Bulletin UASVM, Horticulture 65(1)/2008
pISSN 1843-5254; eISSN 1843-5394

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

Lorentz JÄNTSCHI[1], Sorana D. BOLBOACA[2], Radu E. SESTRAS[3]

[1] Technical University of Cluj-Napoca, 103-105 Muncii Bvd, 400641 Cluj-Napoca, Romania
lori@chimie.utcluj.ro
[2] "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 E. Isac, 400023 Cluj-Napoca, Romania, sbolboaca@umfcluj.ro
[3] University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Mănăştur, 400372 Cluj-Napoca, Romania, rsestras@usamvcluj.ro

Keywords: genetic algorithms; hard problems; evolution; phylogenetic trees; gene sequences

Abstract: 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 are discussed.

INTRODUCTION

Genetic algorithms know today a great expansion in terms of target applications, including biology, chemistry and medicine. They are a sort of algorithms belonging to a more general category, so called meta-heuristics, and an entire field of research is devoted to it: evolutionary programming. Genetic algorithms (GAs) are inspired from primary observations in nature (Lamarck, 1809; Darwin, 1859; Mendel, 1865; Fisher, 1918).

First computer simulations of evolution (today known as GAs) started with a work of Nils Aall Barricelli (Barricelli, 1954). Shortly, Alex Fraser (Fraser, 1957) published a series of papers on simulation of artificial selection of organisms with multiple loci controlling a measurable trait. Fraser's simulations included all of the essential elements of modern genetic algorithms.

Four categories of problems may be the subject of a GA: decision, classification, optimization and simulation. According to Falkenauer (1998) these categories are equivalent at least in theory (a problem of decision can be transformed into an optimization problem, and so on). In fact, a genetic algorithm is suitable in finding solution for hard problems. A hard problem is defined as having an exponential complexity and for these problems even the best classical algorithm will probably be unusable on real-world instances because the search for the optimum often goes into out of time. A large set of hard problems encountered in practice do not necessarily call for the optimum. For these problems heuristics are the available alternative. These are rules of thumb which recipes for solving a particular problem, usually based on common sense, avoiding obvious mistakes.

Together with GAs (Bosworth et al., 1972), other two meta-heuristics proved their feasibility in practice: TS - tabu search (Glower, 1977) and SA - simulated annealing (Davis, 1987). All three

are stochastic in nature and two of them (GAs and SA) are based on natural processes that have been taking place around us ever since.

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics and were invented to mimic some of the processes observed in natural evolution with the idea to use this power of evolution to solve optimization problems. GAs are designed to simulate processes in natural systems necessary for evolution, specially those follow the principles laid down by the pioneers of the modern genetics; since in nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker ones:

÷ Jean-Baptiste Lamarck's "soft inheritance" (Lamarck, 1809);
÷ Charles Darwin's "survival of the fittest" (Darwin, 1859);
÷ Gregor Mendel's "particulate inheritance of genes" (Mendel, 1865);
÷ Sir Ronald Aylmer Fisher's "genetic model" (Fisher, 1918).

Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype of the genome) of candidate solutions (called individuals, creatures, or phenotypes) subject to an optimization problem, which evolves toward better solutions. GAs simulates the survival of the fittest among individuals over consecutive generation for solving a problem. Each generation consists of a population of character strings analogous to the DNA chromosomes. Each individual represents a point in a search space and a possible solution. The individuals in the population are then made to go through a process of evolution. GAs is based on an analogy with the genetic structure and behaviour of chromosomes within a population of individuals.

Thus, the GA's search space is composed from genes (similarly with letters), chromosomes (similarly with words), and genotype (a family of words). The operators presented in Table 1 are fundamental for GA's.

Table 1

Gas operators

Operator	Individual(s)/Population	How operate
Crossover	Two (identified by their chromosomes)	Genes interchangement
Mutation	One (identified by its chromosomes)	Gene mutation(s)
Fitness	One (identified by its chromosomes)	Its strength in the population (identified by the genotype) assessed trough a fitness (score) function
Selection	Population	Selection of the individuals for reproduction (crossover and mutation) through a selection process based on the fitness of the individuals: (**1**) Rational: the chance of reproduction is related with the fitness (using a probability mass function); (**2**) Deterministic: Reproduction are made with best or worst individuals (elitism); (**3**) Tournament: Pairs of individuals compete for selection

If the reproduction is based on chance (using a probability mass function) then the chance may be proportional with the fitness (Proportional), a fixed scale may be used to normalize fitness between different generations (Normalization), or the chance of reproduction is proportional with the rank of fitness (Ranking).

