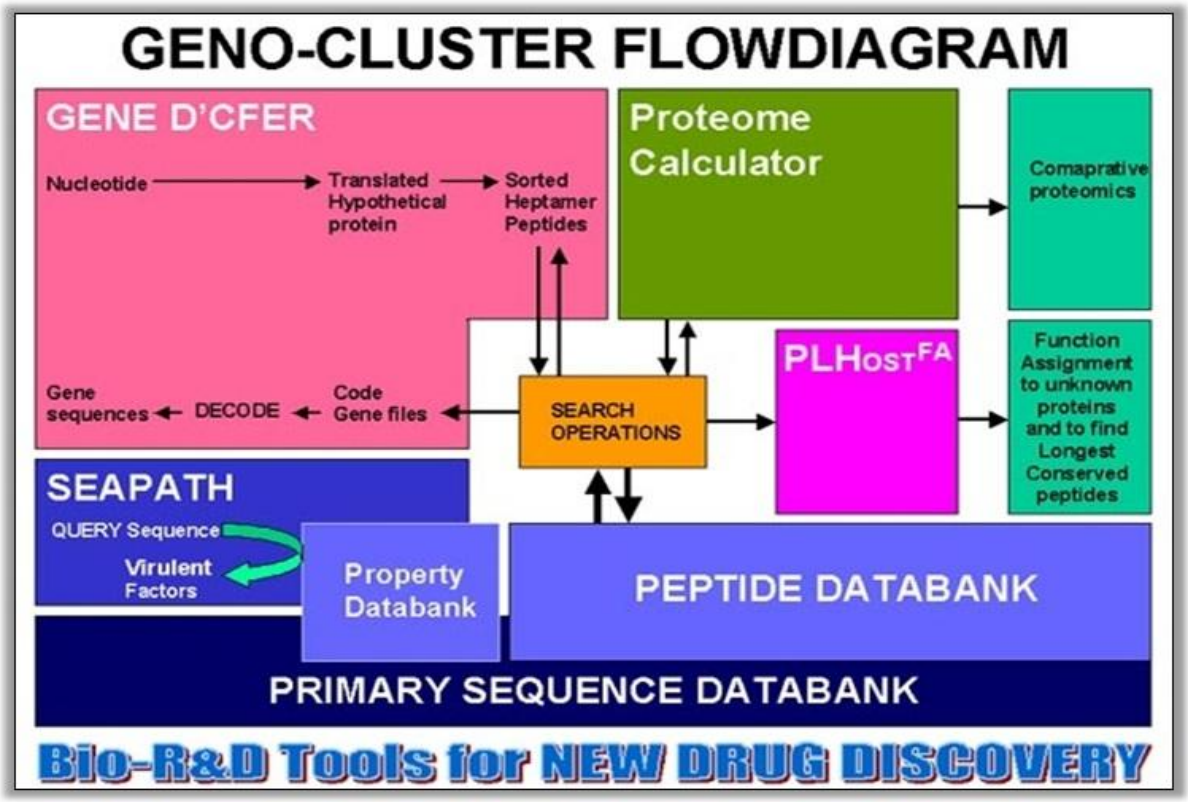
Geno-Cluster
A novel platform software tool for facilitating new drug discovery



Motivation: The availability of complete sequences of more than 280 genomes provides novel opportunities for in depth understanding of various biological phenomena through *in silico* comparative genomics. Identification of novel genes, assignment of function to gene products and their evaluation as potential drug targets is considered to be of prime importance. We have developed a suite of software programs GENE'D'CFER, PROTEOME CALKULATOR, PLHOST^{FA}, and SEAPATH and porting them into LINUX cluster to harness the enhanced computational power that aids in the prediction of prokaryotic genes, functional assignment of encoded products, identification of adhesins with the help of Artificial Neural Network based algorithms.

