Multi- Objective Genetic Algorithm for De Novo Drug Design

R. Vasundhara Devi, S. Siva Sathya, Mohane Selvaraj Coumar

***Abstract*— Genetic algorithms, can be used to solve NP-hard problems in various domains, including computer-aided drug design (CADD). As design & development of a drug molecule takes a number of man years and is also an expensive process, use of computer-aided techniques could help to reduce the time required and the cost of developing drugs. De novo drug design (DNDD) is one of the CADD technique used to design drug-like molecules virtually from smaller fragments/building blocks. This paper proposes a multi-objective genetic algorithm for the de novo design of novel molecules similar to a known reference molecule, possessing drug-like properties from a given set of input fragments and reference molecules. It could be used to design a variety of other virtual drug-like molecules by varying the input fragments and reference molecules based on the user requirement.**

***Index Terms*— computer-aided drug design, de novo drug design, multi-objective genetic algorithm.**

I. INTRODUCTION

In the drug development industry, design and synthesize of a novel drug molecule consumes lot of time and man effort. This in-turn results in high cost for developing new drugs for treating diseases. Traditionally, drug-like molecules were discovered by searching large chemical libraries for active molecules, followed by chemical modification in the laboratory to improve the activity. This process is iterative requiring several rounds of chemical synthesis and testing of molecules in the laboratory, leading to the identification of drug-like molecules suitable for testing in humans and final approval as a drug for human use [1]. Advances in computers during the past two decades, both in software and hardware, has now helped in reducing the cost and time required to develop a drug and made computer-aided drug design (CADD), a popular technique in the drug industry for drug development. In CADD, in silico experimentations are carried out to predict the potent molecules, before performing the actual laboratory experimentation, thereby decreasing the number of laboratory experimentation required to identify a potent drug-like molecule [2], [3].

De novo drug design is one of the CADD techniques, where drug-like molecules are designed from pre-existing fragments or from atoms [4]–[6]. *De novo* drug design requires optimization of several objectives simultaneously and independently to obtain a better optimal solution. Here, in the proposed system, multi-objective genetic algorithm is applied for finding optimal drug-like molecules, which are similar to a known reference molecule using *de novo* drug design.

Genetic algorithm (GA) is a randomized, stochastic technique which is suitable for a number of optimization problems [7], including the identification of a drug-like molecule from large chemical space, as in the case of de novo drug design [6].

Application of the proposed system is shown with the de novo design of drug-like molecules from a fragment library of acids and amines extracted from known drugs. The design of the molecules were guided using two objective functions, a similarity score (tanimoto similarity) to a known reference molecule (Lidocaine & Furano-pyrimidine) and an oral bioavailability score (Lipinski's Rule of 5). The proposed multi-objective de novo drug design system could be used to design drug-like molecules for variety of diseases. The rest of this paper is organized as follows: Section II describes *De novo* drug design, Section III describes the multi-objective genetic algorithm and Section IV details the proposed *De novo* drug design work along with the algorithm. Section V describes the implementation of the proposed work and section VI describes the experimental results and analysis. Finally, section VII concludes the paper.

II. DE NOVO DRUG DESIGN

The latin term "*De novo*" means "from the beginning", "afresh" or "anew". De novo drug design is used to design drug-like molecules virtually from scratch. This CADD technique can be used to effectively explore large chemical space to virtually design novel drug-like molecules. In de novo drug design, the drug-like molecules are constructed using two methods namely, atom-based method and the fragment-based method [4]–[6]. When the molecule construction is done using atom-based method, the molecules are constructed atom by atom. Even though it produces novel molecules with lot of diversity, it takes more time to arrive at the needed solution; moreover, the resultant molecules may lack synthetic feasibility. Hence, fragment-based molecule construction method is preferred in recent de novo drug design programs. In fragment-based method, the molecules are constructed from fragments instead of individual atoms. The fragments themselves can be derived from commercially available drugs. Even though the search space is less in this method, the solution obtained would be drug-like molecules, which could be easily synthesized in the laboratory.

III. MULTI-OBJECTIVE GENETIC ALGORITHM

Genetic algorithm (GA) is a stochastic, randomized search technique that mimics the Darwin's idea of natural evolution process [7].

Manuscript Received on May 2014.