There are many variants and adaptations of GAs in order to improve its performances for a given type of problem. Some problems and Gas solution are as follows: 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), and so on (Davis, 1991).

MATERIALS AND METHODS

Several works reporting results obtained by using of GA retained our attention. These reports deal with the subject of population and evolution, and gene expression analysis respectively.

Population and evolution key field where subject of investigation using GA, main subject being phylogenetic trees investigation. Thus, (Diaconis and Holmes, 1998) discussed phylogenetic trees study methodology, (Bouskila et al., 1998) approached ontogenetic shift during development under predation risk in food and habitat use, (Wiegmann et al., 2003) phylogenetic tree analysis on major brachyceran lineages, (Brady et al., 2006) evolution and diversification of ants and their implication on agricultural systems (Busch et al., 2008). Ortiz-Martınez and co-authors (2008) discussed spider and howler monkeys population distributions.

On field of gene expression analysis, Busch and co-authors (2008) studied gene expression kinetics - DNA microarray data of hepatocyte growth factor-induced migration of primary human keratinocytes, Kikuchi and co-authors (2003) modeled the dynamics of regulator and effector genes in inducible circuits, Del Carpio and co-authors (2002) presents a bioinformatic analysis and design of peptides, and Kadirvelraj and co-authors (2006) analyze the understanding of bacterial polysaccharide antigenicity of *Streptococcus agalactiae* versus *Streptococcus pneumoniae*.

RESULTS AND DISCUSSION: WHAT ARE DONE WITH GENETIC ALGORITHMS?

Modeling the ontogenetic shift during development under predation risk in food and habitat use was subject of Bouskila and co-authors (1998). Authors was able to compare results under four scenarios of increasing complexity: (1) no predation; (2) constant predation; (3) frequency-dependent predation (predation risk diluted at high prey density); and (4) frequency-dependent predation as in (3) but with predators allowed to respond adaptively to prey behavior. Two components of the strategies were identified and quantified: the mean and the variance of the switching time distribution. Selection for decreased variance is selection for dilution. With these quantities the following decision tree to select the most suitable model were obtained:

- Is there selection for decreasing the variance of the switching time distribution when the predation risk is increased?
  - Yes, then is a victim model
  - No, then selection increases will increase the mean switching with the predation risk?
    - Yes, then is a model with constant predation risk
    - No, then is a victim-prey game model.

Phylogenetic trees study using a correspondence with the set of perfect matching in the complete graph (for complete graphs see Jäntschi and Diudea, DOI) was subject of (Diaconis and Holmes, 1998). Authors proved that the correspondence produces a distance between phylogenetic trees, and became a way of enumerating all trees in a minimal step order. As the main issue of the phylogenetic trees study, finding of the optimal tree is a hard problem, based on this correspondence authors showed that using a method for making a product of two matching in what is known as the Brauer algebra (Brauer, 1937), which enables a simple implementation of a genetic algorithm. The problem of large taxon samplings in phylogeny estimation is discussed in (Lemmon and Milinkovitch, 2002), when a meta-population genetic algorithm (metaGA), involving several populations of trees that are forced to cooperate in the search for the optimal tree was found suitable. An important result of Lemmon and Milinkovitch (2002) is that the frequencies with which trees

and clades are sampled by using the metaGA might correspond to unbiased estimates of their posterior probabilities (Huelsenbeck et al., 2001).

Another phylogenetic tree analysis on major brachyceran lineages was published by Wiegmann and co-authors (2003), and indicates that the Brachycera originated in the late Triassic or earliest Mesozoic and that all major lower brachyceran fly lineages had near contemporaneous origins in the mid-Jurassic prior to the origin of flowering plants (angiosperms). Authors obtained an increased resolution of brachyceran phylogeny, and the revised estimates of fly ages improve the temporal context of evolutionary inferences and genomic comparisons between fly model organisms. Nucleotide sequences were aligned manually with the on-screen multiple alignment editor of Genetic Data Environment 2.2 (Smith et al., 1994). The phylogenetic data included 2,220 characters from the 28S rDNA (608 variable and 294 parsimony informative among all taxa; 493 variable and 296 informative within Brachycera) and 101 morphological characters (Yeates, 2002). Phylogenetic analysis of the combined data set was carried out via parsimony with the program PAUP (Fink, 1986).