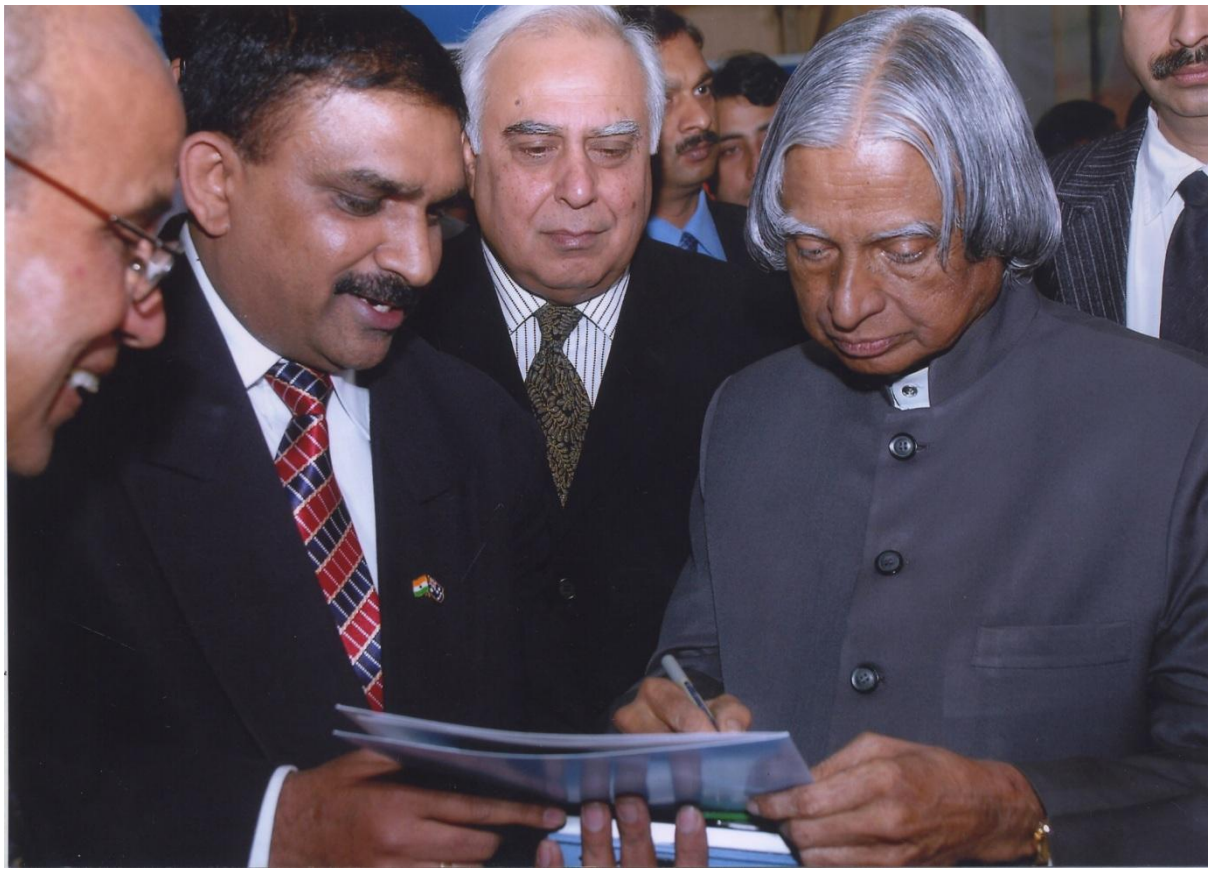
Results: We have developed a generic and versatile new approach, designated Gene'D'cfer (GDC), for prokaryotic gene identification. Unlike other existing methods, this approach employs peptides as markers for protein coding DNA sequences. GDC determines candidate genes among all possible ORFs in a given DNA sequence through the use of Artificial Neural Network (ANN) trained on a set of known peptide library. Potential ORFs are ranked according to a scoring scheme based on the abundance and distribution pattern of heptapeptides along the ORF. ORFs identified by GDC can be overlaid with other features using complementary software programs for ribosomal binding sites, promoter sequences, transcription start sites, or codon biases for further examination. An analysis of 18 completely sequenced prokaryotic genomes has been carried out to demonstrate the capabilities of GDC.

In addition, GDC has been applied on various strains of SARS virus and 4 new genes were predicted.

Delineating Conserved and Variable regions in sequences is of fundamental biological importance. Conserved regions are strong indicators for phylogenetically conserved functional roles whereas variable regions are generally implicated in auxiliary roles, often related to specific cases. The traditional approaches towards this objective involve comparing the homologous sequences using multiple sequence alignment algorithms. This approach although sound in theory is limited in terms of its speed and is not suited for high capacity. Although this limitation can be overcome in principle using powerful computers with enlarged memory, the results need careful scrutiny by the user. In most cases, users simply wish to know, in a first pass, the conserved and variable regions. PROTEOME CALKULATOR meets this need by offering a rapid approach to compare all the proteins (proteome) of a species with proteomes of other species using a peptide library approach.