Mrs. R. Vasundhara Devi, M. Tech Scholar, Department of computer Science, School of Engineering & Technology, Pondicherry University, Puducherry, India.

Dr. S. Siva Sathya, Associate Professor, Department of computer Science, School of Engineering & Technology, Pondicherry University, Puducherry, India.

Dr. Mohane Selvaraj Coumar, Assistant Professor, Centre for Bioinformatics, School of Life Sciences, Pondicherry University, Puducherry, India.

GA is the most popular evolutionary algorithm technique and the GA heuristic is routinely used to generate best optimal solutions for the NP-hard search problems. GA involves: initial population generation, fitness function evaluation, selection and breeding using genetic operators (crossover and mutation), termination condition. GA is classified into single objective and multi-objective genetic algorithm. When the GA uses single objective function to arrive at the optimal solution, it is the single objective genetic algorithm. When the GA uses multiple, sometimes conflicting objectives to arrive at the best optimal solution, it is the multi-objective GA. In multi-objective GA [8], the fitness functions are evaluated with the help of either weighted-sum approach or the pareto-ranking approach. In weighted-sum approach, each fitness function is given a weightage value and the multi-objective is converted into a single objective GA to obtain best optimal solution. In pareto-ranking approach, the entire population is ranked as per the dominance rule and then each solution is assigned a fitness value based on its rank in the population, instead of its actual objective function value.

IV. PROPOSED MULTI-OBJECTIVE GA BASED DE NOVO DRUG DESIGN

The idea of the proposed *De novo* drug design is as follows: *De novo* drug design will be carried out with the help of multi-objective genetic algorithm using a weighted-sum approach. In *De novo* drug design, the solution/molecule need not be the best one. Instead, the molecule should be the best optimal one which satisfies two design objectives: Drug-likeness and similarity to a known reference molecule. Both the objectives were evaluated for the solutions (molecules) using oral bio-availability score, as defined by Lipinski's rule of 5 [9], [10] and Tanimoto similarity coefficient [11], [12].

A. Initial population generation

The initial population consists of a set of 50 chromosomes, wherein each chromosome is represented as a vector of integers, their value representing their identity in the acid and amine fragment libraries. For the current design purpose, only two gene chromosomes are used. Each chromosome represents a possible drug-like molecule that could be synthesized from the two fragments (an acid and amine) that make up the individual genes. Here, the first gene represents the acid fragment and the second gene represent the amine fragment. The fragment library used in the design of chromosomes consists of 28 acids and 162 amines extracted from known drugs [13]. Few representative examples of acid and amine fragments used in this study and chromosome representation are shown in Figure 1. More number of genes (fragments) could also be added to the chromosome to make up a complex drug-like molecule as the need arises.

B. Objective functions and fitness evaluation

The parameters (objectives) to be optimized for deriving the drug-like molecules are oral bio-availability score (drug-likeness score) and Tanimoto similarity coefficient (similarity to a known reference molecule). Hence, these two parameters are chosen in the fitness function to evaluate the suitability of the designed molecules. The oral bio-availability score (OBA score or drug-likeness score) of the generated molecule is calculated with the help of the Lipinski's Rule of 5 [9], [10]. It states that the molecule is more likely to be orally bioavailable (drug-like) if:

- The number of Hydrogen bond donors do not exceed 5
- The number of Hydrogen bond acceptors do not exceed 10
- The molecular weight is not more than 500 Daltons
- And the octanol-water partition coefficient (LogP) value is not greater than 5

The OBA score for calculating the fitness based on the Lipinski's Rule of 5 is shown below:

- When all the 4 rules are satisfied, OBA score = 1
- When 3 rules are satisfied, OBA score = 0.75
- When 2 rules are satisfied, OBA score = 0.50
- When only 1 rule is satisfied, OBA score = 0.25
- When all the rules are violated, OBA score = 0

Tanimoto similarity coefficient is the 2D similarity with a reference molecule. It is a measure to assess chemical/structural similarity between two molecules [11], [12]. It helps to design molecules with a predefined set of properties, in this case chemical similarity to a known reference molecule, so that the newly designed molecules will have function/activity similar to that of the reference molecule. Tanimoto similarity coefficient score ranges from 0 to 1. A score of one means high similarity and zero refers to the least similarity of the designed molecule to that of the reference molecule.

