BioSuite: A comprehensive bioinformatics software package (A Unique Industry-Academia Collaboration)

The NMITLI-BioSuite Team: *Tata Consultancy Services:* Vidyasagar M, Mande S, Rajgopal S, Gopalkrishnan B, Srinivas STPT, Uma Maheswara Rao C, Kathiravan T, Mastanarao K, Narendranath S, Rohini S, Irshad A, Murali T, Subrahmanyam C, Mona T, Sankha S, Priya V, Suman D, Raja Rao, VV, Nageswara Rao P, Issaac R, Yashodeep H, Arundhoti B, Nishant G, Jignesh S, Chaitanya KS, Prasad Reddy SPV; *Bose Institute:* Chakraborty P*;* *Centre for DNA Fingerprinting and Diagnosis*: Hasnain SE, Mande S, Nagarajaram A, Ranjan A, Acharya MS, Anwaruddin M, Arun SK, Gyanrajkumar, Kumar D, Priya S, Ranjan S, Reddi BR, Seshadri J, Sravan Kumar P, Swaminathan S, Umadevi P, Vindal V, Vijaykrishnan S; *Central Drug Research Institute*: Saxena AK, Dixit A, Prathipati P, Kashaw SK; *Indian Institute of Chemical Biology*: Mandal C, Bag S; *Indian Institute of Science:* Balakrishnan N, Bansal M, Chandra NR*, Murthy MRN, Ramakumar S, Sekar K, Srinivasan N, Suguna K, Vishveshwara S*, Anandhi, Bhadra R, Das S, Hansia P, Hariharaputran S, Jeyakani J, Karthikeyan R, Pandey RK, Swamy CS, Vasanthakumar B; *Indian Institute of Technology Bombay:* Balaji PV, Patel RY; *Indian Institute of Technology, Delhi:* Jayaram B, Shaikh SA; *Indian Institute of Technology, Kharagpur*: Chakrabarti PP, Banerjee A, Chakrabarti A; *Indian Statistical Institute*: Karandikar RL (Delhi) and Chaudhuri P (Kolkata); *Institute of Microbial Technology* : Raghava GPS, Ghosh A; *Institute of Bioinformatics and Applied Biotechnology*: Bansal M, Paramsivam N; *Institute of Genomics and Integrative Biology*: Brahmachari SK, Dash D, Balasubramaniam C, Basu A, Biswas P, Hariharan M, Mathur R, Sandhu KS, Scaria V, Shankar R; *International Institute of Information Technology*: Narayanan PJ, Jain V, Nirnimesh; *Madurai Kamaraj University*: Krishnaswamy S,. Alaguraj V, Marikkannu R, Mohan Katta AVSK, Krishnan N, Srividhya KV, Eswari PJ; *National Institute of Pharmaceutical Education and Research*: Bharatam PV, Iqbal P; *Saha Institute of Nuclear Physics*: Bhattacharyya D; *University of Hyderabad:* Desiraju GR, Kumar JJ, Ravikumar M; *University of Madras*: Gautham N, Prasad PA, Bharanidharan D

*** Corresponding authors**

Dr. Nagasuma Chandra	Prof. Saraswathi Vishveshwara
Bioinformatics Centre	Molecular Biophysics Unit
Indian Institute of Science	Indian Institute of Science
Bangalore 560 012	Bangalore 560 012
Tel: +91-80-23601409	+91-80-22932611
Fax: +91-80-23600551	+91-80-23600535
E-mail:nchandra@physics.iisc.ernet.in	sv@mbu.iisc.ernet.in

Abstract

The last decade has witnessed an exponential growth of information in the field of biological macromolecules such as proteins and nucleic acids and their interactions with other molecules. Computational analysis and predictions based on such information are increasingly becoming an essential and integral part of modern biology. With rapid advances in the area, there is a growing need to develop versatile bioinformatics software packages, which are efficient and incorporate the latest developments in this field. In view of this, the Council of Scientific and Industrial Research, (CSIR) India, undertook an initiative to promote a unique Industry-Academia collaboration, to develop a comprehensive bioinformatics software package, under its New Millenium Initiative for Technology Leadership in India (NMITLI) programme. BioSuite, a product of that effort, has been developed by Tata Consultancy Services who took the primary coding responsibility with significant backing from a large academic community who participated on advisory roles through the project period.

BioSuite integrates the functions of macromolecular sequence and structural analysis, chemoinformatics and algorithms for aiding drug discovery. The suite organized into four major modules, contains 79 different programs, making it one of the few comprehensive suites that caters to a major part of the spectrum of bioinformatics applications. The four major modules, (a) Genome and Proteome Sequence analysis, (b) 3D modeling and structural analysis, (c) Molecular dynamics simulations and (d) Drug design, are made available through a convenient graphics-user interface along with adequate documentation and tutorials. The unique partnership with academia has also ensured that the best available methodology has been adopted for each of the 79 programs, which has been thoroughly evaluated in several stages, leading to high scientific value of the suite. The codes have been written by the TCS team for every individual program with strict adherence to CMMi Level 5 quality processes, all within a record time of 18 months. The software, apart from having the advantage of running on a Linux platform on a personal computer, is also flexible, modular, and allows for newer algorithms to be plugged into the overall framework. The package will be valuable for high quality academic research, industrial research and development and for teaching purposes, both locally within the country as well as in the international arena.

1. Background

Genesis of BioSuite: Council of Scientific and Industrial Research, Government of India, (CSIR), proposed a new millenium initiative, in 2000, where in India could acquire leadership positions in key technology areas (NMITLI). Development of versatile, portable bioinformatics software was recognized as one such area, taking into account the expertise available in the Indian academic community. Such a project, promoted by CSIR, was therefore flagged off in partnership with the industry, where Tata Consultancy Services (TCS) took the major responsibility of developing the software with significant scientific support from the major academic institutions in the country. The objectives of the project have been to develop indigenously, a set of software tools, that would assist the academic research, R&D and applications in industry, in the rapidly emerging field of bioinformatics and rational drug design.

Algorithm design, Code writing Coding Quality checks, Graphic-user interfaces & performance benchmarking	Tata Consultancy Services, team (individual names on the first page) led by Drs. Vidyasagar M, Sharmila Mande and Rajagopal Srinivasan
Algorithm/Module design suggestions & Scientific evaluations	Academic partners (individual and institution names on the first page)
Project Monitoring committee	Profs. Narasimha R, Padmanaban G Desiraju GR, Balasubramanian D
Project co-ordination	Drs. Yogeswara Rao and Vibha Sawhney, CSIR
Project funding	CSIR, NMITLI Scheme, Govt. of India
Manuscript preparation	Coordinated by Dr. Nagasuma Chandra & Prof. Saraswathi Vishveshwara , IISc

Box-1: Roles played by different groups for ensuring successful development of BioSuite

The need for such a software suite is exemplified by two main factors: (a) increase in bioinformatics activities at all levels - education, research, industry, rapid growth of primary data and methods in computational biology and (b) limitations of existing suites- such as very high cost and not being comprehensive under a single framework, as discussed later. A team of 35 members from TCS worked on this project.