A comprehensive study about the early evolution and diversification of ants was reported in (Brady et al., 2006). An important part of this study is represented by the used methods (presented as supporting information on the journal website). Thus, authors used a series of programs all having GA implemented (see Table 2).

Table 2

GAs Programs

Program	Feature
Clustal X (Larkin et al., 2007)	sequence alignment
PAUP* v4.0b10 (Fink, 1986)	divergence dating & phylogenetic inference analysis
ModelTest v3.06 (Posada and Crandall, 1998	nucleotide substitution models
GARLI v0.94 (Schultz et al., 2006)*	nonparametric bootstrap maximum-likelihood analyses
MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003)	Bayesian analyses
r8s v1.7 (Sanderson, 2002; Sanderson, 2003)	divergence dating (using the penalized likelihood approach)

* derived from GAML (Lewis, 1998)

Two of the authors of the previous study continued the research on ants and reported (Schultz and Brady, 2008) an identification of relict, extant attine ant species that occupy phylogenetic positions that are transitional between the agricultural systems. The used methodology includes phylogenetic analyses (parsimony, maximum likelihood and divergence dating), Bayesian nucleotide-model and codon-model Markov chain Monte Carlo, and phylogenetic mapping of agricultural systems:

- Terminal taxa were assigned states for a single six-state character representing the four attine agricultural systems and leaf-cutter agriculture (no, lower, yeast, higher, leaf-cutter, coral-fungus);
- Five species (*Myrmicocrypta* n. sp. Brazil, *Mycetagroicus triangularis*, *Cyphomyrmex* n. sp., *Cyphomyrmex morschi*, *Trachymyrmex irmgardae*, and *Pseudoatta* n. sp.) states were assigned to 'unknown', and *Trachymyrmex papulatus* received a 'lower agriculture' state assignment based on a single garden collection from Argentina (a second colony from the same locality cultivated a typical higher attine garden);
- Character evolution was optimized onto the Bayesian codon-model consensus tree (with branch lengths) under both parsimony using MacClade (Maddison and Maddison, 2000) and maximum likelihood using the StochChar module provided in the Mesquite package (Maddison and Maddison, 2006);
- Under parsimony, ancestral-state optimizations were unambiguous. Under the Markov k-state 1-parameter model (Lewis, 2001), the likelihood that each agricultural system arose in the most

recent common ancestor of the corresponding ant clade was, as a proportion of the total probability distributed across the six character states, 0.9831 for lower, 0.9995 for yeast, 0.9905 for higher, 0.9924 for leaf-cutter, and 0.9998 for coral-fungus agricultures.

Other phylogenetic tree analysis was conducted using a GA for rule-set production in order to model spider and howler monkeys population geographic distributions by characterizing their ecological niches (Ortiz-Martınez et al., 2008). Due to the random processes involved in model development, each model produced with a single occurrence dataset are different; to capture the variability authors developed 100 replicate models for each species and then selected the ten models that gave the smallest errors of commission and omission, following the procedures described in (Anderson et al., 2003). Authors were able to obtain that that spider monkeys occupy a wider area and elevational range than howler monkeys. Validation of the model was done for spider monkeys, being enough field data for this species; the validation indicated that the predicted distribution of the species was statistically better than expected by chance.

Shifting to gene sequence applications of GA, (Busch et al., 2008) repeatedly fitted using a genetic algorithm the experimental data from time-series measurements of DNA microarray data of hepatocyte growth factor-induced migration of primary human keratinocytes. Authors succeeded to prove that the inverse modeling of a gene network establish an abstract model with predictive power of a complex biological system able to capture its dynamic properties. The analysis of gene expression kinetics deciphers the dynamics of a small, but extremely complex gene regulatory model of migration by inverse modeling. As authors' underlines, the design of the experiment (for design of experiments see Bolboacă and Jäntschi, 2007) is crucial and was the key element of the success revealing the complex orchestration of multiple pathways controlling cell migration:

- The choice of experimental time window focusing on the decision to migrate and measure at a sufficiently high sampling rate to capture the gene network dynamics;
- The choice of gene candidates from ranking;
- Biological function and the cross-validation against proliferative stimuli;
- The limited number of genes dominating the network dynamics;
- The combination of fitness and robustness criteria in the search for a biologically plausible gene interaction model.