In the proposed work, both oral bio-availability and Tanimoto coefficient scores are used to guide the design of drug-like molecules towards a known chemical scaffold. Hence, equal weightage for calculating the fitness of the molecule is given to the OBA score and tanimoto similarity, so that the newly designed molecule will be drug-like as well as with similar activity/function to that of the reference molecule.

Fitness score = Tanimoto similarity score + oral bioavailability (OBA) score.

C. Steps in the proposed work

The steps involved in the de novo drug design are as follows:

1. Initialize population from the acid and amine fragment library
2. Evaluate the fitness function based on weighted sum of the two objectives for the initial population

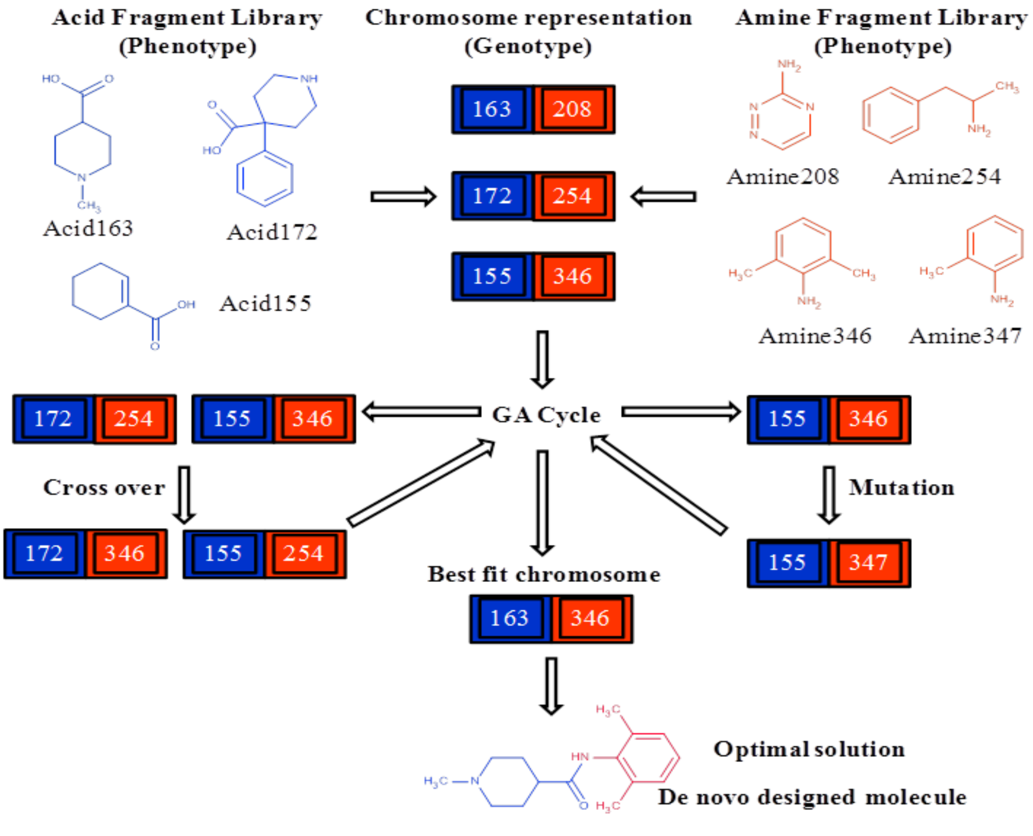

Figure 1. Overall GA workflow involved in the de novo drug design software.

3. Select individuals using tournament selection for the mating pool
4. Apply one-point crossover to the individuals from mating pool
5. Mutate individuals to increase diversity
6. Repeat the steps from 3 to 6 until termination criteria i.e till the number of generations are 50 or 100 is completed.
7. Output the virtually designed molecule with optimal fitness function objectives.

Application of crossover and mutation operators on the chromosome would lead to a number of better solutions during the GA cycle. During the GA cycle the crossover ratio was set at 85% and mutation rate at 10% for each execution. The solutions obtained were evaluated for fitness (oral bio-availability and tanimoto similarity to the reference molecule, Lidocaine and Furano-pyrimidine) to identify optimal solutions. The overall workflow of the de novo drug design program is shown in Fig. 1.