Mode of operation

To ensure the smooth functioning of the project, the following management structure was put in place: (a) *A Monitoring Committee*, monitored the progress of the project through periodic meetings with TCS and the academic partners providing timely focus, (b) *A Steering Committee*, consisting of scientists from academic institutions and TCS, coordinated the activities of the group of (c) *Domain experts and consultants*, consisting of all academic partners, helped in arriving at a basic structure for the suite. Given the large size of the group and the involvement of 18 institutions, the efforts from CSIR and the monitoring committees have played a significant role in fostering the unique partnership to ensure success of this project. The domain experts have advised TCS on the individual modules and individual programs required in each module, identified appropriate algorithms at each step, as also the features required for each program, as per the current research trends and requirements. Further, (d) *a team of pseudo-code developers of 6 people at* TCS, have interacted with domain experts and directed their (e) in-house *team of code developers*, consisting of 27 software engineers, who have written the actual code. The (f) *Software Project Management Committee* from TCS has ensured the overall activities at that end and ensured appropriate benchmarking and in-house quality checks from the software perspective. The scientific performance of the codes developed have been further evaluated by the academic partners, who have tested and reported bugs to Project Management Committee, after which codes have been improved/modified where required. Further, an autonomous assessment of the suite has been obtained by an independent expert in the area.

Operational schedules

A glimpse of the schedules and the various milestones reached are given below: (a) Identification of the modules, the required programs in each module and the appropriate algorithm(s) for each program, was completed in the first 4 months, following which a (b) Software Requirement Specification (SRS) document was developed and reviewed in the next 2 months. Next, the pseudo-codes were developed in about 5 months and converted into final code in the next 12 months. In parallel with alpha-testing that was carried out simultaneously with code development, the documentation, creation of a user guide took about 7 months. Bug reporting and bug fixes were carried out in iterations through the testing phases and a beta-version was produced by June 2004, taking a total of 24 months. Evaluation and bug fixing of this version was carried out in 5 months, leading to the first full version, soft-launched in July 2004 and product released in December 2004.

2. Overview of the organization of the suite

The entire package, consisting of 79 different programs is organized into four major modules, all linked through three common graphics-user interface workbenches, as illustrated in Figure 1. The four modules are: (a) Genome and Sequence analysis, (b) 3D Modeling and Structure analysis, (c) Molecular dynamics simulations and (d) Drug design. They are accessible through central GUIs for file handling, sequence and structure windows.

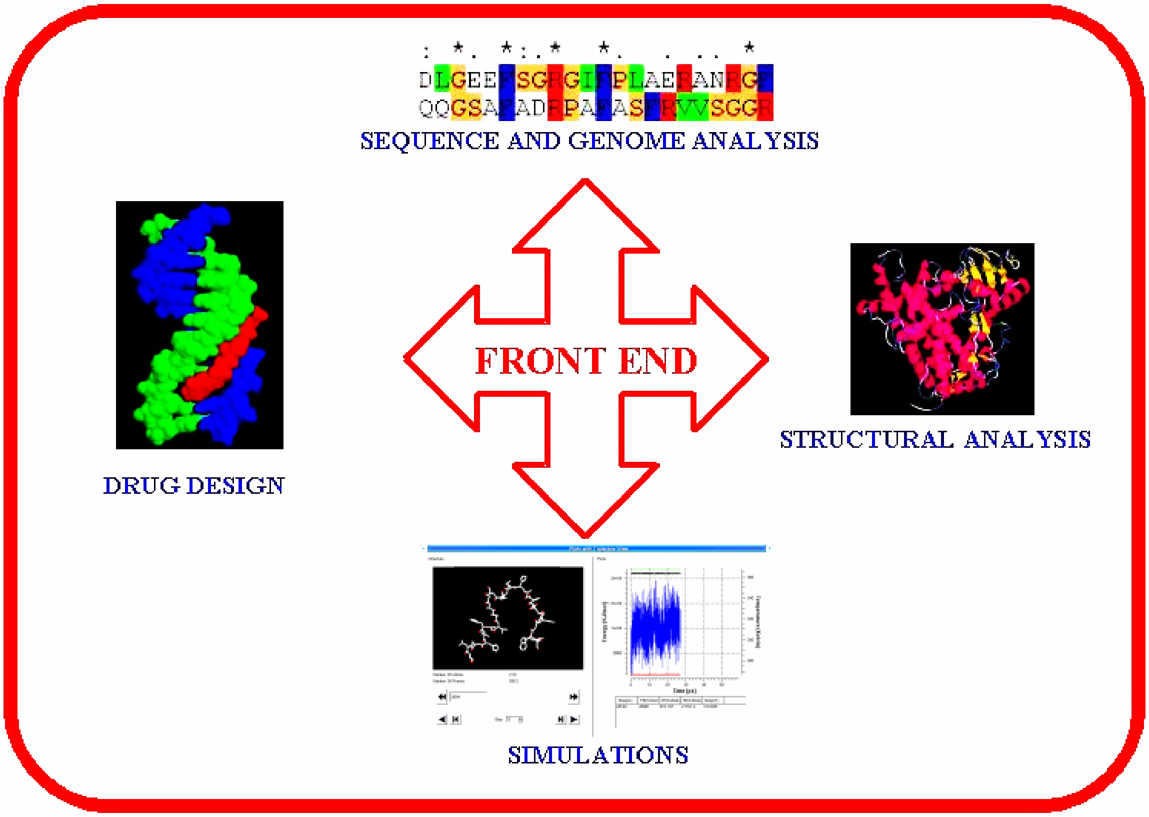


Figure 1: The modular organization of BioSuite

Table-1 lists the programs in each module. Combination of the four modules makes BioSuite a comprehensive package, covering much of the activities of the bioinformatics spectrum, starting from genome sequences to individual and multiple protein sequences, different levels of structure prediction, analysis of the structures, molecular mechanics calculations, molecular dynamics simulations, cahemoinformatics and finally integration with the application of the sequence and structural analyses in rational drug design through algorithms for QSAR, pharmacophore identification and docking processes, for facilitating rational drug design.