An interesting approach is the usage of the S-system (Savageau, 1976) formalism (a type of power-law formalism involved in stoechiometric pre-equilibrium reactions) for modeling of the dynamics of regulator and effector genes in inducible circuits (Hlavacek and Savageau, 1996). Because the basic method (Tominaga and Okamoto, 1998) was able to predict only a very small number of parameters (Tominaga et al., 2000) and the convergence rate was low, a series of improvements were applied in (Kikuchi et al., 2003) on fitness function, crossover method, and optimization strategy as follows:

- Fitness function, as a sum of:
  - Fitness of the basic method (Tominaga and Okamoto, 1998) reproducing given time-courses (it converges to multiple local minima and rarely attains skeletal structures);
  - A pruning term (allowing skeletal structures detection with finding expectation of unknown pathways) being the sum of the absolute values of model parameters (Savageau, 1976).
- Simplex crossover as (Tsutsui et al. 1999) proposed, and (Higuchi et al., 2000) showed that it improved the optimization speed, having the following features:
  - Inherit independence of coordinate systems;
  - Offspring vector values inherit the characteristics of parents and sampling reflect a certain linkage among the parameters;
  - Balances between exploration and exploitation in generating offspring, and it works well on functions with multimodarity and/or epistasis among the parameters;

- Is a simple and non-time consuming operator.

÷ Gradual optimization strategy: Although it is difficult to optimize all the parameters at once, parameters of comparatively lower importance, which become almost 0, are detectable. These values are fixed to 0 and optimization is again done from the beginning, more parameters of lower importance are detected.

Del Carpio-Muñoz and co-authors (2002a) starting from available experimental data on peptide binding affinity to class I MHC (major histocompatibility complex) molecules obtained a new method to compute the antigenic degree of peptides by using a GA, facilitating the design of clinically useful and immunologically silent peptidic drugs, as well as immunotherapeutics and vaccines for autoimmune diseases and cancer. A multilevel information processing methodology to assess peptide affinity for MHC class I molecules is proposed:

÷ The first level: relate the primary structure of peptides to their experimentally measured activity. It involve a GA combined with the profile analysis for detection of related proteins (Gribskov et al., 1987) and leads to a set of coefficients that express quantitatively the contribution of each amino acid in the peptide sequence to its binding activity. Carpio-Muñoz and co-authors (2002a) reports results for 9-mer peptides and the construction of a profile of peptide binding activity to MHC class I molecules, and suggest that the resulting profile can be used to predict binding affinity of any 9-mer sequence.

÷ The second level: analyze of the complex formed by a peptide and the receptor region of the MHC molecule, performed using a system (which uses another meta-heuristic, SA) for assessment of bio-macromolecular interaction reported elsewhere (Del Carpio-Muñoz et al., 2002b). The analysis is carried over a set of several types of MHC class I-binding peptides, and tendencies in hydrophobic correlation as well as electrostatic complementarities are ascertained that drive the binding of the molecules and formation of the complex. The tendencies derived are then applied to assess the probability of ligand-receptor binding at the atomic level.

An adaptation of the GA, Lamarckian GA were reported and used by Kadirvelraj and co-authors (2006) to dock the GBSIII tri-saccharide fragment into the antigen binding site of Fv1B1 succeeding to provide a comprehensive interpretation for a large body of biochemical and immunological data related to Ab recognition of bacterial polysaccharides and should be applicable to other Ab-carbohydrate interactions.

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).

REFERENCES

1. Lamarck, J. B. P. A., 1830, An Exposition of Zoological Philosophy (…, in French). Paris: JB Baillère, pp. 420+450.
2. Darwin, C. N. R., 1859, On the origin of species by means of natural selection. London: J Murray, pp. 459.
3. Mendel, G., 1865, Experiments in Plant Hybridization (in German), Meetings of the Brünn Natural History Society, February 8th and March 8th, Brno.
4. Fisher, R. A., 1918, The Correlation Between Relatives on the Supposition of Mendelian Inheritance, Edinb Roy Soc Trans 52:399-433.
5. Barricelli, N. A., 1954, Numerical examples of evolution processes (in Italian), Methodos 1954:45-68.
6. Fraser, A., 1957, Simulation of genetic systems by automatic digital computers, I. Introduction, Aust J Biol Sci 10:484-491.
7. Falkenauer, E., 1998, Genetic Algorithms and Grouping Problems, New York:Wiley, pp. 220.
8. Bosworth, J., F. Norman, B. P. Zeigler, 1972, Comparison of Genetic Algorithms with Conjugate Gradient Methods, NASA Contractor Reports, CR-2093.

