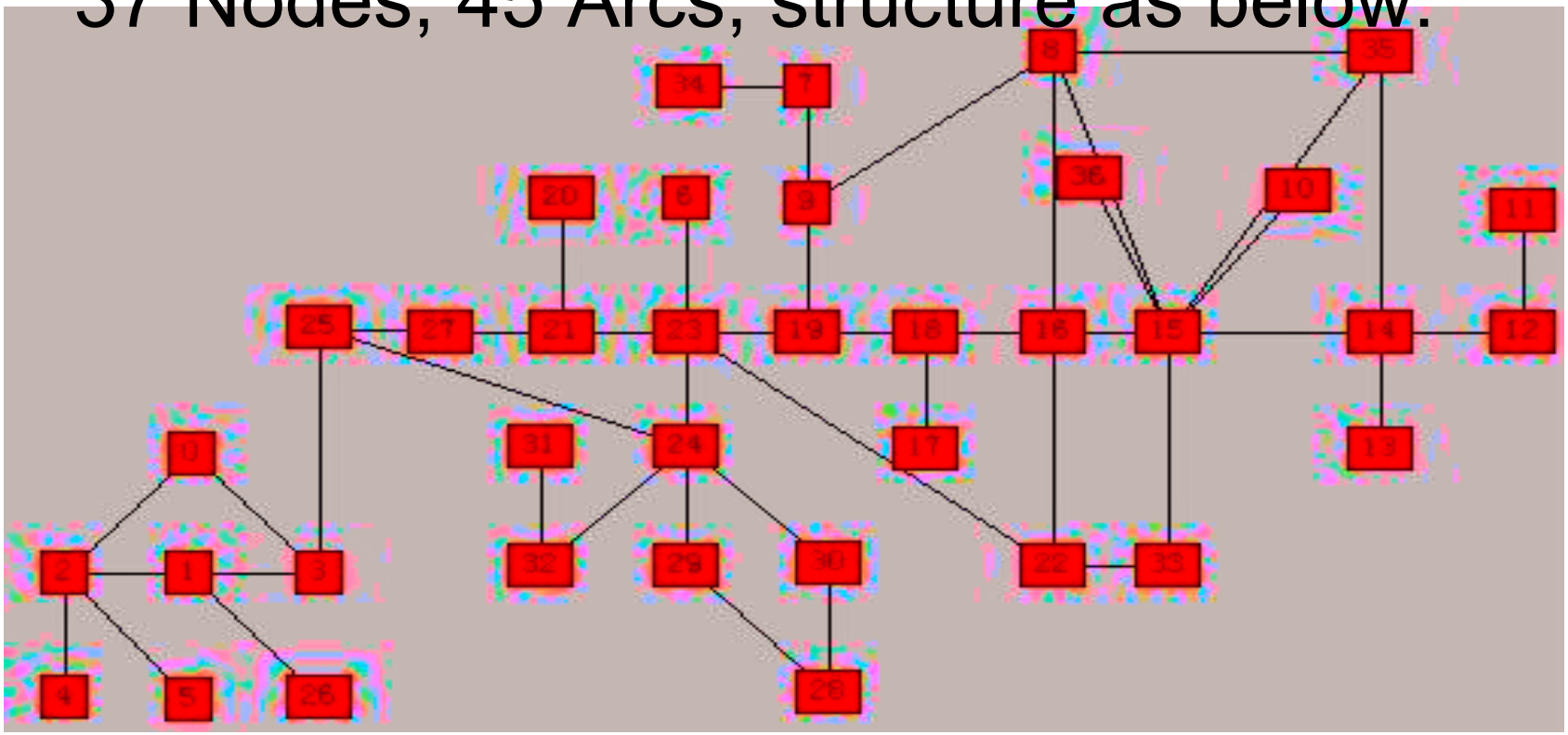
37 Nodes, 46 arcs



Experiment Results

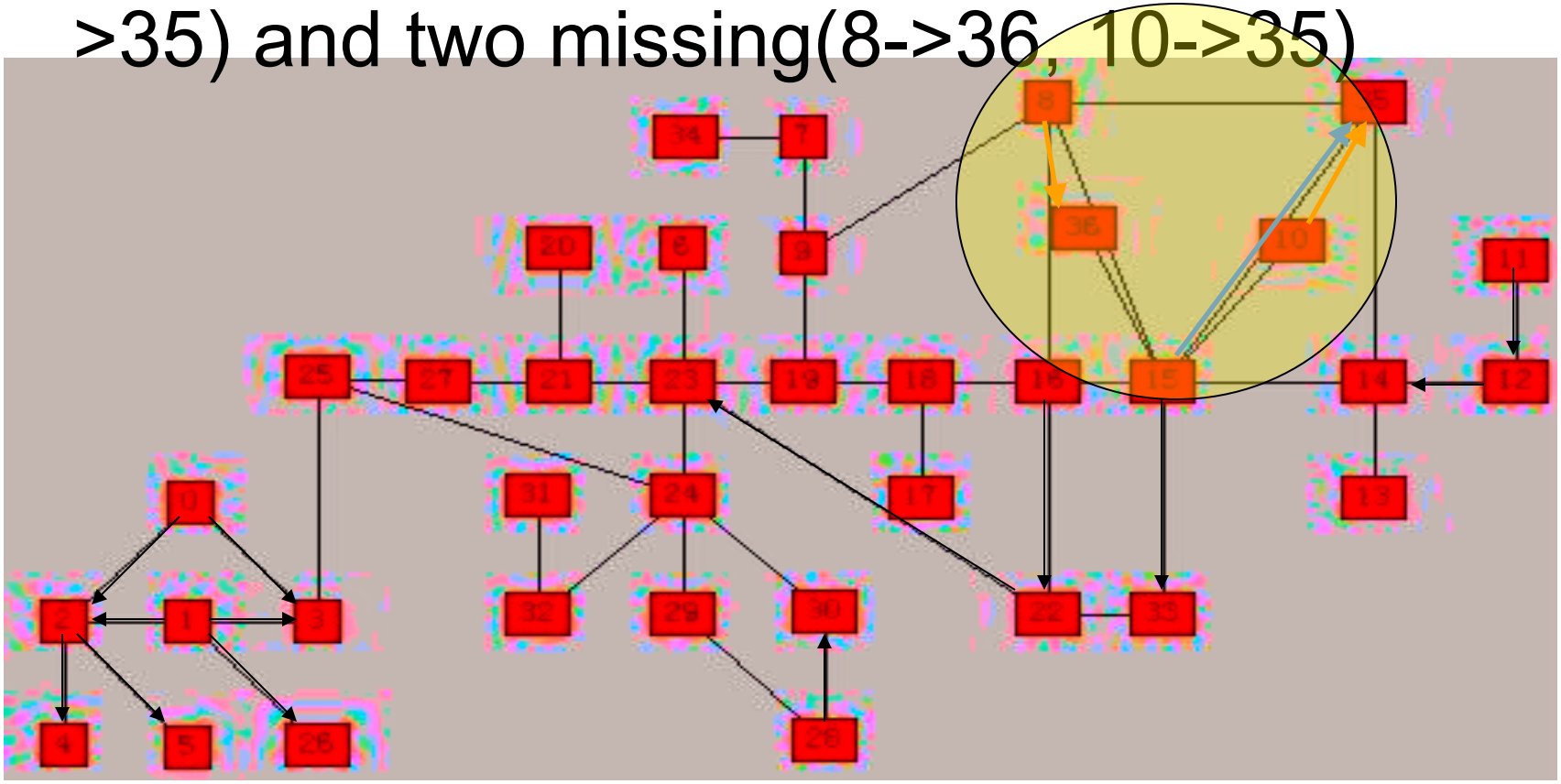
- Our Results:

  37 Nodes, 45 Arcs, structure as below:



Result Analysis

Our structure has one redundancy(15->35) and two missing(8->36, 10->35)



Result Analysis

- Possible reasons for the three problematic arcs are reasonable; since the possible attribute combination of all 37 variables, we use only 10000 record database rather than this enormous size and complexity: it is relatively small.
- Our E-algorithm is feasible.

Step 2: ALGAE

ALGAE 1

- ALGAE Goal:

  AGene Data Collection (used by BN for analysis)
- Design:

  Develop ALGAE to mimic natural selection and create a dataset related to the selection of the best/fittest gene resulting from artificial evolution

ALGAE 2

- Experiment: based on using GA to develop AL competitive environment

Why choose Artificial Life?

Why AL?

- AL concept is based on Evolutionary Biology and AI.
- Genes and chromosomes artificially emulate real organisms and living systems.
- Goal is to survive through genetic fitness.
- Can perform testing to speed up evolution time, and can control environment and create rules which control species in it.

Why GA?

- Based on evolution and Darwinism's natural selection (Gene Selection) and applies it to AL genes.
- GA is adaptive search algorithm that improves and optimizes outcomes for each generation by building on previous, sub-optimal solutions.
- Reaches best solution by learning as it goes along.