Prediction of surface proteins involved in virulence from the complete sequences of proteomes of pathogens can greatly facilitate the development of ant-infectives towards eradicating infectious diseases. ANN was used to develop SEAPATH, which predicts the probability of a protein being an adhesin (Pad) based on 105 compositional properties of a sequence. SEAPATH draws upon the base algorithm SPAAN, which had optimal sensitivity of 89% and specificity of 100% and could identify 97.4% of adhesins from a wide range of bacterial pathogens causing a broad range of diseases in humans and other hosts. In the case of Severe Acute Respiratory Syndrome (SARS) associated Human corona virus, the spike glycoprotein, and nsps (nsp2, nsp5, nsp6 and nsp7) of SARS virus were identified with adhesin-like characteristics and offer new leads for rapid experimental testing.

GENO-CLUSTER Achievements



1. 283 proteins have been identified and patented by using Seapath and novel adhesins were identified in SARS and other 17 pathogenic organisms causing diseases in humans and plants.
2. 4 New SARS virus genes identified using GeneDcfer
3. 15 protein-coding regions in 18 strains of SARS-CoV were predicted.
4. Predicted new genes in bacterial genomes
5. New SARS virus gene annotation done using PLHOSTFA
6. Functional Signatures Identified for 2605 bacterial and 112 human hypothetical proteins.
7. Novel adhesins were identified in SARS and other pathogenic organisms to enable vaccine development using Reverse Vaccinology with Seapath

Why Pharmaceutical companies use these product in any of their on-going research as this will be an added advantages especially for those who are into Re-engineering Vaccines, and looking at new drug discovery or need to find drug able drug targets or for the following needs:

1. To increase the efficacy of drug based on population genetics.
2. To decrease the number of Adverse Drug Reactions (ADR).
3. For targeting only those populations capable of responding to a drug will reduce the cost and risk of clinical trials.
4. To reduce the number of medicines patients must take to find an effective therapy.
5. To revive previously failed drug targets, as they are matched with the niche population they survey.
6. To shorten the length of time patients are on medication.
7. To increase the range of possible drug targets will promote a net decrease in the cost of health care.
8. To discover potential therapies more easily using genome targets.
9. To facilitate the drug approval process, as trials are targeted for specific genetic population groups providing greater degrees of success.

Why it will be a successful story

Future: The practice of studying genetic disorders is changing from investigation of single genes in isolation to discovering cellular networks of genes, understanding complex interactions, and identifying their role in disease. As a result of this, a whole new age of individually tailored medicine will emerge. Bioinformatics will guide and help molecular biologists and clinical researchers to capitalize on the advantages brought by computational biology. On the horizon: more effective and affordable medicines, new research that leads to treatment and cures, and healthcare decisions based on a person's genes

Collaborations: between small biotech companies and larger drug development organizations, such as pharmaceutical companies, can be mutually beneficial. Under such agreements, smaller companies can gain financing to carry on with their R&D programs, while the bigger company will supplement its new drug pipeline with an innovative product.

Scientists rely on bioinformatics during every step of the drug discovery process in an effort to comprehend biological and disease mechanisms, identify new targets and to select and design novel drugs. But while methods for sequencing, measuring expression, and assessing structure have achieved high-throughput capacity via automation, the means by which data is analyzed are lagging behind.

Indian Bioinformatics Product Success story

NMITLI is the largest public-private-partnership R&D initiative of the Govt. of India. In a short span of time, the programme has several significant achievements to its credit. These include the TB molecule, herbal formulations for Psoriasis, low cost computer, weather forecast system, Bio-informatics products etc, with GENO-CLUSTER being one of them that has been developed by Institute of Genomics and Integrative Biology (IGIB) and the Council of Scientific and Industrial Research (CSIR) and now further development and hosting it is supported by Indian Centre for Social Transformation (Indian CST) a public charitable Trust. All the applications hold US patents, and have been installed in the leading academic and research institutions all over India and across the world. The software has already proved to be of tremendous use in the discovery of novel genes of the SARS virus and has several papers credited to its findings.