V. IMPLEMENTATION DETAILS

The object oriented programming language Java was used for the implementation. The library of 28 acid and 162 amine fragments used for evolving the new drug-like molecules were derived from known drugs and obtained from e-LEA3D: ChemInformatics Tools and Databases website (http://chemoinfo.ipmc.cnrs.fr/) [13]. They were stored in MySql database. New drug-like molecules were generated by making linear connection between the acid fragment as the first part and the amine fragment as the second part of the chromosomes and their fitness values (OBA score and tanimoto similarity) were evaluated using the Chemical development kit (CDK) library [14], [15]. Fragment library, reference molecule and newly designed molecule outputs are stored/ written in MOL2 file format [11]. The MOL2 files were read using MarvinSketch chemical structure drawing tool [16].

VI. EXPERIMENTAL RESULTS AND ANALYSIS

The ability of the de novo drug design program to generate virtual drug-like molecules were tested using two test cases. In test case - 1, the design of the molecules were based on the reference molecule Lidocaine (local anesthetic drug used by dentist) [17]. In the test case – 2, an experimental anti-cancer molecule (Furano-pyrimidine) reported in the year 2010 was used as the reference molecule [18]. In both the test cases, the initial population size was set as 50 and the GA process was run for 50 cycles. In both cases, the molecules were evolved using single objective as well as multi-objective GA, ie., using tanimoto similarity alone (objective function) and the other with a combination of OBA score and tanimoto similarity (objective function). The entire GA process was run 5 times each and the optimal solutions (drug-like molecules) obtained are shown in Fig. 2 for test case -1 and in Fig. 3 for test case - 2. The OBA score and tanimoto similarity score for the newly designed compounds were compared with the reference compounds and shown in Table I and Table II for test case - 1 and test case - 2, respectively.

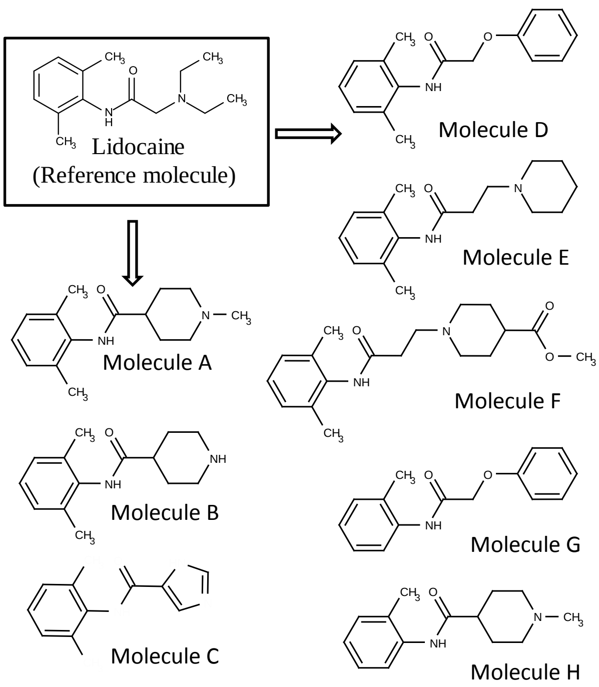

Figure 2. Reference molecule (Lidocaine, local anesthetic used by dentist) and de novo designed molecules for test case - 1.