9. Glover, F., 1977, Heuristics for Integer Programming Using Surrogate Constraints. Decision Sci 8(1):156-166.
10. Davis, L., 1987, Genetic Algorithms and Simulated Annealing. San Francisco: M Kaufmann, pp. 216.
11. Bouktir, T., L. Slimani, 2005, Optimal Power Flow of the Algerian Electrical Network using an Ant Colony Optimization Method, Leonardo J Sci 4(7):43-57.
12. Benoit, B., F. Fleurey, J. M. Jézéquel, Y. Le Traon, 2005, Automatic Test Case Optimization: A Bacteriologic Algorithm. IEEE Software 22(2):76-82.
13. De Boer, P. T., D. P. Kroese, S. Mannor, R. Y. Rubinstein, 2005, A Tutorial on the Cross-Entropy Method. Ann Oper Res 134(1):19-67.
14. Kobti, Z., R. G. Reynolds, T. Kohler, 2004, Agent-Based Modeling of Cultural Change in Swarm Using Cultural Algorithms, SwarmFest 8:8p.
15. Schwefel, H. P., 1995, Evolution and Optimum Seeking. New York: Wiley & Sons, pp. 456.
16. Fogel, L. J., 1999, Intelligence Through Simulated Evolution: Forty Years of Evolutionary Programming. New York: Wiley Interscience, pp. 162.
17. Bak, P., K. Sneppen, 1993, Punctuated equilibrium and criticality in a simple model of evolution, Phys Rev Lett 71(24):4083-4086.
18. Kjellström, G., 1991, On the Efficiency of Gaussian Adaptation. J Optim Theor Appl 71(3):589-597.
19. Banzhaf, W., P. Nordin, R. E. Keller, F. D. Francone, Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications, San Francisco: M Kaufmann Publishers, pp. 450.
20. Smith, J.E., 2007, Coevolving Memetic Algorithms: A Review and Progress Report, IEEE Trans Syst Man Cy B 37(1):6-17.
21. Davis, L., 1991, The Handbook of Genetic Algorithms. New York: VN Reinhold, pp. 385.
22. Bouskila, A., M. E. Robinson, Roitberg B. D., B. Tenhumberg, 1998, Life-history decisions under predation risk: Importance of a game perspective. Evolut Ecol 12(6):701:715.
23. Busch, H., D. Camacho-Trullio, Z. Rogon, K. Breuhahn, P. Angel, R. Eils, A. Szabowski, 2008, Gene network dynamics controlling keratinocyte migration. Molec Syst Biol 4:199(16).
24. Bolboacă, S. D., L. Jäntschi, 2007, Design of Experiments: Useful Orthogonal Arrays for Number of Experiments from 4 to 16, Entropy 9(4):198-232.
25. Diaconis, P. W., S. P. Holmes, 1998, Matchings and phylogenetic trees. Proc Natl Acad Sci USA 95(25):14600-14602.
26. Jäntschi, L., M. V. Diudea DOI, Subgraphs of Pair Vertices, J Math Chem DOI:10.1007/s10910-008-9411-6.
27. Brauer R., 1937, On algebras which are connected with the semisimple continuous groups, Ann Math 38(4), 857-872.
28. Lemmon, A. R., M. C. Milinkovitch, 2002, The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation, Proc Natl Acad Sci USA 99(16):10516-10521.
29. Huelsenbeck, J. P., F. R. Ronquist, R. Nielsen, J. P. Bollback, 2001, Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology, Science 294(5550), 2310-2314.
30. Kikuchi, S., D. Tominaga, M. Arita, K. Takahashi, M. Tomita, 2003, Dynamic modeling of genetic networks using genetic algorithm and S-system, Bioinformatics 19(5):643-650.
31. Savageau, M. A., 1976, Biochemical System Analysis: a Study of Function and Design in Molecular Biology, Reading: Addison-Wesley, pp. 199.
32. Hlavacek, W. S., M. A. Savageau, 1996, Rules for coupled expression of regulator and effector genes in inducible circuits, J Mol Biol 255(1):121-139.
33. Tominaga, D., M. Okamoto, 1998, Design of canonical model complex nonlinear dynamics, Proc Int Conf Comp Appl Biotechn 1998:85-90.
34. Tominaga, D., N. Koga, M. Okamoto, 2000, Efficient numerical optimization algorithm based on genetic algorithm for inverse problem, Proc Genet Evolut Comput Conf 2000:251-258.
35. Tsutsui S., M. Yamamura, T. Higuchi, 1999, Multi-parent recombination with simplex crossover in real coded genetic algorithms, Proc Genet Evolut Comput Conf 1999:657-664.
36. Higuchi, T., S. Tsutsui, M. Yamamura, 2000, Theoretical analysis of simplex crossover for real-coded genetic algorithms, Proc Int Conf Parall Probl Solv Nature 2000:365-374.
37. Ortiz Martınez, T., V. Rico-Gray, E. Martınez-Meyer, 2008, Predicted and verified distributions of Ateles geoffroyi and Alouatta palliata in Oaxaca, Mexico, Primates 49(3):186-194.
38. Anderson R. P., D. Lew, A. T. Peterson, 2003, Evaluating predictive models of species distributions: criteria for select optimal models, Ecol Model 162(3):211-232.

