ON ABOUT WHAT CAN BE DONE AND WHAT CANNOT BE DONE WITH GENETIC ALGORITHMS IN PHYLOGENETIC TREE AND GENE SEQUENCE ANALYSES

Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ

Technical University, "Iuliu Haţieganu" University of Medicine and Pharmacy, and University of Agriculture Sciences and Veterinary Medicine, Cluj-Napoca, Cluj, Romania

Genetic algorithms, a sort of algorithms belonging to a more general category, so called meta-heuristics, know today a great expansion in terms of target applications, including biology, chemistry and medicine. They are inspired from primary observations in nature (Lamarck, 1809; Darwin, 1859; Mendel, 1865; Fisher, 1918), and started with simulation of artificial selection of organisms with multiple loci that controls a measurable trait (Fraser, 1957). Genetic algorithms evolved into complex and strong informatics tools capable to deal with hard problems of decision, classification, optimization, and simulation in fields as biology, chemistry and medicine. The aim of the present article was to introduce genetic algorithms and to present their suitability for biological hard problems. Some important results reported in the literature about the use of genetic algorithms for phylogenetic and gene sequence analysis is discussed.

Genetics discoveries - the artefacts of genetic algorithms:

(Linnaei, 1735)	Systematic classification of vegetables - first phylogenetic trees
(Lamarck, 1809)	Systematic studies of vegetables treats and characters - first hard problem
	Soft inheritance theory
(Darwin, 1859)	Evolution theory
	Natural selection
	Survival of the fittest
(Mendel, 1866)	Characters inheritance through crossover
(Weismann, 1893)	Hard inheritance theory
(Morgan, 1915)	Chromosomal theory
(Muller, 1927)	Mutagens
(Fisher, 1914-1962)	Evidences for evolution theory

Genetic algorithm - it realizes an evolution of individuals from a population (sample) by iterating successive generations:

Design	a chromosome (as a series of genes)
Construct	a transformation between a chromosome (genotype) and a individual (phenotype)
Chouse	a fitness score function
Establish	the condition for a solution to be a good solution (stop condition)
Build	an initial population (random or determinist) of genotypes
Repeat	obtain phenotypes from genotypes ÷ assess their fitness
	do selection to pick up 2 (or more) individuals ÷ do crossover of the selected individuals (offspring's obtaining)
	decide (using a low probability) if mutation applies to a offspring ÷ then chouse (random or determinist) gene/genes for mutation and mutate gene/genes
Until	at least one phenotype satisfies an imposed condition (usually of fitness)

Genetic algorithms offspring's:

Ant colony optimization	(Bouktir and Slimani, 2005)
Bacteriologic algorithms	(Benoit et al., 2005)
Cross-entropy method	(De Boer et al., 2005)
Cultural algorithms	(Kobti et al., 2004)
Evolution strategies	(Schwefel, 1995)
Evolutionary programming	(Fogel et al., 1966)
Extremal optimization	(Bak and Sneppen, 1993)
Gaussian adaptation	(Kjellström, 1991)
Genetic programming	(Banzhaf et al., 1997)
Memetic algorithm	(Smith, 2007)
Others	(Davis, 1991)

Genetic algorithms based software:

Clustal X	Multiple alignment of nucleic acid and protein sequences
GABRIEL	Genetic Analysis By Rules Incorporating Expert Logic
GAL	Genetic Algorithm Library 1.3 Open Source
GAML	Genetic Algorithm for Maximum Likelihood
GARLI	Genetic Algorithm for Rapid Likelihood Inference
GDE	Genetic Data Environment (v. 2.2)
GOAL	Genetic Optimization Algorithm
LamarckianGA	Lamarck type Genetic Algorithm
MacClade	Clade Analysis
MEGA	Molecular Evolutionary Genetic Analysis
Mesquite	Modular system for evolutionary analysis
metaGA	Meta-Population Genetic Algorithm
ModelTest	v. 3.06, tool to select the best-fit model of nucleotide substitution
MrBayes	v. 3.1, Bayes type searches
MS BLAST	Mass Spectrometry Basic Local Alignment Search Tool
NCBI BLAST	National Center for Biotechnology Information Basic Local Alignment Search Tool
PAUP*	v. 4.0, Phylogenetic Analysis Using Parsimony
r8s	v. 1.7, estimating absolute rates
RAGA	RNA sequence Alignment by Genetic Algorithm
SAGA	Sequence Alignment by Genetic Algorithm
S-System	Savageau System