SlNo	**Name of the program**	**Algorithm/reference**	**Description**
1.	Blast	Altschul *et al.,* 1990	Search Tool for Finding Locally Optimal regions from sequences in a database
2.	PsiBlast	Altschul *et al.,* 1997	Search Tool for Finding Locally Optimal regions from sequences in a database
3.	Local Alignment	Gotoh, 1982	Finds an optimal local alignment of a pair of sequences using a dynamic programming method
4.	Global Alignment	Needleman & Wunsch, 1970	Finds an optimal global alignment of a pair of sequences using a dynamic programming method
5.	Dot Plot	Simple string matching	Aligns two sequences and displays the output as dotplots.
6.	Multiple Alignment	Thompson *et al.,* 1994	Aligns multiple sequences.
7.	Composition	Simple String matching	Finds the composition of nucleotide(s)/amino acid(s)/N-mers in a DNA or protein sequence
8.	Word Search	Simple String matching	Identifies the locations where a user given pattern is found.
9.	Restriction Site Analysis	Knuth-Morris-Pratt's pattern-matching algorithm, 1977	Identifies the locations where specific restriction enzyme(s) will cut a given DNA sequence
10.	Repeat Analysis	Landau *et al.,* 2001	Scans a DNA/protein sequence for potential tandem repeats up to a specified size
11.	Inverted Repeats	Naïve string matching algorithm.	Finds the hairpin structures and single strand inverted repeats in a given DNA sequence.
12.	DNA Structure Motifs	String Matching Algorithm	Finds Cruciform DNA, Z-DNA form, Triplex helical sites and potential quadruplex structural sites in a given DNA sequence
13.	Protein Secondary Structure	Chou-Fasman (Chou-Fasman, 1978) GOR1 (Garnier *et al.,* 1978) GORIII (Gibrat *et al.,* 1987) GORIV (Garnier *et al.,*	Predicts the protein secondary structural elements in a given protein sequence

		1996) NeuralNetwork (Jones, 1999)	
14.	Transmembrane Region	TMAP algorithm (Persson and Argos, 1997); DAS algorithm (Cserzo *et al*., 1997); β-strand prediction Gromiha *et al*., 1997	Predicts likely transmembrane, alpha-helical and beta barrel regions in a protein
15.	RNA Secondary Structure	Jaeger *et al*., 1989; Zuker *et al*., 1999	Predicts RNA secondary structures for a given RNA sequence.
16.	Antigen Binding Site	Kolaskar and Tangaonkar, 1990	Predicts the potential antigenic regions with in a protein sequence
17.	Peptide Map	Simple Pattern Matching	Finds the potential cleavage sites of proteolytic enzyme or reagent
18.	Property Profile		Plots the properties of a protein sequence over the length of the sequence. There are 32 protein properties that can be plotted in BioSuite.
19.	Isoelectric Point		Isoelectric Point enables to calculate the isoelectric point and plots the pH versus charge for a given protein sequence

20.	Domain Build	Rabiner, L. R. 1989;
		Durbin *et al*., 1998;
		Eddy, 1998.
			Creates a Profile Hidden Markov model (HMM) from a set of multiple aligned nucleotide or protein sequences

21.	Domain Calibrate	Rabiner, L. R. 1989;
		Durbin *et al*., 1998;
		Eddy, 1998.
			Calibrates the HMM

22.	Domain Search	Rabiner, L. R. 1989;
		Durbin *et al*., 1998;
		Eddy, 1998.
			Searches sequence database to align with profile HMM

23.	Profile Search	Rabiner, L. R. 1989;
		Durbin *et al*., 1998;
		Eddy, 1998.
			Searches pfam database to find domains that are present in the query

24.	Motif Build	Expectation Maximization algorithm (Bailey and Elkan, 1995)	Finds conserved motifs in a group of unaligned sequences
25.	Motif Search	QFAST algorithm by Bailey and Gribskov, 1998	Searches sequence database to align with motif model
26.	Helix-Turn-Helix	Dodd & Egan, 1990	Finds the Helix-Turn-Helix motifs with in a protein sequence.
27.	Coiled Coil Prediction	Lupas *et al.*, 1991	finds the coiled coil motifs in a selected protein sequence
28.	Evolutionary Distance Estimation DNA Distance Estimation	Uncorrected Distance or P Distance Jukes-Cantor Distance for DNA distance (Jukes and Cantor, 1969); Tajima-Nei Distance (Tajima and Nei, 1984); Kimura Two-Parameter Distance (Kimura, 1980); Tamura Distance (Tamura, 1992) Tamura-Nei distance	Estimates pairwise evolutionary distances between nucleotide or protein sequences using different approaches of distance correction measures.

	Protein Distances	(Tamura and Nei, 1993); Felsenstein F81 Distance (Felsenstein, 1981) logDet Distance (Barry and Hartigan, 1987) Simple Distance or p distance Similarity Jukes Cantor Protein Distance(Jukes and Cantor, 1969) Poisson Distance KimuraProtein Distance (Kimur, 1983) PAM (Dayhoff, 1978) BLOSUM (Henikoff and Henikoff, 1992) JTT (Jones *et al*., 1992	
29.	Tree Construct	UPGMA (Sneath and Sokal, 1973) WPGMA (Sneath and Sokal, 1973) Neighbor-Joining *(*Saitou and Nei, 1987) Fitch Margoliash (Fitch, W. M. and Margoliash,. 1967).	Constructs a tree based on distances estimated from sequence dissimilarities
30.	Maximum Parsimony Assessing tree reliability	Swofford, 1993, Swofford et al., 1996., Fitsch, 1971. Bootstrapping (Felsenstein, 1985) Jackknifing Consensus (Swofford, 1993*)*	Constructs evolutionary trees for nucleotides or protein sequences using maximum parsimony as the tree construction approach. Associates a reliability estimate value to every node in the constructed tree
31.	Translate		Converts a given DNA sequence into the corresponding protein sequence in all the six frames or any user specified frame
32.	Back Translate		Converts a given protein sequence into the corresponding DNA sequence.
33.	DNA to RNA		Converts DNA sequences into RNA sequences
34.	RNA to DNA		Converts RNA sequences into DNA sequences
35.	Primer Design	Nearest Neighborhood Thermodynamic Method for T_m estimation by SantaLucia, 1996.	Designs both forward and reverse primers for a given DNA sequence
36.	Probe Design	Nearest Neighborhood Thermodynamic Method for T_m estimation by SantaLucia, 1996.	Designs probes for a given DNA sequence
37.	Vector Trimming	String Matching Algorithm (BLAST) Altschul *et al.*, 1990	Finds matching regions with in a given string from a database of vectors.

38.	Contig Assembly	Huang, 1992	Converts consensus sequence from a set of contigs
39.	EST Mapping	Modified Smith Waterman algorithm.	Map a given EST to a specific location in the genome sequence.
40.	ePCR	Schuler, G. D. 1997	Finds Sequence Tagged sites (STSs) in a given DNA sequence
41.	ORF Prediction	a)Frequency based method (Fickett) . Inhomogeneous Markov model (Borodovsky *et al*.) Interpolated Markov model (Delcher *et al*)	Locates the putative coding regions in a given prokaryotic genome sequence.
42.	Intron Exon Boundary	Logitlinear model (Kleffe *et al.*, 1996)	Locates the putative junction regions between introns and exons in a given eukaryotic DNA sequence.
43.	Whole Genome Alignment	Suffix Trees (Arthur *et al.*, 2003)	Aligns two similar genomes
44.	Orthologs	Tatusov *et al*., 2000	Assigns given protein sequence to existing Clusters of orthologous genes
45.	Unique Gene	Enright *et al*. 2000	Finds unique genes between two genomes
46.	Fused Protein	Enright *et al*. 1999	Finds fused proteins in one genome wrt the other
47.	Phylogenetic Profile	Marcotte *et al*., 1999	Finds evolutionary profiles of a given protein sequence
48.	Gene Order	Mazumder *et al.,* 2001	Finds the order of genes between two genomes
49.	Format Converter		Converts one sequence file format to another sequence file format
50.	PDB to FASTA		Extracts sequence information from PDB files and writes sequence as FASTA format.
51.	Sequence Randomizer		Randomizes given sequences
52.	Simplify		Reduces the size of the alphabet.
53.	Genetic Code Editor		Edits and Saves genetic code
54.	Codon Usage Editor		Edits and Saves Codon code
55.	Fold Classification	SSAP(Sequential Structure Alignment Program) (Orengo *et al* (1996))	Detects the 3-D fold for the three - dimensional structure of a protein
56.	Interactions	Baker *et al*. (1984). McGaughey *et al*. (1998)	Checks for Van der Waals, hydrophobic, hydrogen, saltbridge, aromatic – aromatic and amino – aromatic interactions
57.	Nucleic Acid Analysis	Bansal *et al* (1995)	Evaluates the stereochemical properties of a nucleic acid structure
58.	Binding Site Detection PASS and Evolutionary Trace	PASS(Putative Active Sites with Spheres) (Brady *et al* (2000))	Identifies probable active sites
59.	Quality Check	Laskowski *et al* (1993)	Checks for geometric, stereochemical correctness of a molecule
60.	Structural Superposition	Sutcliffe *et al* (1987)	Performs Structural Superposition for given set of molecules. Superposes multiple set of molecules based on the equivalences