ALGAE Design - Frame

A Simple Ecology System:

- Certain resources (plant) exist, distributed in a two dimensional grid.
- Two agents in this virtual world: Species 1 and Species 2.
- Compete for resources to survive.
- Certain behaviors, as: eat, mate, fight.
- Ages increase until maximum, then natural death.
- Barriers exist to constrict their movement.

ALGAE Design - Factors

In ALGAE, we consider several aspects, such as:

- Living environment (or lifespace)
- Population
- Food resources
- Barriers
- Competition
- Behavior patterns & preferences
- Physical status

ALGAE Design - AChromosome

- Coding of artificial chromosomes using standard behaviors such as motion, contact with other individuals (either species) such as attack, mating, defense
- Cf. the following table, showing 32-bit chromosome coded at start of run.

AChromosome Design

32-bit AChromosome *Gi* descriptor:

[*SP* , *SL* , *VF* , *TM* , *CM* , *LM* , *CA* , *CR* , *SA* , *DA* , *LA* , *EF*]

Gene	Gene Name	Bit Site	Gene	Gene Name	Bit Site
SP	SPecies type	0	CA	Action Characteristic	13-15
SL	Life Span	1-4	CR	Capricious Rate	16-18
VF	Vision Field	5-6	SA	Attack Speed	19-21
TM	Transition Movement	7-8	DA	Defend Ability	22-24
CM	Motion Characteristic	9-11	LA	Attack Loss	25-27
LM	Motion Loss	12	EF	Food Efficiency	28-31

AGenes - 1

[*SP , SL , VF , TM , CM , LM , CA , CR , SA , DA , LA , EF*]

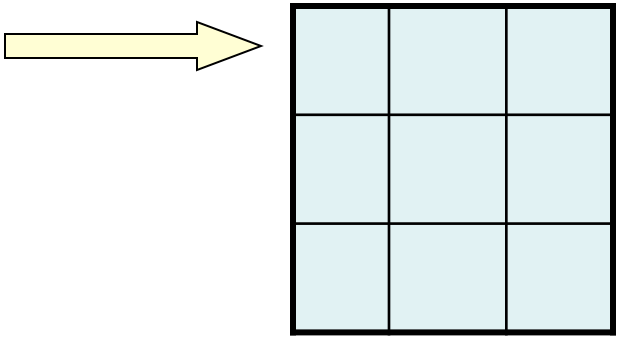
- *SP* (*Species type):* 0-species 1 / 1-species 2.
- *SL* (*Life span): Age =SL_MIN+SL.*
- *VF (Vision field*):

