agent-based model of genotype editing

on static and dynamic evolutionary scenarios

luis m. rocha
Indiana university
school of informatics
1900 East Tenth Street, Bloomington IN 47406

rocha@indiana.edu
http://informatics.indiana.edu/rocha
agent-based model of genotype editing

team

chien-feng huang
StatSoft and National Tsing Hua University
USA/Taiwan

ana maguitman
Universidad Nacional del Sur
Bahía Blanca, ARGENTINA

jasleen kaur
Indiana university, Department of Informatics, USA

acknowledgements

shetaal narayanan, Indiana university, Department of Computer Science
vitorino ramos, Instituto Superior Tecnico, Portugal

rocha@indiana.edu
http://informatics.indiana.edu/rocha/editing
traditional view of gene function

genotype/phenotype mapping in artificial life

Genotype DNA

RNA

Translation (Code)

Phenotype Amino acid Chains

Development

Environmental Ramifications

Traditional Genetic Algorithm

Genotype Variations

Phenotype Selection

INFORMATION

luis rocha 2006

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY
ncRNA: a regulatory hidden layer in Eukaryots

- Evidence for non-protein coding RNA (ncRNA) in complex organisms (higher eukaryotes)
  - “ncRNA dominates the genomic output of the higher organisms and has been shown to control chromosome architecture, mRNA turnover and the developmental timing of protein expression, and may also regulate transcription and alternative splicing.”
  - A Hidden Layer of Non-protein-coding RNAs in Complex Organisms.

- Two types of genetic information
  - mRNA for proteins
  - ncRNAs for RNA products
    - concomitant networking and other functions
    - parallel trans-acting signals for gene regulation

- Three types of genes in eukaryotes
  - Encoding only proteins
  - Encoding only ncRNA
  - Encoding both

- Many types of ncRNA
  - tRNA, rRNA, SnoRNA, miRNA, siRNA, eRNA, etc.

Mattick, J. S. [2001]. EMBO Reports 2, 11, 986–991
Mattick, J. S. And V. Makunin [2005]. Human Molecular Genetics 14, 11, R121-R132
u-insertion (via gRNA)

```
Ser    Gly    Glu
AuGuuuCGuuGuuAGAuuuuuAuAuuuuuuuuuAuAu
MerPhe Arg Cys Arg Phe Leu Leu Phe Phe Leu Leu
```

```
Gln    Glu    Gly    Arg    Gly    Lys ... 
CAGGAGGGC CGUGGAuAAG
Gln    Glu    Gly    Arg    Gly    STOP
```

- RNA Editing: post-transcriptional alteration of genetic information
  - can be performed by ncRNA structures (and proteins).
- U-Insertion/deletion RNA Editing (mitochondria of kinetoplastid protozoa)
  - involve small guide RNAs (gRNA) complementary to the target mRNA
  - insertion and deletion of Uracil (U) residues, usually in coding regions of mRNA transcripts
    - e.g. creation of open reading frames)
A-to-I and C-to-U substitution in humans

- **Adenosine (A) to inosine (I) Substitution**
  - Inosine (I) is read as guanosine (G) in translation
  - Involve enzymes: adenosine deaminases acting on RNA (ADAR) / RNA Editase
    - Implicated in epilepsy, Parkinson's Disease, depression, etc.

- **C-to-U Substitution**
  - Humans have a single locus encoding the APOB gene.
  - Gene expressed in both liver and intestine cells
  - Liver
    - No edition
    - apolipoprotein B-100 is produced
    - essential for the transport of cholesterol and other lipids in the blood
  - Intestine
    - Editing changes a codon from the amino acid glutamine (Gln) to a STOP codon (UAA)
    - Via the enzyme cytidine deaminase: C to U.
    - Produces apolipoprotein B-48 protein
    - that aids in the absorption of dietary lipids from the contents of the intestine.

John W. Kimball
http://biology-pages.info/
RNA editing

U-insertion and A-to-I substitution

Transcription

DNA

mRNA transcript

gRNA

Base-pairing

Edited mRNA

Translation

Protein

Transcription

DNA

ADAR

ds Pre-mRNA

substitution

Edited mRNA

Translation

Protein

Edition

rocha@indiana.edu
http://informatics.indiana.edu/rocha
RNA editing modulates gene expression

Genotype DNA

Transcription

RNA

Code

Amino Acid Chains

Development

Phenotype Environmental Ramifications

Only mutations that occur during DNA replication can become permanent and heritable.