61.	Symmetry	Rossmann and Arnold	Generates symmetry related molecules based on space group
62.	Threading	Contact based, 3D – 1D, Consensus Bowie *et al* (1991). Zhang *et al* (2000)	Prediction of three-dimensional fold of a protein
63.	Solvent Accessible Area and Volume	Shrake & Rupley (1973) Lee and Richards (1971)	Calculates the solvent accessible area and volume of a molecule using numerical calculations
64.	Molecular Surface Area and Volume	Conolly (1985)	Uses an analytical method to compute molecular surface area and volume.
65.	Homology Modeling		Builds a three-dimensional model of a protein from its sequence based on the structure of homologous proteins
66.	Loop Modeling	McLachlan A. D (1982)	Identifies the loops in a molecule
67.	Side chain Modeling	Dunbrack, Jr. and M. Karplus (1993)	Identifies the side chains for the molecule
68.	Create and Edit molecules		Creates small molecules and biological macro molecules, provides various editing options such as adding hydrogens, geometric transformations
69.	Binding Site detection using Evolutionary Trace	Evolutionary trace (Brady *et al* (2000) and Lichtarge *et al* (1996))	The Evolutionary Trace is a novel predictive technique that identifies active sites and functional interfaces in proteins with known structure.
70.	Energy Minimization	Steepest Descents Minimizer Conjugate Gradient Minimizer (Gilbert *et al* (1992), Watowich *et al* (1988)) Polak-Ribiere Plus CG Method (Polak (1969)) Shanno's CG method (Shanno (1978)) More *et al* (1994) Forcefields: Weiner *et al* (1984)	Minimizes the energy of the molecule
71.	Electrostatics	Bottcher(1973), Debye *et al* (1923), Fogolari *et al* (1999),Jayaram *et al* (1989), Klapper *et al* (1986), Nicholls *et al* (1991))	Computes electrostatic potential using Poisson and Boltzman equation for molecules
72.	Molecular Dynamics	Integrator : Velocity Verlet, Leapfrog Constraints : Shake, Rattle Temperature Control (Andersen 1980), Andersen(1983), Berendsen *et al* (1984)), Pressure Control (Berendsen *et al* (1984)) Periodic Boundary : Minimum Image	Simulates the dynamic behavior of molecular system under various conditions
73.	Molecular Dynamics	RMSD	Analyzes trajectories obtained from

	Analysis	Standard Deviation Average Position Plots of system properties.	MD runs
74.	PBE Analysis	Surface Potential display Contours	Analyzes electrostatic potential maps
75.	Conformation Search	Random, Systematic Simulated AnnealingJonathan M. Goodmann (1998)	Explores Conformation space of a molecule
76.	Docking	Simulated Annealing (Goodsell *et al* (1990)) Genetic algorithms (Morris *et al* (1998))	Finds favourable binding configurations between a flexible ligand and a macromolecular target (usually a protein molecule)
77.	QSAR	Over 80 descriptors. Regression analysis	Computes structure activity relationship
78.	Alignment	Steric & Electrostatic algorithms (Good *et al* (1992)) RMSD calculation algorithm (Jones *et al* (1995) Genetic Algorithms (Morris *et al* (1998))	Calculates the molecular similarity of a group of molecules with reference to a template molecule.
79.	Pharmacophore Identification	3D structure similarity searches (Kurogi *et al* (2001)) Identification of features (donors/acceptors/rings) (Jones *et al* (1995))	Determines pharmacophore in a set of structures
80.	Database Generation & Search		Creates and searches through database of molecules
	Structure Viewer		Interactively view/ manipulate structures in 3 – dimensions in variety of renderings
	Sequence Viewer		Interactively view/edit sequences and alignments

3. Choice of algorithms and coding methods

Choice of algorithms was discussed extensively with academic partners and the latest concepts available in literature have been adopted wherever possible. For some programs, more than one algorithm has also been implemented, to suit the current research trends of using multiple methods and studying consensus predictions. In general, about two scientists have analyzed and chosen a particular algorithm for a particular purpose. Table-1 indicates the algorithms chosen for each of the programs. The knowledge and description of each of the algorithms have been captured into detailed

SRS documents by the pseudo-code development team at TCS through extensive interactions with the academic partners as well as with a detailed study of the appropriate literature. The pseudo-code generated for each algorithm and its linkages have been developed using formal software engineering methods, so as to guarantee correctness. The pseudo-code was then converted into actual code by another set of programmers who have ensured strict adherence to well-established quality processes such as CMMi Level 5.

All codes have been written in C++. A total of 170 algorithms and about 100 QSAR descriptor calculators, have been implemented in 79 programs, with about 700,000 lines of code. The suite is modular, which not only facilitates seamless updation of the modules but also enables integration of new programs by the end users.

4. Description of the modules

The functionalities of the programs contained within the four major modules are briefly described below.

4.1 The Genome And Proteome Sequence Analyis module of BioSuite deals with the applications relating to the analysis of the nucleic acid and protein sequences, not only of individual molecules, but also of complete genome and proteome sequences. This module would enable researchers to annotate genomes, predict protein secondary structures, derive a phylogenetic relationship among organisms and compare two genomes for similarities at the gene or protein level, along with a range of other applications. This module is further divided into four sub-modules: Sequence Analysis, Genome Analysis, Comparative Genomics and Utilities.

Sequence analysis of individual molecules are enabled through the sequence analysis modules, while the programs in ‘Genome analysis’ sub-module enable comparison and analysis of full genomes and proteomes. Two database searching tools, BLAST and PSI-BLAST are interfaced with the suite, that will enable searching databases to identify a given sequence or find conserved domains or even find distantly related homologues from some other species. An option of building custom-made databases is also provided. Alignment of sequences, a crucial task in sequence analysis, is provided for, through two well-established algorithms for global and local alignments using dynamic programing algorithms (Needleman-Wunsch & Smith-Waterman). Further, a hierarchical clustering-based multiple alignment algorithm (ClustalW) is included for aligning a set of sequences. Besides, pattern identification and matching tasks such as finding composition, inverted repeats, DNA structure motifs, restriction site analysis and repeat analysis, are part of this module.