  0:*Area*= 3×3;

  1:*Area*=5×5;

  2:*Area*=7×7;

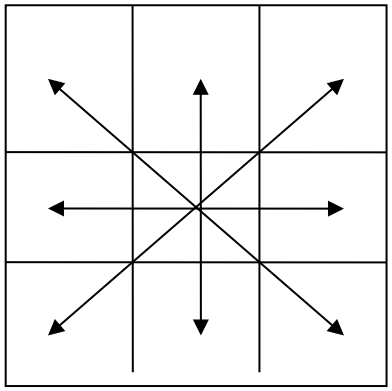
  3:*Area*=9×9;



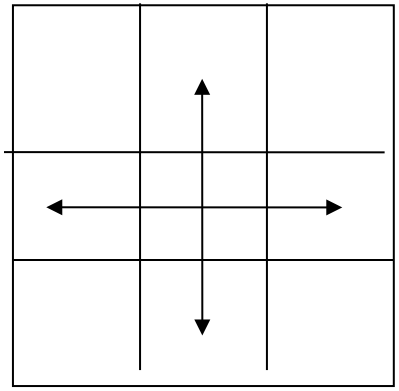
AGenes - 2

[*SP , SL , VF , TM , CM , LM , CA , CR , SA , DA , LA , EF*]

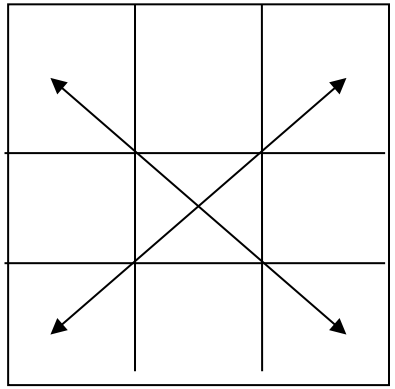
- *TM* (*Transition Movement*):



Pattern1: Move randomly



Pattern2: Move across



Pattern3: Move diagonally

AGenes - 3

[*SP , SL , VF , TM , CM , LM , CA , CR , SA , DA , LA , EF*]

- *CM* (*Motion Characteristic*):

  1st: homogeneous biological motion (0: Neg./1: Pos.);

  2nd: heterogeneous biological motion (0: Neg./1: Pos.);

  3rd: food motion (0: Neg./1: Pos.).

- *LM (Motion Loss):*

  Energy Loss＝LM＋1 .

- *CA (Action Characteristic*):
  - *CA* simulates the biological drive for three different behaviors: 1. Attack 2. Hunting 3.Copulation.
  - It also describes behavioral preferences and their sequence, as follows:

AGenes - 4

[*SP , SL , VF , TM , CM , LM , CA , CR , SA , DA , LA , EF*]

1. Attack 2. Hunting 3.Copulation.
   - *if CA* =0/1/2; sequence:(1; 2; 3)
   - *if CA* =3; sequence:(1; 3; 2)
   - *if CA* =4; sequence:(2; 1; 3)
   - *if CA* =5; sequence:(3; 1; 2)
   - *if CA* =6; sequence:(2; 3; 1)
   - *if CA* =7; sequence:(3; 2; 1)

- *CR* (*Capricious Rate*):
  - *CR* is the probability that individual does not comply with *CM, s*pecies behavior (1;2;3).
  - However, individual may switch to another choice since fickleness is an implicit factor in any decision-making.