RNA Editing may produce different mRNA’s (and thus proteins), but editions are not inherited.

- What is inheritable, and subjected to variation, is the genetic material (both coding and non-coding) which is ultimately selected and transmitted to the offspring of the organism.

traditional genetic algorithms

artificial models of neo-Darwinian natural selection

**Traditional Genetic Algorithm**

- **Genotype**
  - $S_1$, $S_2$, ..., $S_n$
  - $X_1$, $X_2$, ..., $X_n$
- **Variation**
- **Selection**

**Contextual Genetic Algorithm**

- **Variation**
  - $S_1$, $S_2$, ..., $S_n$
- **Pre-translation Info-processes**
- **Development**
  - $Y$, $X_1$, $X_2$, ..., $X_n$
- **Selection**

Used for *optimization* of solutions for different problems. Uses the syntactic operators of *crossover* and *mutation* for variation of encoded solutions, while selecting best solutions from generation to generation. Holland, 1975; Goldberg, 1989; Mitchell, 1995.

- Indirect Encoding
- Contextual Dependencies
  - Allows regulation
- Development
  - Smaller Genotypes, Self-Organization
- RNomics (Syntactic/Info) Operators
  - Extra Variation, One-to-many encoding, network regulation

Rocha, Luis M. [2001]. *Biosystems*. 60. pp. 95-121

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Indiana University
genetic algorithms with edition


![Diagram of genetic algorithm components](image-url)

- **Population of Strings $P$**: The population consists of multiple strings $S_1, S_2, ..., S_{n_p}$.
- **Variation**:
  - Family of $l$ editors ($\mathcal{E}$, $\mathcal{F}$)
  - Concentrations: $0.5$
- **Edited Population**: $S'_1, S'_2, ..., S'_{n_p}$
- **Selection**:
  - Solution alternatives in $X$: $X_1, X_2, X_3, ..., X_{n_p}$

Contact:
- rocha@indiana.edu
- [Informatics, Indiana University](http://informatics.indiana.edu/rocha)
contextual genetic algorithms

linking editor concentrations to environment: phenotypic plasticity


Semantic relations

Editors Change

Population of Strings $P$

Concentrations

Ministrings functions

Symbolic Descriptions

Solutions

Solution alternatives in $X$

Fitness

Environmental demands change
agent-based model of genotype editing

population of agents with codotype and editype

- Genome contains both coding and non-coding portions
  - Codome and Editome (Editosome)
- For each agent
  - Codotype edited by editype before “translation”
- Modeling pre-translation information (syntactic) processes
  - No RNA/DNA distinction
  - A process of *non-inheritable alteration of genotypes via edition*, not any specific type of RNA Editing.
  - Not mutation
- Co-evolution of editype and codotype
  - Not in the EC sense of independent populations
  - Independent variation


editype and codotype variation in the ABMGE

separate crossover and mutation for codotype and editype
Testing the ABMGE

\[
F(x) = \sum_{s_i \in S} c_i \sigma_{s_i}(x)
\]

- Miniature of "Royal Road" function (Forrest and Mitchell, 1993)
  - Schemata \(S = (s_1, ..., s_8)\)
  - \(c_i\) is a value assigned to each schema \(s_i\)
  - \(\sigma_{s_i}(x) = 1\) if \(x\) is an instance of \(s_i\) and 0 otherwise
  - Fitness of the global optimum string (40 1's) is \(10 \times 8 = 80\)
    - Class of building-block-based functions, in which search advancements depend entirely on the discovery and exploitation of building blocks. This serves as an idealized testbed for observing how editing improves the GA's search power by tracing the origin of each advance in performance.
- Evolutionary algorithm Parameters
  - Binary tournament selection
  - population size 40 over 200 generations for 50 runs
- Codotype
  - Length: 40 bits
  - one-point crossover rate: 0.7
  - mutation rate: 0.005
- Editype
  - Size of editor Family: sampled from \{1, \ldots, 5\}
  - Editor length: sampled from \{2, \ldots, 4\}
  - Concentration: sampled from [0,1]
  - Editing function: inserts or deletes a number of random bits sampled from \{1, \ldots, 3\}
    - Fixed at the start
  - Editype Mutation: 0.01, 0.05
  - Editype Crossover: 0, 0.3, 0.5, 0.7, 0.9
- Averaged best-so-far for 50 runs
  - vertical bars represent the 95-percent confidence intervals
- ABMGE
  - optimum found in all 50 runs
    - Best-so-far = 80
- Traditional GA
  - optimum found in 17 out of 50 runs
    - Best-so-far approx 70
    - search power is limited by hitch-hiking