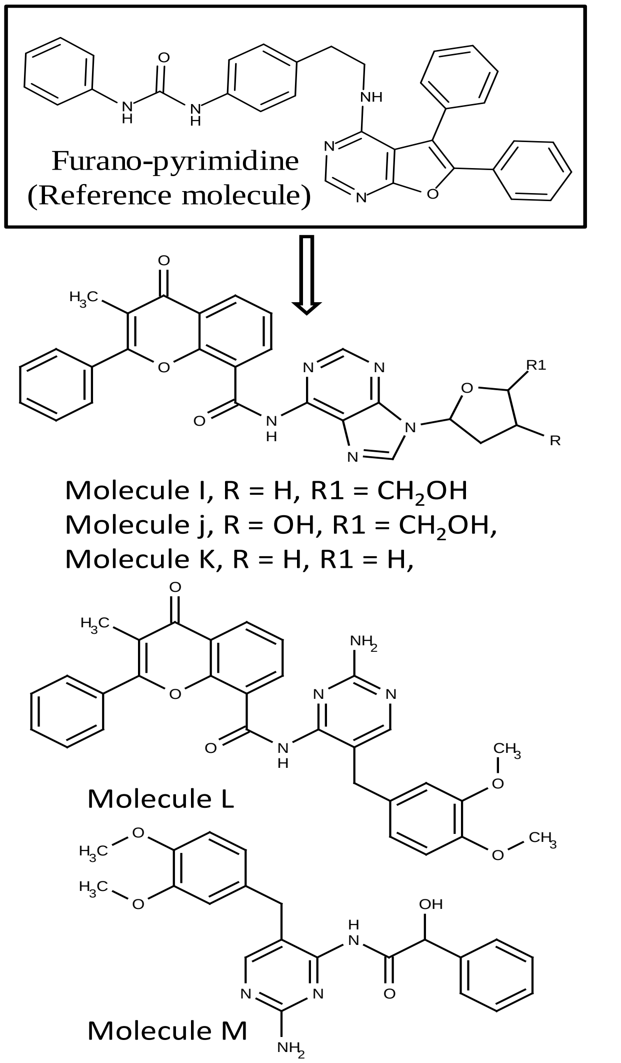
Furano-pyrimidine
(Reference molecule)
Molecule I, R = H, R1 = CH₂OH
Molecule j, R = OH, R1 = CH₂OH,
Molecule K, R = H, R1 = H,
Molecule L
Molecule M

Figure 3. Reference molecule (Furano-pyrimidine, anti-cancer molecule) and de novo designed molecules for test case - 2.

Table I. Test case - 1

Molecule	*GA Process*	*OBA score[a]*	*Tanimoto similarity score[b]*
Lidocaine	-	1.00	1.000
A	Single objective	1.00	0.573
B	Multi-objective	1.00	0.573
C	Single objective	1.00	0.534
D	multi-objective	1.00	0.533
E	multi-objective	1.00	0.550
F	Single objective	1.00	0.502
G	Single objective	1.00	0.511
H	Single and multi-objective	1.00	0.529

[a]Calculated based on Lipinski’s rule of 5 and is a measure of drug-likeness. A score of 1 means drug-like and a score of 0 means non-drug-like molecule. [b]Similarity score is a measure of structural similarity between the reference molecule (Lidocaine) and the newly designed molecule. A score of 1 means higher the similarity and a score of 0 means lower the similarity between the molecules.

Table II. Test case - 2

Molecule	*GA Process*	*OBA score[a]*	*Tanimoto similarity score[b]*
Furano-pyrimidine	-	1.00	1.000
I	Multi-objective	1.00	0.478
J	Single objective	0.75	0.481
K	Multi-objective	1.00	0.476
L	Single objective	0.75	0.497
M	Single and multi-objective	1.00	0.487

[a]Calculated based on Lipinski’s rule of 5 and is a measure of drug-likeness. A score of 1 means drug-like and a score of 0 means non-drug-like molecule. [b]Similarity score is a measure of structural similarity between the reference molecule (Furano-pyrimidine) and the newly designed molecule. A score of 1 means higher the similarity and a score of 0 means lower the similarity between the molecules.

In test case - 1, a total of 8 different de novo designed molecules were produced during the single and multi-objective runs. Analysis of all the 8 molecules using Lipinski’s OBA criteria showed that all the compounds are drug-like. However, in the test case – 2, molecules J and L generated by single objective GA process has a molecular weight greater than 500 (OBA score = 0.75), suggesting that multi-objective GA process is able to generate drug-like molecules more efficiently than single-objective GA process in the de novo design of molecules.

VII. CONCLUSION AND FUTURE IMPROVEMENTS

Discovery of a drug is a time consuming and effort intensive process, requiring huge amount of monetary investment.