New training courses should be initiated by Universities to promote the idea of the advantages of using Indigenous Bioinformatics tools in bio-informatics departments across the country. These courses would include the fundamentals of Operating Systems, Parallel Programming, Hardware Design, Architectures and how to use these tools etc. Interdisciplinary research should be encouraged and students across departments should be allowed to take up such courses, thus allowing them to get the exposure of the recent emerging trends in the field and the experience of using a supercomputing environment.

These tools will help students in generating and developing new ideas and concepts, which could revolutionize the bio-informatics research. At the postgraduate level; it could be used for carrying out project work and publishing papers.

In the larger national interest Bioinformatics tools developed in India should be taught to the students by all our Universities, Colleges and Bioinformatics centers to strengthen the employability of the qualifiers and also to create manpower familiar with such tools.

GeneD'cfer

This software tool for predicting genes in Prokaryotes determines gene candidates amongst all possible ORF's of a given DNA sequence by using a peptide library and an ***Artificial Neural Network (ANN).***

Background:

Development of ***GeneD'cfer*** is based upon the observation that difference between total number of theoretically possible peptides of a given length and those which are actually observed in nature, grows drastically as this length of the peptide increases. Moreover, it is interesting to note that most of these peptides selected by nature are found only in coding regions and very rarely in theoretically translated non- coding regions. Prediction of a given ORF as a coding region/gene is based upon the number of heptapeptides present and the distribution of these heptapeptides along the ORF.

Method:

The method can be divided into five major steps:

1) Generation of a peptide library.
2) Artificial translation of a given genome into six reading frames.
3) Conversion of each translated sequences into an integer coded sequence.
4) Training of ***ANN***.
5) Deciphering genes using trained ***ANN***.

Features:

1) Powered by a database of conserved Heptapeptides across organisms.
2) Based on Artificial Neural Networks (ANN) using evolutionary principles.
3) Cross validation of proteomic information to explicate its protein coding sequences.
4) Good for both small as well as large genomes unlike HMM based methods.
5) Parallel algorithms for creating faster library.
6) Statistical interpretation.
7) Interactive Graphical user interface (GUI).
8) Customization options.
9) Flexibility to build your own peptide library.
10) Excellent circular genome result visualizer.
11) It is a follows a combinatorial approach by taking both compositional as well as database similarity into consideration.

Distinctions:

1) Four new SARS genes were discovered using this software after some customization.
2) GeneD'cfer has got a high accuracy of more than 90% on an average.
3) It has got a high level on sensitivity and specificity.
4) It is a high end quality product of the combined effort of CSIR, IGIB & now Indian Centre for Social Transformation
5) It has got a long list of licensed users that include highly esteemed institutes like IIT ,IICB etc.

PLHost

This software tool is based on invariant peptide motif signatures and assigns putative functions to unknown proteins. It is a complementary tool to blast and is an auto – annotator unlike BLAST.

Background:

The knowledge of conserved invariant peptides in a protein can be useful in assigning functions to hypothetical proteins, identifying critical amino acids, structural determinants and so on. The software ***PLHost*** and the database ***COPS*** (Comprehensive Peptide Signature) were developed to perform this task. The database provides information about function, structure and occurrence in biochemical pathways of the proteins containing these signature peptides. This database also facilitates the identification of folding nucleus / structural determinants in proteins and functional assignment to novel proteins.

Concepts & Methods:

1) ***PLHost*** is based on the novel peptide library based approach for the identification of 'functional signatures'. This approach is independent of alignment methods, which does not require any priori classification of protein functional families and hence is applicable in case of proteins with weak degrees of overall sequence similarity.

2) ***PLHost*** provides a novel method for simultaneous comparison of multiple proteomes comprising of millions of peptides and retrieves functional signatures without a prior classification of protein functional families.

Features:

1) Annotation and homology of small peptides.
2) Octapeptide library and cross validation.
3) Longest conserved peptide sequences.
4) Annotation of unknown proteins.
5) Peptides involved in active site formation.
6) Homology in invariant peptides.
7) User friendly and no usage of complicated sequence alignment tools.
8) Customizable and interactive graphical user interface (GUI).

Distinctions:

1) 69 potential antibacterial drug targets found using ***PLHost.***
2) 112 human proteins annotated.
3) 12076 invariant peptides predicted as functional signatures using ***PLHost*** to make ***COPS*** database.
4) 4 new SARS genes annotated using ***PLHost***.
5) It is a high end quality product of the combined effort of CSIR, IGIB & now Indian Centre for Social Transformation
6) It has got a long list of licensed users that include highly esteemed institutes like IIT ,IICB etc.