**F3**

- simple, unimodal step function
  - global optimum easily located
- compare mean function evaluations required for locating the optimum

**Evolutionary algorithm Parameters**
- population size 50 over 400 generations for 100 runs

**Codotype**
- Length: 50 bits

**Editype**
- Size of editor Family: sampled from \{1,2\}, \{1,...,5\}, \{1,...,8\}
**ABMGE**

- **Evolutionary algorithm Parameters**
  - population size 50 over 400 generations for 100 runs
- **Codotype**
  - Length: 50 bits
- **Editype**
  - Size of editor Family: sampled from \{1,2\}, \{1,...,5\}, \{1,...,8\}

---

**De Jong Static generations=200 runs=100**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Editype Parameters</th>
<th>Mean (Std. Err.) eval. to optimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA</td>
<td>N/A</td>
<td>3237.25 (194.15)</td>
</tr>
<tr>
<td>ABMGE</td>
<td>No Crossover and (r \in {1, 2})</td>
<td>1422.96 (178.37)</td>
</tr>
<tr>
<td>ABMGE</td>
<td>No Crossover and (r \in {1, \cdots, 5})</td>
<td>898.74 (172.19)</td>
</tr>
<tr>
<td>ABMGE</td>
<td>No Crossover and (r \in {1, \cdots, 8})</td>
<td>586.62 (72.82)</td>
</tr>
<tr>
<td>ABMGE</td>
<td>Crossover and (r \in {1, 2})</td>
<td>770.09 (122.77)</td>
</tr>
<tr>
<td>ABMGE</td>
<td>Crossover and (r \in {1, \cdots, 5})</td>
<td>329.15 (50.75)</td>
</tr>
<tr>
<td>ABMGE</td>
<td>Crossover and (r \in {1, \cdots, 8})</td>
<td>252.68 (46.98)</td>
</tr>
</tbody>
</table>

---

Optimal control problems involve nonlinearity in:
- Existence of multiple local optima in the area of interest.
- Premature convergence.
  - Height of the hill is much lower than the spikes, but it occupies most of the search space
  - Most of the population individuals attracted to the hilltop

---

Experiment results

Optimal Control - Static - 200 Generations

- **ABMGE**
  - Explores more of the search space
- **GA**
  - Best-so-far is 27.01 (the fitness at the hilltop) in nearly 60 out of 100 runs.

**Evolutionary algorithm Parameters**
- Population size 50 over 200 generations for 100 runs

**Codotype**
- Length: 100 bits

**Editype**
- Size of editor Family: sampled from {1,2}, {1,...,5}, {1,...,10}
multimodality

- Many local optima
  - To attain the global optimum at the center of the search space, the population has to cross over many deep wells and high barriers.

experiment results

- **Evolutionary algorithm Parameters**
  - population size 50 over 200 generations for 100 runs

- **Codotype**
  - Length: 100 bits

- **Editype**
  - Size of edit Family: sampled from \( \{1, \ldots, 5\} \), \( \{1, \ldots, 10\} \), \( \{1, \ldots, 20\} \)

---


Schwefel function

difficult path to optima

- **Deceptive testbed**
  - global maximum is geometrically distant, over the parameter space, from the next best local optima
  - EC tend to converge in the wrong direction
Schwefel function

Experiment results

- **Evolutionary algorithm Parameters**
  - Population size 30 over 1000 generations for 100 runs

- **Codotype**
  - Length: 100 bits

- **Editype**
  - Size of editor Family: sampled from \{1, \ldots, 5\}, \{1, \ldots, 10\}, \{1, \ldots, 20\}