Recent advances in computational methods have made the whole drug discovery process, in particular the design of a new drug more efficient and faster. De novo drug design, a computer-aided drug design (CADD) technique helps in virtually designing drug-like molecules from molecular fragments. Few of the virtually designed molecules could be synthesized and tested in the laboratory, there by alleviating the need for actual synthesis and testing of a huge number of molecules. Here we have applied multi-objective genetic algorithm to carry out de novo design of molecules which are similar to a known molecule (reference molecule), at the same time possessing drug-like characters. We have successfully designed drug-like molecules similar to Lidocaine (local anesthetic used by dentist) and Furano-pyrimidine (experimental anti-cancer compound), from a set of 28 acids and 162 amine fragments using the GA process, by applying a combination of oral bio-availability score (based on Lipinski's rule of 5) and tanimoto similarity as objective functions. The de novo program could be used to design novel drug-like molecules by varying the fragment sets and the reference molecules used in this study. In future, the de novo program could be modified to use more than two gene chromosomes for genotypic representation of the solution, as well as use more than two objective functions for the fitness function evaluation in the GA process.

ACKNOWLEDGMENT

Financial support from Science & Engineering Research Board (SR/FT/LS-64/2011), Govt. of India, to MSC is gratefully acknowledged. RVD thank Mr. Hemant Arya, Centre for Bioinformatics for his help to collect the fragment library.

REFERENCES

1. R. Ng. Drugs: From disocovery to approval. 2nd ed. New Jersey: John Wiley & Sons, Inc., 2009, pp. 1-52.
2. T.T. Talele, S.A. Khedkar, A.C. Rigby, Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Curr. Top. Med. Chem., 10, 2010, 127-141.
3. D.E. Clark, What has computer-aided molecular design ever done for drug discovery? Expert Opin. Drug Discov., 1, 2006, pp. 103-110.
4. G. Schneider, U. Fechner. Computer-based de novo design of drug-like molecules. Nature Rev. Drug Discov., 4, 2005, pp. 649-663.
5. K. Loving, I. Alberts, W. Sherman, Computational approaches for fragment-based and de novo design. Curr. Top. Med. Chem., 10, 2010, pp. 14-32.
6. C.A. Nicolaou, C. Kannas, E. Loizidou, Multi-objective optimization methods in de novo drug design. Mini Rev. Med. Chem., 12, 2012, pp. 979-987.
7. D.E. Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Boston: Addison-Wesley Longman Publishing Co., Inc, 1989.
8. K. Deb, Multi-objective optimization using Evolutionary algorithms. London: Wiley, 2001.
9. C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev., 23, 1997, pp. 3-25.
10. C. A. Lipinski. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov. Today Technol., 1, 2004, pp. 337-341.
11. N. Brown, Chemoinformatics - An introduction for Computer Scientists. ACM Computing Surveys, 41, 2009, 8.
12. T. Tanimoto. An Elementary Mathematical theory of Classification and Prediction. "IBM Internal Report," IBM technical report series, 1957.
13. E. Pihan, L. Colliandre, J.F. Guichou and D. Douguet. e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design, Bioinformatics, 28, 2012, pp. 1540-1541.
14. C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J. Chem. Inf. Comput. Sci., 43, 2003, pp. 493-500.
15. C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, E. L. Willighagen. Recent developments of the chemistry development kit (CDK) — an open-source java library for chemo- and bioinformatics. Curr. Pharm. Des., 12, 2006, pp. 2111-2120.
16. https://www.chemaxon.com/products/marvin/
17. http://www.drugbank.ca/
18. M.S. Coumar, C.Y. Chu, C.W. Lin, H.Y. Shiao, Y.L. Ho, R. Reddy, et al. Fast-forwarding hit to lead: aurora and epidermal growth factor receptor kinase inhibitor lead identification. J. Med. Chem., 53, 2010, pp. 4980-4988.

AUTHORS PROFILE

Mrs. R. Vasundhara Devi received Master's degree in Computer Applications from Presidency College, Madras University, Tamil Nadu. Currently pursuing M.Tech. Degree in Computer Science at Pondicherry University.

Dr. S. Siva Sathya is an Associate professor in the Department of Computer Science, Pondicherry University. Her areas of interests include Evolutionary Algorithms, Bioinformatics, Intrusion detection, etc. She has to her credit a number of research papers in International journals and conferences.

Dr. Mohane Selvaraj Coumar is an Assistant Professor at Centre for Bioinformatics Pondicherry University. He received his M. Pharm & Ph.D degrees in Pharmaceutical Sciences from Punjab University, Chandigarh, India. His research interests include Computer-aided drug design and Medicinal chemistry, with focus on Cancer drug discovery. He has to his credit over 40 national & international publications and 2 US patents.