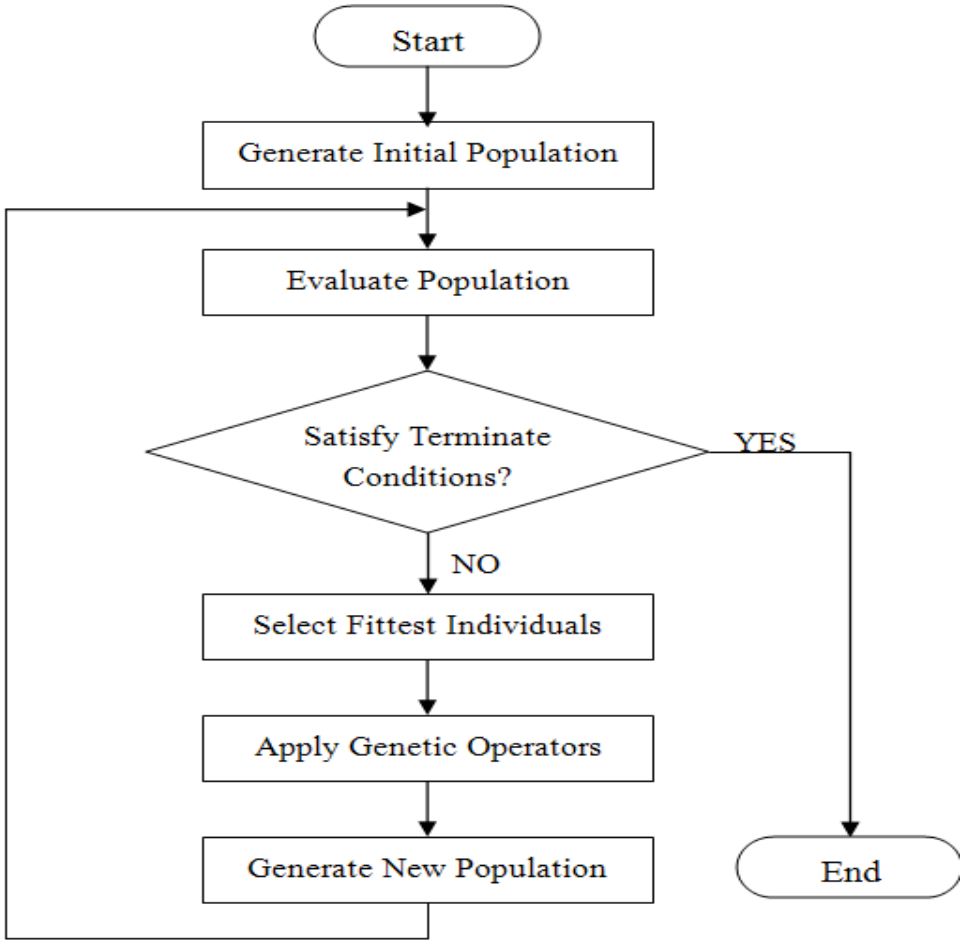
AGenes - 5

[*SP , SL , VF , TM , CM , LM , CA , CR , SA , DA , LA , EF*]

- *SA*
- *DA*
- *LA*
- *EF*

ALL Related to Fighting

GA Diagram


Start
Generate Initial Population
Evaluate Population
Satisfy Terminate Conditions?
YES
NO
Select Fittest Individuals
Apply Genetic Operators
Generate New Population
End

ALGAE Run Process - 1

- Step 1: Initialize AWorld environment:
  - set up barriers and vegetative food supply;
- Step 2: Initialize a population of AChromosomes randomly:
  - each individual *i* $Energy_i$: (70,100), Age_i :(0,*SL_MIN)*;
- Step 3: Evolutionary process start :
  - population of AChromosomes are ready to evolve;
- Step 4: According to individuals' AGene and status, certain activity is to command either "*Move*" or "*Act*":
  - *Move*: change to another spot;
  - *Act*: any one of *attack, eating, and mating*

  Within individual' vision field, no attractive thing or food exists, then individual can only choose to *Move*;

ALGAE Run Process - 2

- Step 5: Increase each individual *Age* 1;
  - if anyone's *Lifespan* surpasses Max., then eliminate it from population, also use cadaver as animal food;
- Step 6: Increase vegetative food *Fresh Level* 1;
  - eliminate the expired food supplies which have surpassed its *Time Limit*;
- Step 7: Increase generation number 1;
  - if all species extinct or over Max. given *Generation Number*, then go to step 3, Loop.

ALGAE Business Model App.

- ALGAE works also for two corporations who exist in competitive market conditions
- Code into the corporate entity “genes” for certain marketing abilities and functions
- Program imposes accelerated evolution on each business, mimics environment of real world conditions

Business App.

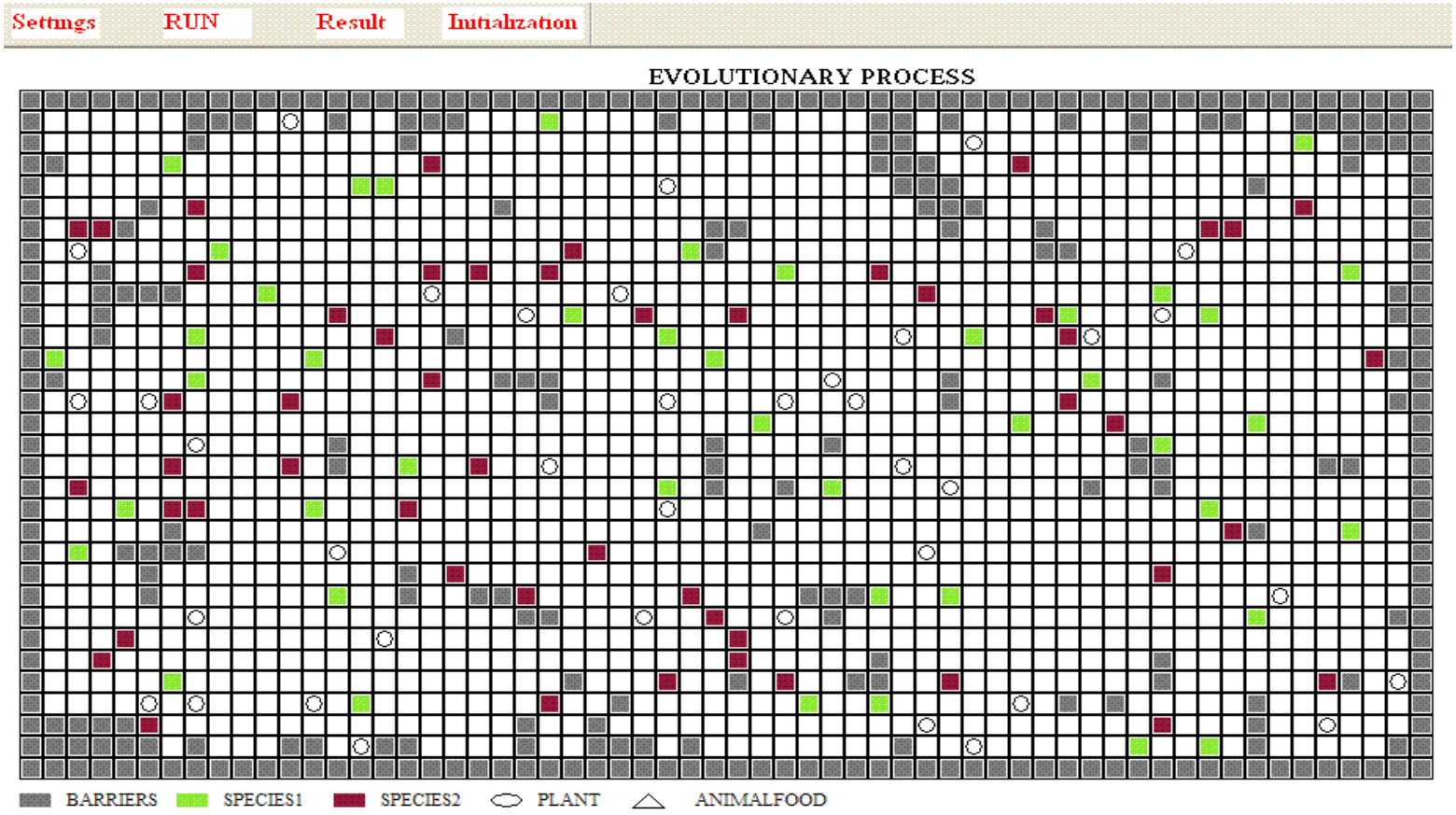
- Results show ALGAE predicts best qualities of successful business
- Allows business to plan ideal strategy for profitable operation
- Strategy is based on understanding the precise factors which contribute to survival and success of the enterprise