Huang, C-F, J. Kaur, A. Maguitman and L.M. Rocha [2007].
*Evolutionary Computation*. In Press.
Oscillation period
- 100 (50, 200) generations

First 1000 generations
- Same parameters as in static case

Small Royal Road 1 (SRR1)

Small Royal Road 0 (SRR0)


rocha@indiana.edu
http://informatics.indiana.edu/rocha
dramatic environmental changes


**Oscillation period**
- 100 generations

**Last 1000 generations from 4000**
- Same parameters as in static case
Oscillatory royal road

behavior of 3 algorithms

GA

ABMGE no Edt Cross

ABMGE
dynamic Shaffer function

- **Severity** parameter that changes fitness function every p generations
  - 100 (50, 200)
  - Linear and jumping dynamics
- 50 agents
  - 1000 generations
  - 100 runs

**Dynamic Shaffer Function**

- **Severity parameter** that changes fitness function every $p$ generations
  - $100$ (50, 200)
  - Linear and **jumping** dynamics
- **50 agents**
  - 1000 generations
  - 100 runs

Contact: rocha@indiana.edu
http://informatics.indiana.edu/rocha
Dynamic optimal control function


- **Severity parameter that changes fitness function every p generations**
  - 100 (50, 200)
  - Linear and jumping dynamics
- **50 agents**
  - 1000 generations
  - 100 runs
Dynamic Schwefel function

- **Severity** parameter that changes fitness function every p generations
  - 50 (100, 200)
  - Linear and jumping dynamics
- 50 agents
  - 1000 generations
  - 100 runs


rocha@indiana.edu
http://informatics.indiana.edu/rocha
exploration and exploitation with genotype editing

single run fitness distribution (Schwafel function example)
100 run fitness distribution

dynamic Schwafel function example (GA)
100 run fitness distribution

dynamic Schwafel function example (ABMGE)
exploration and exploitation with genotype editing

dynamic Schwefel function fitness distribution videos

GA 1 run

ABMGA 10 run

rocha@indiana.edu
http://informatics.indiana.edu/rocha/editing
Agents search fitness space with a stochastic “cloud” of phenotypes
  - Rocha (1995)

Stochastic phenotypic diversity
  - The same genotype repeated in the population, produces different phenotypes
    - even though only one genotype is inherited and subjected to variation
  - In dynamic environments, when the environment changes, the population of agents is capable of producing a few higher fitness phenotypes.

Better exploitation/exploration balance than GA
  - Evolution leads to exploitation of good volumes of genotype space
    - Only genotype is inherited
  - But genotype editing allows one single genotype to explore different areas of the fitness space.
    - Stochastic “expression” of phenotypes
effects of codotype mutation rates on ABMGE and GA

- On Royal Road
  - Best mutation for GA is 0.005 (with crossover 0.7)
effects of codotype mutation rates on GA

- **On De Jong**
  - Best mutation for GA is 0.005 (with crossover 0.7)
effects of codotype mutation rates on dynamic Royal Road

- On Dynamic Royal Road
  - Best mutation for GA is 0.05 (with crossover 0.7)
effects of codotype mutation rates on dynamic Royal Road

On Dynamic Royal Road
- Best mutation for ABMGE is 0.005

0.05

0.005
effects of codotype mutation rates on static optimal control

- On Dynamic Royal Road
  - Best mutation for ABMGE is 0.005

rocha@indiana.edu
http://informatics.indiana.edu/rocha
effects of codotype mutation rates
on dynamic optimal control

Optimal Control - Dynamic - 1000 Generations, Update freq 100, Severity 0.1, Runs 100, Size 50

- GA Pm 0.05
- ABMGE Pm 0.05

rocha@indiana.edu
http://informatics.indiana.edu/rocha
effects of codotype mutation rates on static Schaffer

- On Static Schaffer
  - Best mutation for GA is 0.05

rocha@indiana.edu
http://informatics.indiana.edu/rocha
effects of codotype mutation rates

- On Dynamic Shaffer
  - Best mutation for GA is 0.05
agent-based model of genotype editing

Discussion and future

- Focus on differences on analysis of search characteristics
- Larger and more realistic search spaces
- Include artificial gene regulatory network

rocha@indiana.edu
http://informatics.indiana.edu/rocha