Algorithms for secondary structure prediction including transmembrane region detection, RNA structure prediction and analysis are also part of this module. The secondary structure prediction algorithms were trained (or re-trained as appropriate) using a comprehensive dataset containing 731 high resolution protein structures (with resolutions ≤ 2 Å) that comprise a non-redundant dataset (Redundancy has been removed through sequence comparisons, using a similarity cut-off of 25% with the Blosum62 substitution matrix). Use of a large dataset in training the prediction algorithms ensures high prediction accuracy. A comprehensive biophysical parameter computation ability has also been built into BioSuite, by extracting 36 different physico-chemical properties for protein molecules from the data set and subsequently using them as training-sets in the prediction algorithms. Algorithms for predicting isoelectric point, peptide cleavage patterns, B-cell antigenicity from protein sequences are also included in this module. Yet another useful feature of this module is the domain building and analysing functionality. Programs are available for identifying domains, building consensus domain sequences, calibrating them and searching across a database. Hidden Markov models using sequence profiles are used for these purposes. In addition, the module has programs for studying molecular evolution, to cluster groups of sequences based on several criteria and to compute phylogenetic trees as well as to calculate evolutionary distances. Finally, algorithms for gene finding, gene assembly, probe and primer design, vector trimming and EST analysis are also part of this module. Some examples of using the various programs of this module are illustrated in Figure 2.

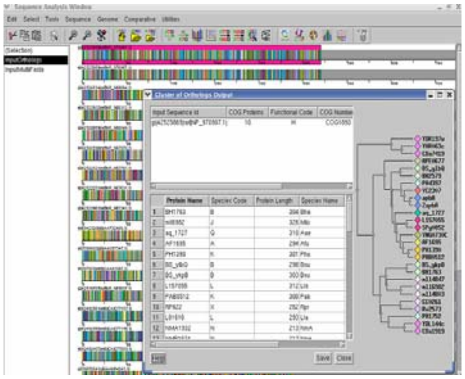
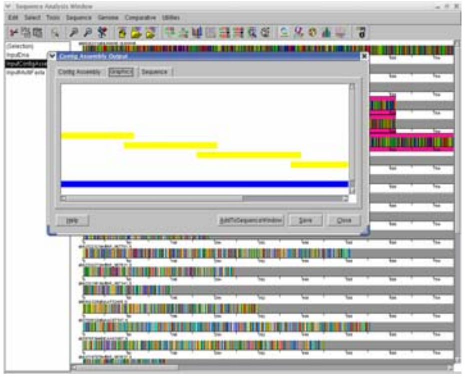
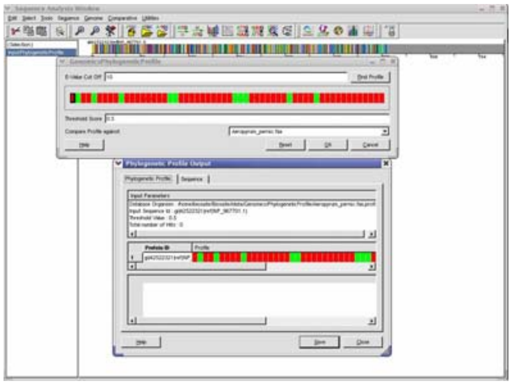


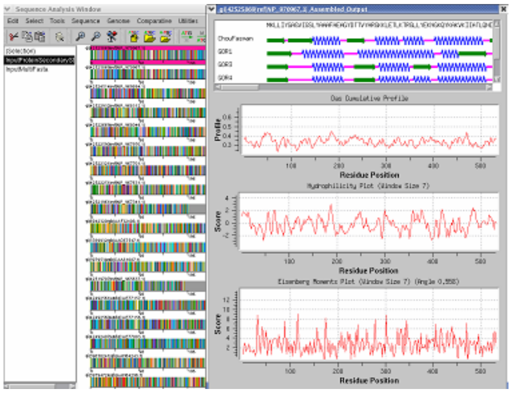
Figure 2a: Genome comparison: **Mapping Protein gi|42525869, from *Bacillus halorudians* to Clusters of Orthologous Groups (COG no. 1893), by using orthologues. A homologue for gi|42525869 from *Bacillus halorudians* was identified**



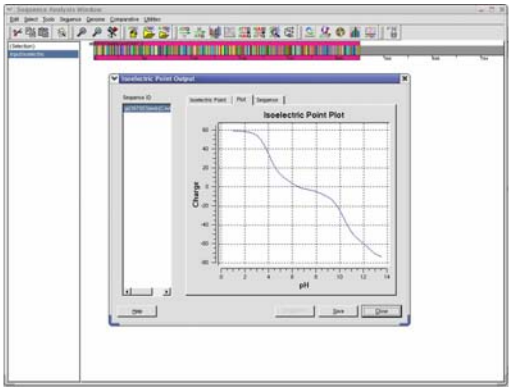
2b: Contig Assembly: Assembly of partial contigs from *E.coli* genome.

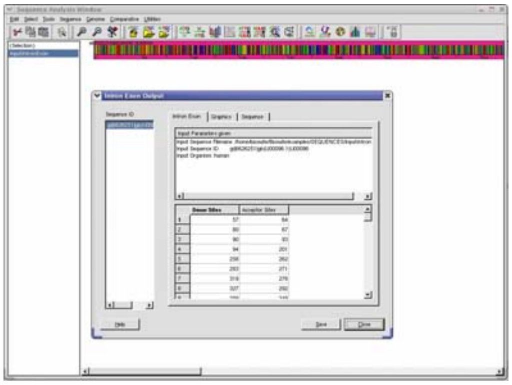


2c: Phylogenetic profiles: For gi|42525869, using phylogenetic profile, which shows similarities with *Mycobacterium tuberculosis*.