Some results reported using GAs:

phylogenetic trees study methodology	(Diaconis and Holmes, 1998)
ontogenetic shift during development under predation risk in food and habitat use	(Bouskila et al., 1998)
phylogenetic tree analysis on major brachyceran lineages	(Wiegmann et al., 2003)
evolution and diversification of ants	(Brady et al., 2006)
ants implication on agricultural systems	(Schultz and Brady, 2008)
spider and howler monkeys population distributions	(Ortiz-Martınez et al., 2008)
gene expression kinetics	(Busch et al., 2008)
dynamics of regulator and effector genes in inducible circuits	(Kikuchi et al., 2003)
bioinformatic analysis and design of peptides	(Del Carpio et al., 2002)
bacterial polysaccharide antigenicity of *Streptococcus agalactiae* vs *Streptococcus pneumoniae*	(Kadirvelraj et al. 2006)

BIBLIOGRAPHY

Anderson RP, Lew D, Peterson AT. 2003. Evaluating predictive models of species distributions: criteria for select optimal models. Ecol Model 162(3):211-232.

Bak P, Sneppen K. 1993. Punctuated equilibrium and criticality in a simple model of evolution. Phys Rev Lett 71(24):4083-4086.

Banzhaf W, Nordin P, Keller RE, Francone FD. Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications. San Francisco: M Kaufmann Publishers, pp. 450.

Barricelli NA. 1954. Numerical examples of evolution processes (in Italian). Methodos 1954:45-68.

Benoit B, Fleurey F, Jézéquel J-M, Le Traon Y. 2005. Automatic Test Case Optimization: A Bacteriologic Algorithm. IEEE Software 22(2):76-82.

Bolboacă SD, Jäntschi L. 2007. Design of Experiments: Useful Orthogonal Arrays for Number of Experiments from 4 to 16, Entropy 9(4):198-232.

Bosworth J, Norman F, Zeigler BP. 1972. Comparison of Genetic Algorithms with Conjugate Gradient Methods. NASA Contractor Reports, CR-2093.

Bouktir T, Slimani L. 2005. Optimal Power Flow of the Algerian Electrical Network using an Ant Colony Optimization Method, Leonardo J Sci 4(7):43-57.

Bouskila A, Robinson ME, Roitberg BD, Tenhumberg B. 1998. Life-history decisions under predation risk: Importance of a game perspective. Evolut Ecol 12(6):701:715.

Brady SG, Schultz TR, Fisher BL, Ward PS. 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. Proc Natl Acad Sci 103(48):18172-18177.

Brauer R. 1937. On algebras which are connected with the semisimple continuous groups. Ann Math 38(4), 857-872.

Busch H, Camacho-Trullio D, Rogon Z, Breuhahn K, Angel P, Eils R, Szabowski A. 2008. Gene network dynamics controlling keratinocyte migration. Molec Syst Biol 4:199(16).

Darwin CR. 1859. On the origin of species by means of natural selection. London: J Murray, pp. 459.

Davis L. 1987. Genetic Algorithms and Simulated Annealing. San Francisco: M Kaufmann, pp. 216.

Davis L. 1991. The Handbook of Genetic Algorithms. New York: VN Reinhold, pp. 385.

De Boer P-T, Kroese DP, Mannor S, Rubinstein RY. 2005. A Tutorial on the Cross-Entropy Method. Ann Oper Res 134(1):19-67.

Del Carpio-Muñoz CA, Hennig T, Fickel S, Yoshimori A. 2002a. A combined bioinformatic approach oriented to the analysis and design of peptides with high affinity to MHC class I molecules. Immun Cell Biol 80(3):286-298.

Del Carpio-Muñoz CA, Ichiishi E, Yoshimori A, Yoshikawa T. 2002b. MIAX: A new paradigm for modeling biomacromolecular interactions and complex formation in condensed phases. Prot Struct Funct Genet 48(4):696-732.

Diaconis PW, Holmes SP. 1998. Matchings and phylogenetic trees. Proc Natl Acad Sci USA 95(25):14600-14602.

Falkenauer E. 1998. Genetic Algorithms and Grouping Problems. New York:Wiley, pp. 220.

Fink WL. 1986. Microcomputers and phylogenetic analysis. Science 234(4780):1135-1139.

Fisher RA. 1918. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Edinb Roy Soc Trans 52:399-433.

Fogel LJ. 1999. Intelligence Through Simulated Evolution: Forty Years of Evolutionary Programming. New York: Wiley Interscience, pp. 162.