Experimental run

- Note that species have *choice* and that GA randomly assigns parameters for each generation and selects genes according to their fitness (ability to survive and adapt)

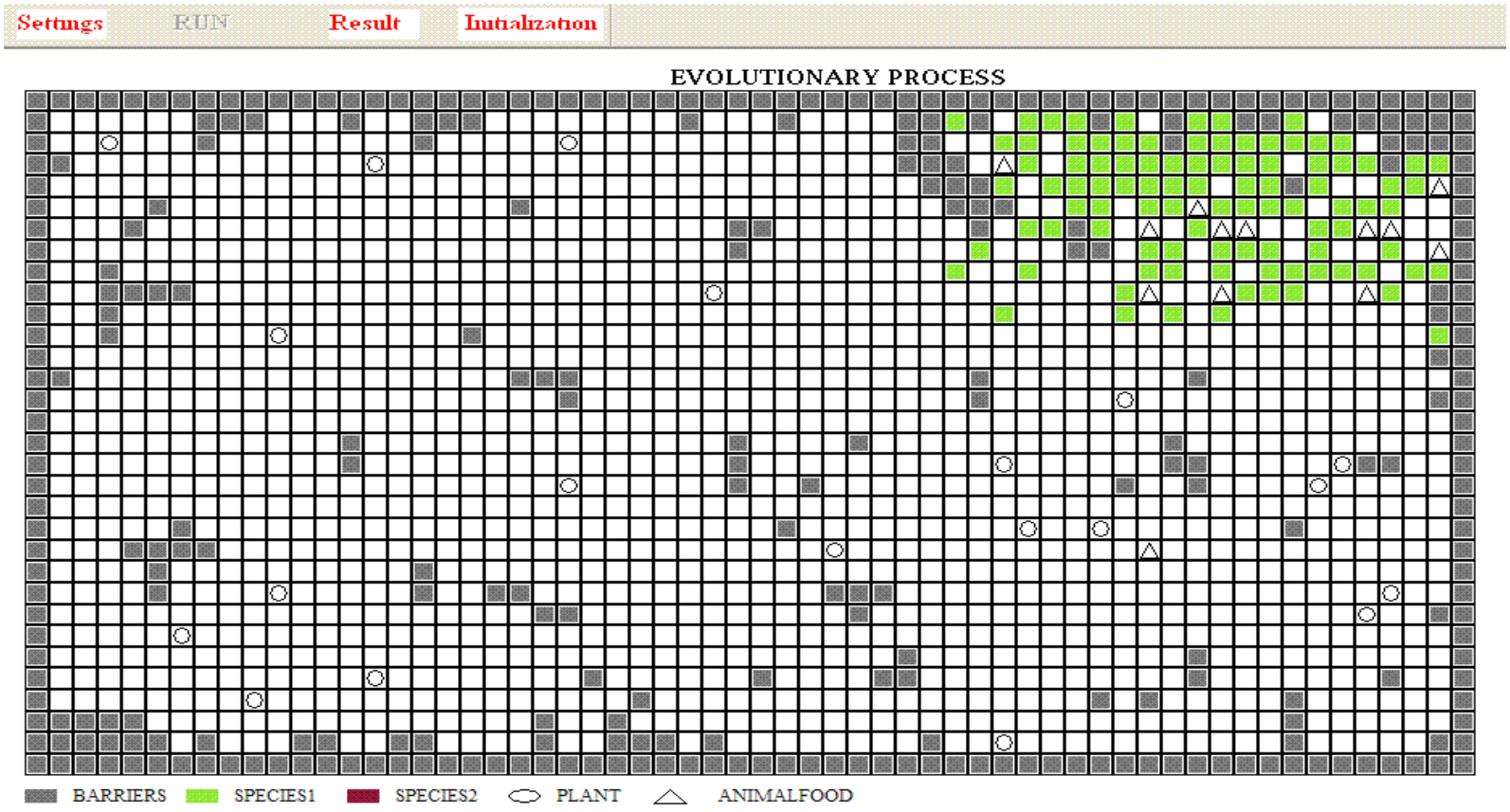
Results

- Initial State:



Results

- Final State:



Results

- Fittest Genes:

Gene	Gene Name	Fittest Genes	Gene	Gene Name	Fittest Genes
SP	SPecies type	1	CA	Action Characteristic	111
SL	Life Span	0111	CR	Capricious Rate	000
VF	Vision Field	11	SA	Attack Speed	111
TM	Transition Movement	01	DA	Defend Ability	111
CM	Motion Characteristic	101	LA	Attack Loss	000
LM	Motion Loss	1	EF	Food Efficiency	0100

Results

Fittest Genes Explanation (a) :

- Species 1 has the bigger chance of survival.
- Long Life span.
- Wide view field.
- Flexible with movement.
- Dislikes fight, but prefers homogeneous entity and food.
- Low energy consumption when move.