Proteome Calculator

Comparative Proteomics play a vital role in analyzing protein sequence of various organisms. It helps in understanding the disease process, develop new biomarkers for diagnosis and accelerate drug development.

Background:

Delineating conserved and variable regions in sequence is of fundamental biological importance. Conserved regions are strong indicators for phylogenetically conserved functional roles whereas variable regions are generally implicated in auxiliary roles, often related to specific cases. The traditional approaches towards this objective involve comparing the homologous sequences using multiple sequence alignment algorithms. In the real time cases, researchers simply wish to know, in a first pass, the conserved and variable regions. ***Proteome Calculator*** meets this need by offering a rapid approach to compare all the proteins (Proteome) of given species with proteomes of the other species using a novel peptide library approach.

Method:

1) ***Proteome Calculator*** is a powerful computational tool to study several proteomes at one go by performing set theory operations like union, intersection, difference and inverse. These operations would help- in identifying the most unique, conserved and clustered regions of proteins across species, which enable us to formulate a specific drug target in pharmaceutical industry.
2) The characteristic feature of the tool is that it carries out multiple analysis on a wide range of bacterial strain. It performs a screening on the pathogenic organisms, narrowing down to a unique disease condition. Its efficient backstitching operation fishes out specific protein functions and domains.

Features:

1) Alphabetically indexed peptide library.
2) Unique set theory operations applied to comparative proteomics.
3) Extensive data mining options.
4) Proteomics comparison in wide spectrum of organisms.
5) Gives high confidence level for invariant peptide by giving its total occurrence in proteins and organisms.
6) Search options based on peptide, occurrence and both in query results.
7) Stitch module and multiple analysis.
8) User friendly and Interactive graphical user interface (GUI).
9) Customizable.

Distinctions:

1) Less computational and accurate.
2) It is a high end quality product of the combined effort of CSIR, IGIB & now Indian Centre for Social Transformation
3) It has got a long list of licensed users that include highly esteemed institutes like IIT ,IICB etc.

Seapath

Prediction of surface proteins involved in virulence from the complete sequences of proteomes of pathogens can greatly facilitate the development of anti-infectives towards eradicating infectious diseases.

Background:

The virulent organisms posses adhesin proteins which bind to the host and resist any defense mechanisms. These proteins help in identifying potential targets (bacterial and viral surface antigens / adhesin) for new vaccine formulations and developing therapeutics against pathogens. The conventional methods for identifying adhesin proteins are time consuming and demand large resources. It uses decisive parameters to assess whether a protein is an adhesin. The software not only identifies the known adhesins but also helps the researcher in narrowing down their search and thus enhances the accuracy percentage in annotations of proteins as adhesins.

Method:

1) ***Seapath*** is based on the base algorithm ***SPAAN*** .
2) The underlying architecture of this tool is based on ***Artificial Neural networks (ANN).***
3) It takes ***105 compositional properties*** of a sequence in consideration so as to predict the adhesins.

Features:

1) Non homology based method.
2) Analysis based on 5 parameters.
3) Only software available for Adhesin prediction.
4) Optimal sensitivity of 89% and specificity of 100% on a defined test set.
5) Adhesin prediction accuracy for known adhesins of 97.4% for wide range of bacteria.
6) Tabulates individual contributions of the parameters.
7) Choice of parameters.
8) User friendly and Interactive graphical user interface (GUI).
9) Customizable.
10) Primarily used for drug discovery.