Fraser A. 1957. Simulation of genetic systems by automatic digital computers. I. Introduction. Aust J Biol Sci 10:484-491.

Glover F. 1977. Heuristics for Integer Programming Using Surrogate Constraints. Decision Sci 8(1):156-166.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. Proc Natl Acad Sci USA 84(13):4355-4358.

Higuchi T, Tsutsui S, Yamamura M. 2000. Theoretical analysis of simplex crossover for real-coded genetic algorithms. Proc Int Conf Parall Probl Solv Nature 2000:365-374.

Hlavacek WS, Savageau MA. 1996. Rules for coupled expression of regulator and effector genes in inducible circuits. J Mol Biol 255(1):121-139.

Huelsenbeck JP, Ronquist FR, Nielsen R, Bollback JP. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. Science 294(5550), 2310-2314.

Jäntschi L, Diudea MV. DOI. Subgraphs of Pair Vertices. J Math Chem DOI:10.1007/s10910-008-9411-6.

Kadirvelraj R, Gonzalez-Outeirino J, Foley BL, Beckham ML, Jennings HJ, Foote S, Ford MG, Woods RJ. 2006. Understanding the bacterial polysaccharide antigenicity of Streptococcus agalactiae versus Streptococcus pneumoniae. Proc Natl Acad Sci USA 103(21):8149-8154.

Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M. 2003. Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics 19(5):643-650.

Kjellström G. 1991. On the Efficiency of Gaussian Adaptation. J Optim Theor Appl 71(3):589-597.

Kobti Z, Reynolds RG, Kohler T. 2004. Agent-Based Modeling of Cultural Change in Swarm Using Cultural Algorithms, SwarmFest 8:8p.

Lamarck JBPA. 1830. An Exposition of Zoological Philosophy (…, in French). Paris: JB Baillère, pp. 420+450.

Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947-2948.

Lemmon AR, Milinkovitch MC. 2002. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. Proc Natl Acad Sci USA 99(16):10516-10521.

Lewis PO. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol Biol Evol 15(3):277-283.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol 50(6):913-925.

Maddison DR, Maddison WP. 2000. MacClade v4.0. Sunderland: Sinauer (software).

Maddison WP, Maddison DR .2006. Mesquite: A modular system for evolutionary analysis v1.1. Tucson: University of Arizona (software).

Mendel G. 1865. Experiments in Plant Hybridization (in German), Meetings of the Brünn Natural History Society, February 8th and March 8th, Brno.

Ortiz-Martınez T, Rico-Gray V, Martınez-Meyer E. 2008. Predicted and verified distributions of Ateles geoffroyi and Alouatta palliata in Oaxaca, Mexico. Primates 49(3):186-194.

Posada D, Crandall KA. 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14(9):817-818.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12):1572-1574.

Sanderson MJ. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. Mol Biol Evol 19:101-109.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301-302.

Savageau MA. 1976. Biochemical System Analysis: a Study of Function and Design in Molecular Biology. Reading: Addison-Wesley, pp. 199.

Schultz A-K, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B, Stanke M. 2006. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC Bioinform 7:265(15).

Schwefel H-P. 1995. Evolution and Optimum Seeking. New York: Wiley & Sons, pp. 456.

Smith JE. 2007. Coevolving Memetic Algorithms: A Review and Progress Report. IEEE Trans Syst Man Cy B 37(1):6-17.

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM. 1994. The genetic data environment and GUI for multiple sequence analysis. Comput Appl Biosci 10(6):671-675.

Tominaga D, Koga N., Okamoto M. 2000. Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. Proc Genet Evolut Comput Conf 2000:251-258.

Tominaga D, Okamoto M. 1998. Design of canonical model complex nonlinear dynamics. Proc Int Conf Comp Appl Biotechn 1998:85-90.

Tsutsui S, Yamamura M, Higuchi T. 1999. Multi-parent recombination with simplex crossover in real coded genetic algorithms. Proc Genet Evolut Comput Conf 1999:657-664.

Wiegmann BM, Yeates DK, Thorne JL, Kishino H. 2003. Time Flies, a New Molecular Time-Scale for Brachyceran Fly Evolution Without a Clock. System Biol 52(6):745-756.

Yeates DK. 2002. Relationships of extant lower Brachycera: A quantitative synthesis of morphological characters. Zool Scripta 31(1):105-121.

ACKNOWLEDGMENTS

Authors acknowledge that the research was conducted being supported in part by national research grants

LJ: ID1051/UEFISCSU

SDB: ID0458/UEFISCSU

RES: PCCP1177/CNMP