Gene	Gene Name	Fittest Genes
SP	SPecies type	1
SL	Life Span	0111
VF	Vision Field	11
TM	Transition Movement	01
CM	Motion Characteristic	101
LM	Motion Loss	1

Results

Fittest Genes Explanation (b):

Behavior preference : (3.Copulation;2.Hunting; 1.Attack); It implies the fittest way to maintain energy.

- Stable with decision.
- Fast attack speed.
- Strong Defense capability.
- Low energy consumption when fight.
- High Food absorption efficiency.

Gene	Gene Name	Fittest Genes
CA	Action Characteristic	111
CR	Capricious Rate	000
SA	Attack Speed	111
DA	Defend Ability	111
LA	Attack Loss	000
EF	Food Efficiency	0100

Result Discussion

- Ten Trials Dataset Logs:

Trial No.	Survivor SP.	Dataset Size	Fittest Genes
1	1	11420	1101001100000101011100001100 1100
2	1	11251	1001000001010111000001100000000 1
3	1	11558	1111101011100101011111001101100
4	0	11248	0110010110111000000110101110011 1
5	1	2977	1111111110110101010001100010000 0
6	0	5281	0100000001110001001011111011010
7	1	10679	1011000100001110010111100010 1101
8	1	4910	1001100001101111001101010101011 1
9	1	7311	1111010110000111001000110000001 0
10	1	11086	1101111100000110001110110111001 1

Results Discussions

- This is competition model therefore energy levels, especially before and after attack, are important. Stronger members survive; others die. Genes carry the information about which characteristics give strength and are useful. ALGAE builds stronger genes over *n* generations.
- Table above shows variation in the composition of each best gene. Each AChromosome has entirely different and unique attributes.

Why?

Result Discussions

- Because ALGAE randomizes the chromosomes for each run, as well as certain environmental factors such as population distribution in relation to resources.

But!!

- Under the same rule of evolution, what can we learn from the total Gene Selection Process?

NEXT, use BN to analyze the AGene datasets.

Step 3: BANANA

BANANA

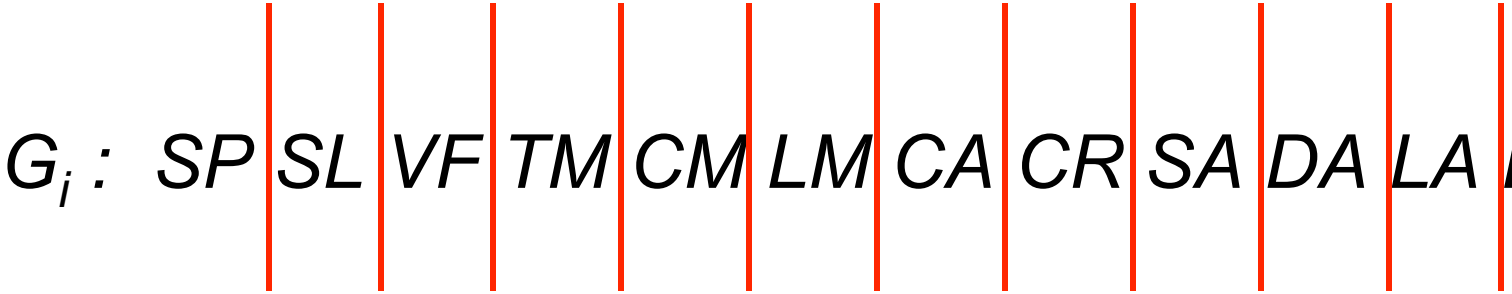
Goals:

- Answers questions about hidden relationship of characteristics coded into artificial genes.
- Describe AGenes in graphical model, in order to account for how to survive during Gene Selection process.

Data Processing 1

To facilitate the processing by BANANA, it required some manipulation:

1. AChromosomes are divided into 12 segments, by bit size, as shown below:



2. Convert Binary coding of the 12 segments into real values (1-4).

Data Processing 2

A conversion principle follows:

- if Seg_i = 00/01; then $Value_i$ = 1;
- if Seg_i = 10/11; then $Value_i$ = 2;
- if Seg_i = 100/101/110/111; then $Value_i$ = 3;
- Otherwise, $Value_i$ = 4.

We use MS ACCESS database to process the real genotype binary values into integers for the BN analysis.

Testing E-algorithm

BANANA program is based on E-algorithm which is tested and verified:

- Test E-algorithm against ALARM:

  Produces acceptable BN for this data, Confirm its usefulness

- Test E-algorithm against Chest Clinic benchmark. Acceptable result.