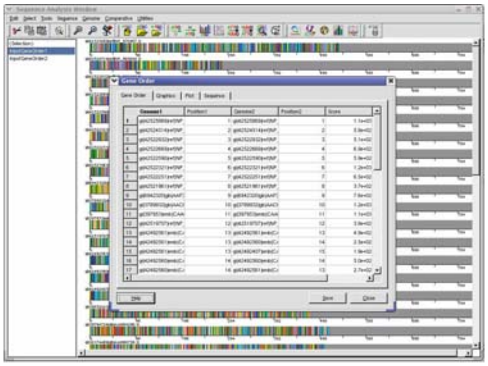
2d: Protein secondary structure prediction using different methods, Property profiles for gi|42525869 protein sequence

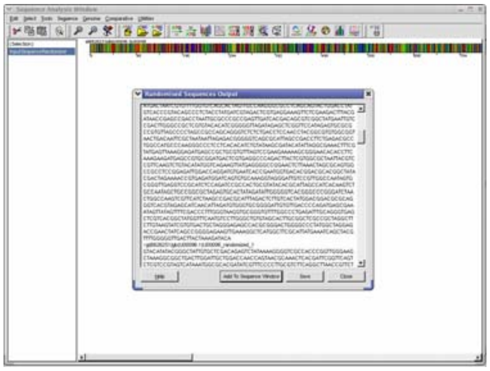




2e: IsoElectric Point :
Isoelectric Point plot for a protein sequence

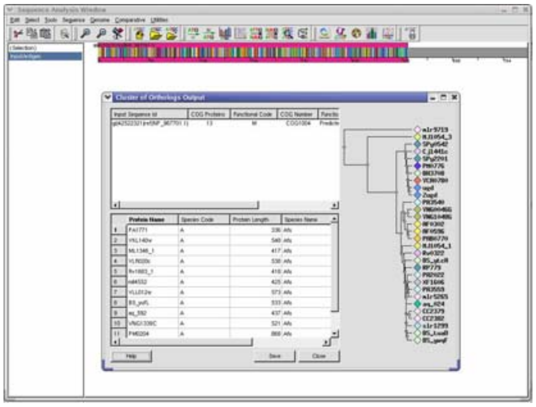
2f: Splice Site Prediction :
Intron / Exon boundary for an EST genome





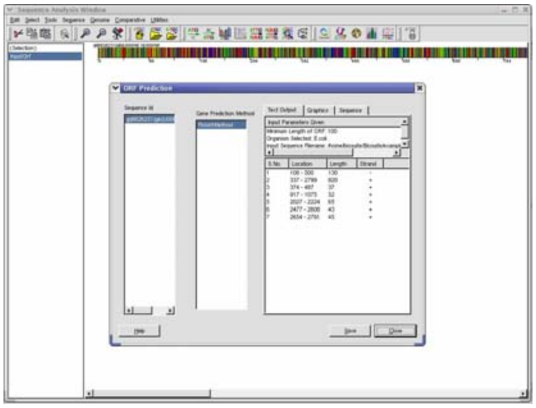
2g: Gene Order :
Gene Order table for 2 genome sequences

2h: Sequence Randomizer :
Randomised Sequences output for a given DNA sequence



Input Sequence Id	COG Proteins	Functional Code	COG Number	Functio
gi|42522321|ref|NP_967701.1|	13	M	COG1004	Predict

	Protein Name	Species Code	Protein Length	Species Name
1	PA1771	A	336	AAs
2	YKL140w	A	540	AAs
3	ML1348_1	A	417	AAs
4	YLR020c	A	538	AAs
5	Rv1683_1	A	418	AAs
6	mll4532	A	425	AAs
7	YLL012w	A	573	AAs
8	BS_yqfL	A	533	AAs
9	aq_592	A	437	AAs
10	VNG1308C	A	521	AAs
11	PM0204	A	868	AAs

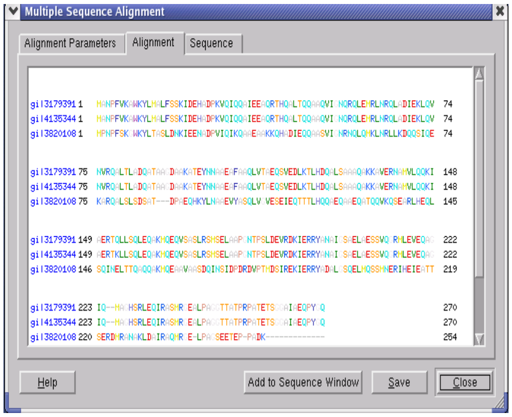


Input Parameters Given
Minimum Length of ORF: 100
Organism Selected: E.coli
Input Sequence Filename: [illegible]

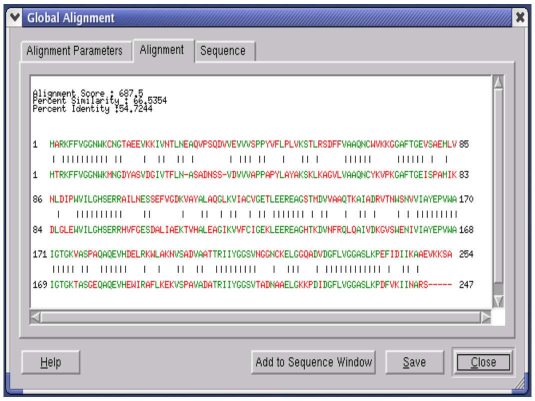
S.No.	Location	Length	Strand
1	108 - 500	130	-
2	337 - 2799	820	+
3	374 - 487	37	+
4	917 - 1075	52	+
5	2027 - 2224	65	+
6	2477 - 2608	43	+
7	2654 - 2791	45	+

2i: Orthologues :
Cluster of Orthologues for an amino acid sequence

2j: ORF prediction for a EST genome file using Hidden Markov Model for gene prediction