39. Brady, S. G., T. R. Schultz, B. L. Fisher, P. S. Ward, 2006, Evaluating alternative hypotheses for the early evolution and diversification of ants, Proc Natl Acad Sci 103(48):18172-18177.
40. Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. Mcgettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, 2007, Clustal W and Clustal X version 2.0, Bioinformatics 23(21):2947-2948.
41. Fink, W. L., 1986, Microcomputers and phylogenetic analysis. Science 234(4780):1135-1139.
42. Posada, D., K. A. Crandall, 1998, MODELTEST: Testing the model of DNA substitution. Bioinformatics 14(9):817-818.
43. Schultz, A. K., M. Zhang, T. Leitner, C. Kuiken, B. Korber, B. Morgenstern, M. Stanke, 2006, A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC Bioinform 7:265(15).
44. Ronquist, F., J. P. Huelsenbeck, 2003, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics 19(12):1572-1574.
45. Sanderson, M. J., 2002, Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach, Mol Biol Evol 19:101-109.
46. Sanderson, M. J., 2003, r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock, Bioinformatics 19:301-302.
47. Lewis, P. O., 1998, A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data, Mol Biol Evol 15(3):277-283.
48. Wiegmann, B. M., D. K. Yeates, J. L. Thorne, H. Kishino, 2003, Time Flies, a New Molecular Time-Scale for Brachyceran Fly Evolution Without a Clock. System Biol 52(6):745-756.
49. Yeates, D. K., 2002, Relationships of extant lower Brachycera: A quantitative synthesis of morphological characters, Zool Scripta 31(1):105-121.
50. Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, P.M. Gillevet, 1994, The genetic data environment and GUI for multiple sequence analysis, Comput Appl Biosci 10(6):671-675.
51. Maddison, D. R., W. P. Maddison ,2000, MacClade v4.0. Sunderland: Sinauer (software).
52. Maddison, W. P., D. R. Maddison, 2006, Mesquite: A modular system for evolutionary analysis v1.1. Tucson: University of Arizona (software).
53. Lewis, P.O., 2001, A likelihood approach to estimating phylogeny from discrete morphological character data, Syst Biol 50(6):913-925.
54. Del Carpio-Muñoz, C. A., T. Hennig, S. Fickel, A. Yoshimori, 2002, 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.
55. Del Carpio-Muñoz, C. A., E. Ichiishi, A. Yoshimori, T. Yoshikawa, 2002, MIAX: A new paradigm for modeling biomacromolecular interactions and complex formation in condensed phases, Prot Struct Funct Genet 48(4):696-732.
56. Gribskov, M., A. D. McLachlan, D. Eisenberg, 1987, Profile analysis: Detection of distantly related proteins, Proc Natl Acad Sci USA 84(13):4355-4358.
57. Kadirvelraj, R., J. Gonzalez - Outeirino, B. L. Foley, M. L. Beckham, H. J. Jennings, S. Foote, M. G. Ford, R. J. Woods, 2006, Understanding the bacterial polysaccharide antigenicity of Streptococcus agalactiae versus Streptococcus pneumoniae, Proc Natl Acad Sci USA 103(21):8149-8154.