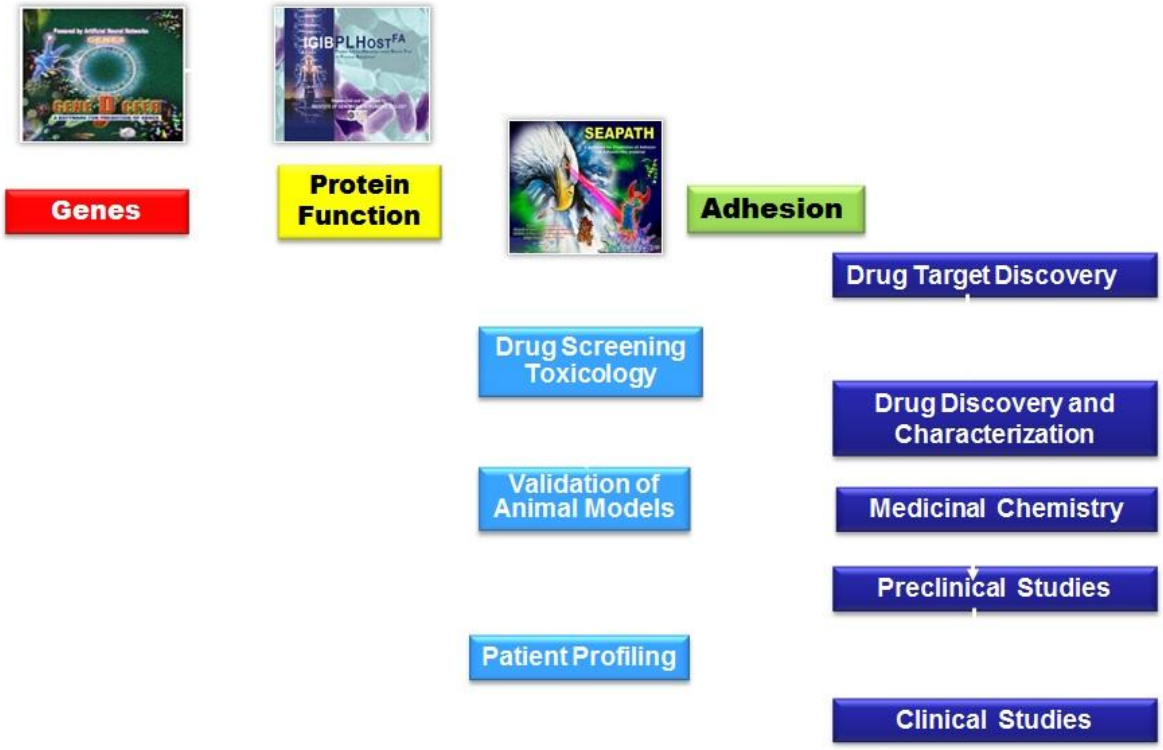
Distinctions:

1) Novel adhesins were identified for different pathogens.
2) In case of SARS associated Human corona virus, the spike glycoprotein, and nsps of SARS virus were identified after some customization.
3) Only counterpart to wet lab in such type of analysis.
4) It is a high end quality product of the combined effort of CSIR, IGIB & now Indian Centre for Social Transformation
4) It has got a long list of licensed users that include highly esteemed institutes like IIT ,IICB etc.

http://www.pnas.org/content/105/14/5555.abstract

INSTITUTES WHO HAVE PURCHASED THE GENO-CLUSTER PRODUCT	
1	INDIAN INSTITUTE OF CHEMICAL BIOLOGY
2	Dr.Naidu's Global Academy
3	Indian Veterinary Research Institute (IVRI)
4	University Of Pune
5	MADURAI KAMARAJ UNIVERSITY
6	Rajiv Gandhi Centre for Biotechnology
7	Fisheries College
8	Vidya Pratishthan's (Baramati)
9	Indian Institute of Technology Madras
10	Amrita Vishwa Vidyapeetham
11	RIKEN - Genome Science Center -Japan
12	DOEACC Kolkata
13	SRI RAMACHANDRA MEDICAL COLLEGE AND RESEARCH INSTITUTE
14	Holy Cross College
15	UNION CHRISTIAN COLLEGE
16	Banasthali Vidyapith
17	TNAU, Coimbatore
18	Bharathidasan University
19	National Institute of Technology
20	IBSD (Imphal)
21	BARC
22	CIMAP
23	West Bengal University of Technology
24	Lyallpur Khalsa College
25	National Jalma Institute
26	University of Kerala
27	University of Allahabad
28	NIPER

GENO-CLUSTER™ in Drug Discovery process



For more details and real time solutions demo experience visit www.indiancst.in

http://genocluster.indiancst.com/login/

Shri. RAJA SEEVAN
Founder Trustee
rajaseevan@indiancst.in, rajaseevan@gmail.com
Mobile: +91 9739047849,
Helpdesk Number: +91 9742979111

Indian Centre for Social Transformation
A Public Charitable Trust
"Grace Mansion," #25 Infantry Road ,
Bengaluru- 5600 01,
Karnataka -India,
Tel: +91 8049389696