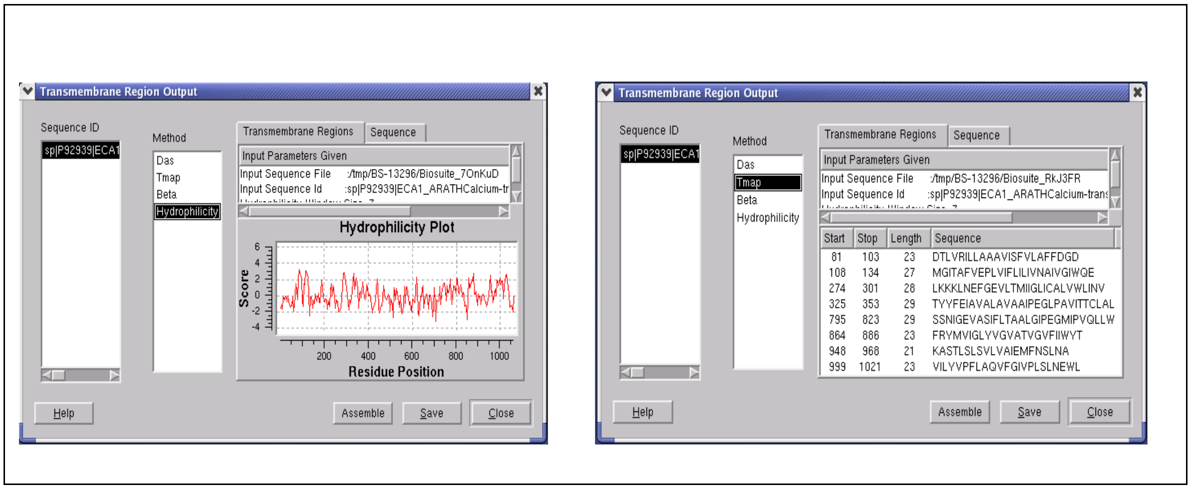
```
gi|3179391 1   MANPFVKAWKYLMALFSSKIDEHADPKVQIQQAIEEAQRTHQALTQQAAQVIGNQRQLEMRLNRQLADIEKLQV  74
gi|4135344 1   MANPFVKAWKYLMALFSSKIDEHADPKVQIQQAIEEAQRTHQALTQQAAQVIGNQRQLEMRLNRQLADIEKLQV  74
gi|3820108 1   MPNPFSKGWKYLTASLDNKIEENADPVIQIKQAAEAAKKQHADIEQQAASVIGNRNQLQMKLNRLLKDQQSIQE  74

gi|3179391 75  NVRQALTLADQATAAGDAAKATEYNNAAEAFAAQLVTAEQSVEDLKTLHDQALSAAAQAKKAVERNAMVLQQKI  148
gi|4135344 75  NVRQALTLADQATAAGDAAKATEYNNAAEAFAAQLVTAEQSVEDLKTLHDQALSAAAQAKKAVERNAMVLQQKI  148
gi|3820108 75  KARQALSLSDSAT---DPAEQHKYLNAAEVYASQLVGVESEIEQTTTLHQQAEQAAEQATQQVKQSEARLHEQL  145

gi|3179391 149 AERTQLLSQLEQAKMQEQVSASLRSMSELAAPGNTPSLDEVRDKIERRYANAIGSAELAESSVQGRMLEVEQAG  222
gi|4135344 149 AERTKLLSQLEQAKMQEQVSASLRSMSELAAPGNTPSLDEVRDKIERRYANAIGSAELAESSVQGRMLEVEQAG  222
gi|3820108 146 SQINELTTQAQQAKMQEAAVAASDQINSIDPDRDVPTMDSIREKIERRYADALGSQELMQSSMNERIHEIEATT  219

gi|3179391 223 IQ--MAGHSRLEQIRASMRGEALPAGGTTATPRPATETSGGAIAEQPYGQ                          270
gi|4135344 223 IQ--MAGHSRLEQIRASMRGEALPAGGTTATPRPATETSGGAIAEQPYGQ                          270
gi|3820108 220 SERDMRANAKLDAIRAQMRGE-LPAGSEETEP-PADK-------------                          254
```



```
Alignment Score : 687.5
Percent Similarity : 66.5354
Percent Identity :54.7244


1   MARKFFVGGNWKCNGTAEEVKKIVNTLNEAQVPSQDVVEVVVSPPYVFLPLVKSTLRSDFFVAAQNCWVKKGGAFTGEVSAEMLV 85
    | |||||||||| ||    |  ||  || |       | ||| ||   |   || |     ||||||     |||||| |  |
1   MTRKFFVGGNWKMNGDYASVDGIVTFLN-ASADNSS-VDVVVAPPAPYLAYAKSKLKAGVLVAAQNCYKVPKGAFTGEISPAMIK 83

86  NLDIPWVILGHSERRAILNESSEFVGDKVAYALAQGLKVIACVGETLEEREAGSTMDVVAAQTKAIADRVTNWSNVVIAYEPVWA 170
     |   ||||||||||    ||      |   ||  | ||  | || ||||||| | ||   |  || |    | | |||||||||
84  DLGLEWVILGHSERRHVFGESDALIAEKTVHALEAGIKVVFCIGEKLEEREAGHTKDVNFRQLQAIVDKGVSWENIVIAYEPVWA 168

171 IGTGKVASPAQAQEVHDELRKWLAKNVSADVAATTRIIYGGSVNGGNCKELGGQADVDGFLVGGASLKPEFIDIIKAAEVKKSA  254
    ||||| ||  ||||||   |  |   ||  ||  |||||||||   |  |||   | |||||||||||| |  || |
169 IGTGKTASGEQAQEVHEWIRAFLKEKVSPAVADATRIIYGGSVTADNAAELGKKPDIDGFLVGGASLKPDFVKIINARS-----  247
```

2k. Multiple sequence alignment of 35kDa Alanine rich protein from *M. bovis, M. tuberculois* and *C. diptheriae*. The residues here are color-coded based on standard physical nature of amino acid.

2l. Global Sequence Alignment of Triose phosphate isomerase enzymes from *Arabidopsis thaliana* and *C.elegans*. Alignment tab showing the output of sequence alignment. The sequences in green represent identical residues and red represent residues that are different.

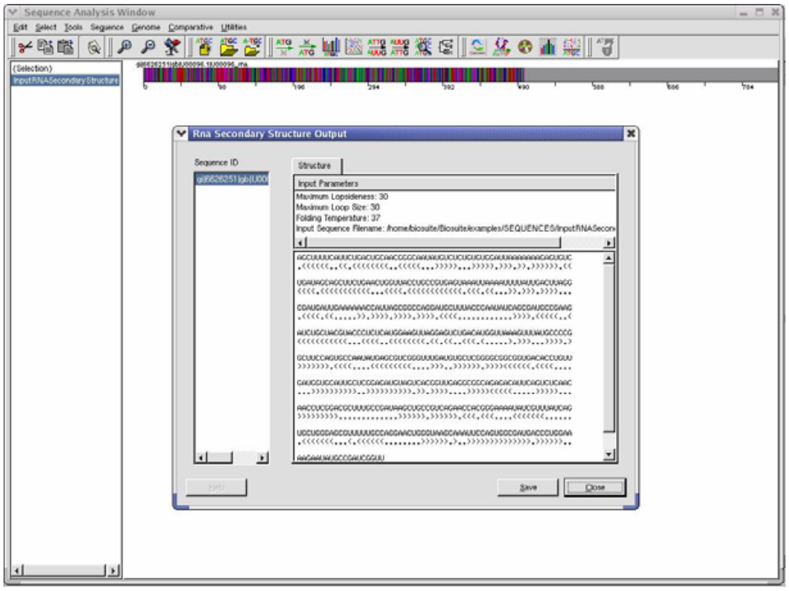


```
Input Parameters Given
Input Sequence File     :/tmp/BS-13296/Biosuite_7OnKuD
Input Sequence Id       :sp|P92939|ECA1_ARATHCalcium-tr
```

```
Input Parameters Given
Input Sequence File     :/tmp/BS-13296/Biosuite_RkJ3FR
Input Sequence Id       :sp|P92939|ECA1_ARATHCalcium-trans
```

Start	Stop	Length	Sequence
81	103	23	DTLVRILLAAAVISFVLAFFDGD
108	134	27	MGITAFVEPLVIFLILIVNAIVGIWQE
274	301	28	LKKKLNEFGEVLTMIIGLICALVWLINV
325	353	29	TYYFEIAVALAVAAIPEGLPAVITTCLAL
795	823	29	SSNIGEVASIFLTAALGIPEGMIPVQLLW
864	886	23	FRYMVIGLYVGVATVGVFIIWYT
948	968	21	KASTLSLSVLVAIEMFNSLNA
999	1021	23	VILYVPFLAQVFGIVPLSLNEWL

2m Transmembrane region in Ca dependent ATPase from *A. thaliana* ,
2n Hydrophilicity plot



```
Input Parameters
Maximum Lopsideness: 30
Maximum Loop Size: 30
Folding Temperature: 37
Input Sequence Filename: /home/biosuite/Biosuite/examples/SEQUENCES/InputRNASecon
```

```
[illegible]
```

2o: Comparison of RNA Secondary Structure prediction module of BioSuite with Vienna RNA Package. Biosuite enables customization of parameters required for the thermodynamic calculation, such as Folding Temperature , Maximum size of Internal Loop and Maximum Lopsidedness of Interior Loop. The presence of a pre-microRNA sequence in one of our query sequences identified through BioSuite was later validated through the miRNA Registry. The results were in full agreement with those obtained from the Vienna RNA Package (Hofacker I. L, 2003; Vienna RNA package: http://www.tbi.univie.ac.at/~ivo/RNA/

4.2 **3D Modelling and analysis**

The 3D Modelling and analysis module has capabilities to build, analyze and predict three dimensional structures of macromolecules and macromolecular complexes. This module is further subdivided into the following sub-modules: (a) Homology Modeling (b) Threading, (c) Building Proteins, (d) Building Nucleic Acids, (e) Building Carbohydrates, (f) Generation of Symmetry Related Molecules, (g) Structural Superposition (h) Surfaces and Volumes (i) Binding Site Analysis, (j) Nucleic Acid Analysis, (k) Interactions, (l) Quality Check and (m) Fold Detection.