AGene Datasets

- Produced ten datasets for ten trials to give valid empirical data for analysis.
- Datasets contains 12 variables (after segment); they are:

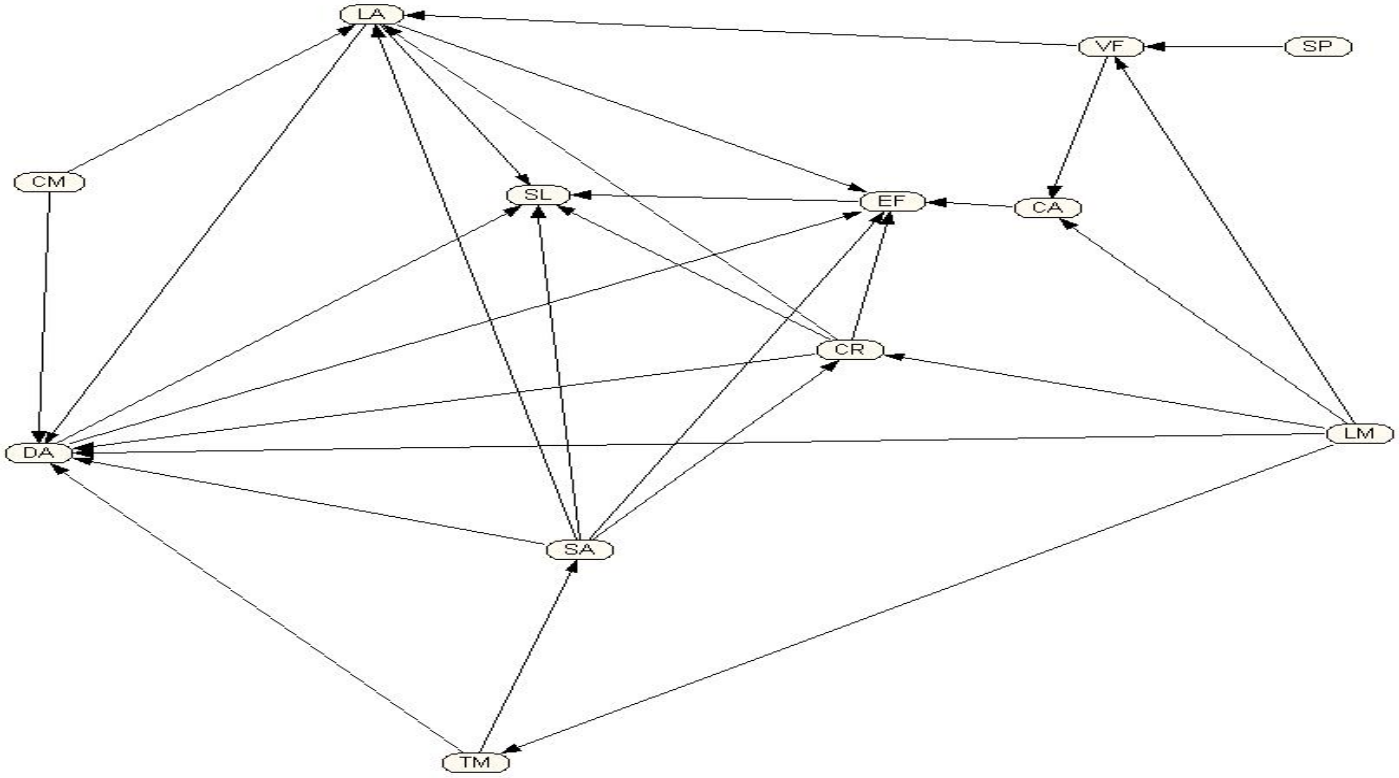
  SP, SL, VF, TM, CM, LM, CA, CR, SA, DA, LA, EF.

BANANA Result 1

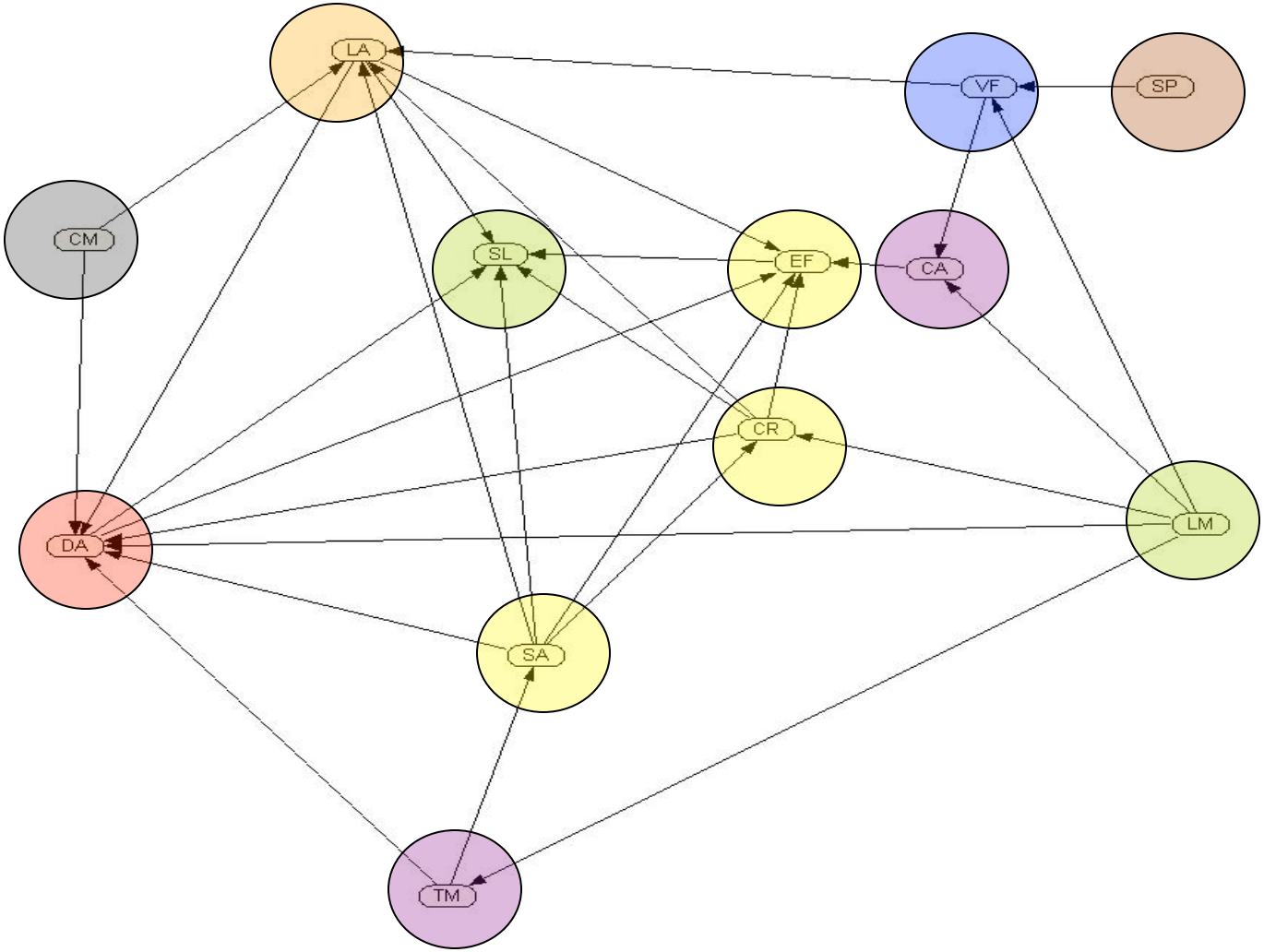
- BN uses data logs for 10 trials survivor genes to establish a graphical structure, to reveal dependencies or hidden attributes of genes in relation to each other.
- Graphical Model see as follows:

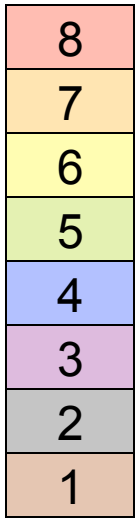
BANANA Result 2

- 24 arcs connecting 12 variables clearly.



BANANA Result Discussion 1





BANANA Result Discussion 2

- The relationships and dependencies indicate that Defense Ability (*DA)* is the key gene in determining survival in the process.
- BN shows the hidden ‘rules’ of survival embedded in the dataset from ALGAE.
- The rule is that only certain gene combinations will allow species to survive.

BANANA Result Discussion 3

- Each gene has a different level of importance in survival and evolutionary process, as indicated by the different colors.
- It is the key to why even generations with weaker genes can somehow adapt to living conditions and live long enough (*DA, SL*) to have offspring to create the next generation.

BANANA Result Discussion 4

- Ten Trials Dataset Logs:

Trial No.	Survivor SP.	Dataset Size	Fittest Genes
1	1	11420	11010011000001010111000011001100
2	1	11251	10010000010101110000011000000001
3	1	11558	11111010111001010111111001101100
4	0	11248	01100101101110000001101011100111
5	1	2977	11111111101101010100011000100000
6	0	5281	01000000011100010010111111011010
7	1	10679	10110001000011100101111000101101
8	1	4910	10011000011011110011010101010111
9	1	7311	11110101100001110010001100000010
10	1	11086	11011111000001100011101101110011

BANANA Result Discussion 5

- Table shows the gene composition but *not the reason for its success.* The data log merely reports the fact; the BN tells the story of *why* this species could continue to live and thrive.

Overall Results

Overall Results

- BN shows best gene structure quite well.
- Allows analysis of variables so that characteristics which are most adaptive for survival are revealed.
- Shows relationship of these characteristics so that their combined effect produces an ideal ‘best’ gene over each generation.

Part 3: Conclusions

Main Research Focus

- Evolution has important lessons for ‘survival of the fittest’ (Darwin).
- Evolution not analyzed efficiently.
- BN will efficiently analyze evolutionary process information.

Contributions

- Bayesian Networks Learning E-Algorithm
- Artificial Model (Genetic Algorithm Based):

  ALGAE (Artificial Life Genetic Algorithm Expression)
- BNs Application in Gene Selection:
  BANANA (BAyesian Networks ANAlysis)

Conclusions

- Experiment in created AL environment is useful in producing a unique reliable dataset, unlike any benchmarks available
- GA could be modified to suit experimental design for either real-world business (or similar) application, or for purely hypothetical experimental purposes
- BANANA worked well with ALGAE dataset and produced acceptable results

References

References

- L. J. Yan, N. Cercone. Bayesian network modeling for evolutionary genetic structures. *Comput. Math. Appl.* 59, 8 (April 2010), 2541-2551. 2010.
- D. Heckerman, D. Gieger, M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Microsoft Technical Report MSR-TR-94-09, 1994.*
- J. Pearl. Constraint-propagation approach to probabilistic reasoning. In L. M. Kanal & J. Lemmer(Eds.). *Uncertainty in Artificial Intelligence. Netherlands: Amsterdam, 1986.*
- L. Qiang, T.Y. Xiao, G.X. Qiao. An Improved Bayesian Networks Learning Algorithm. *J.Computer Research & Development,39(10), 1221-1226, 2002.*

References

- M.L.Wong, S.Y. Lee, K.S. Leung. A Hybrid Approach to Discover Bayesian Networks Learning from Databases Using Evolutionary Programming. *Proc.2002 IEEE Int'l Conf. on Data Mining,498-505, 2002.*
- J. Yan, S. Lv, N. Zhong. Artificial Life Modeling in Corporate Strategy. *Journal of Guangxi Normal University.* 2007.
- J. Ji, C. Liu, J. Yan, N. Zhong. Bayesian networks structure learning and its application to personalized recommendation in a B2C portal. *Proc. IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society Press, 2004, 179-184.*
- J. Yan. *Bayesian Network Structure Learning. Thesis. College of Computer Science, Beijing U. Technology, 2003.*

Thank You.