Building the models of protein molecules by predicting their three dimensional structures by comparative modeling techniques are enabled through the first two sub-modules, for which 6 algorithms are available that incorporate the latest concepts in these areas. Building nucleic acids and carbohydrates using geometric information is enabled through the building modules. A notable feature of the builder programs is the incorporation of 17 geometrical templates for nucleic acids and 12 templates for carbohydrates providing a handle to address the stereo-chemical variability in a large number of sugars. Several programs that can address visualization and analysis of crystallographically derived structures are also included in this module. For example a lattice assembly of a protein molecule, as seen in its crystal structure can be generated easily. Structure validation tools for proteins and nucleic acids are enabled through the Quality check programs. Extensive analysis is possible through the Analysis and Interactions functions, that can be used for analyzing integral features of protein structure, protein-protein interactions as well as protein-ligand interactions. Finally, algorithms for classifying protein structures, in relation to the other protein structures known in literature, are also included in this module through the fold detection routines. Here too, the unique integration of building, analysis and structural bioinformatics tools such as structure classification, all within one framework, significantly enhances the technical value of BioSuite. Some examples of using the various programs of this module are illustrated in Figure 3.

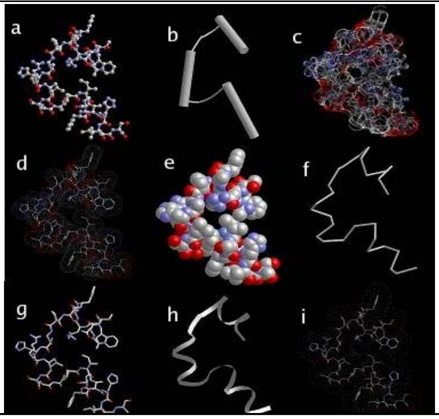
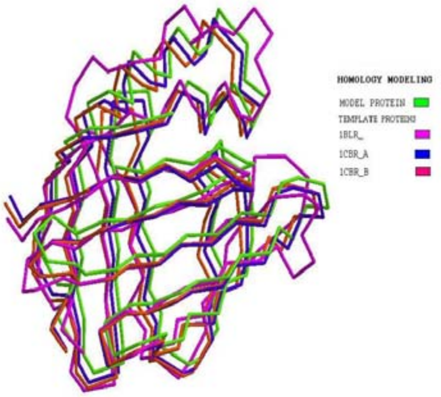
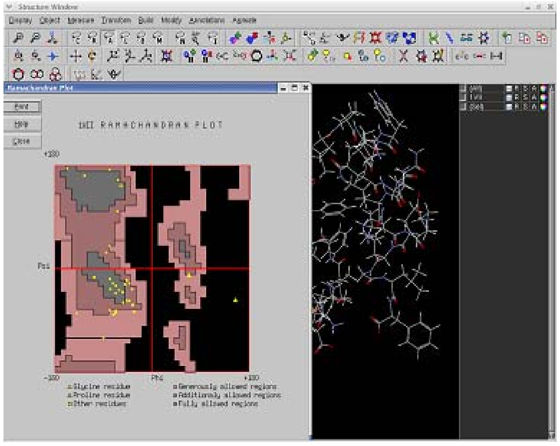
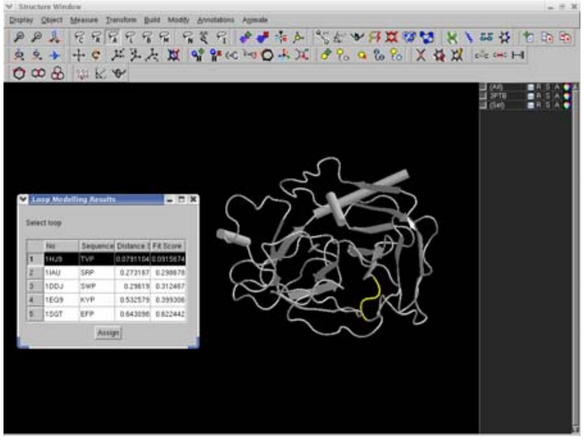


Figure 3a: Different molecular representations in BioSuite
a) Ball-and-stick b) Cartoons c) Molecular surface d)van der Waals surface e) Space fill f) C-alpha trace g) sticks h) ribbons i) solvent accessible surface.



3b: Illustration of structure prediction using Homology modeling methods



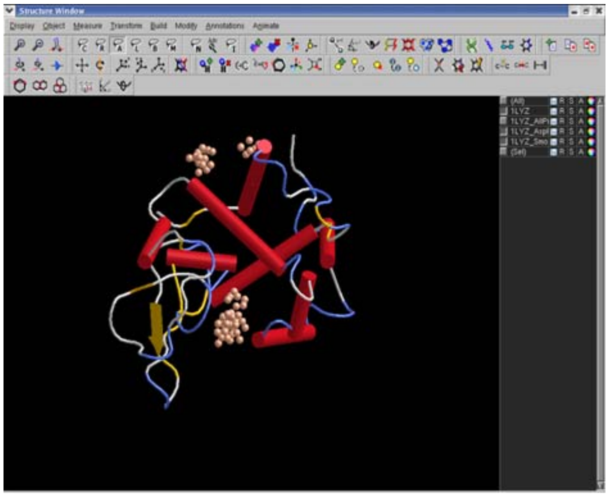

Structure Window
Display Object Measure Transform Build Modify Annotations Animate
Loop Modelling Results
Assign

Select loop

	No	Sequence	Distance S	Fit Score
1	1HJ9	TVP	0.0791104	0.0915874
2	1IAU	SRP	0.273187	0.298678
3	1DDJ	SWP	0.29619	0.312467
4	1EQ9	KYP	0.532579	0.399306
5	1SGT	EFP	0.643096	0.622442

3c: Quality Check-Ramachandran plot

3d: Illustration represents loop modelling between preflex and postflex region (72 ILE -81LYS) of molecule 3PTB –beta trypsin based on the best distance score and fit score of (1.0) first loop number1HJ9 was derived from the loopdatabase was assigned to build the loop. Loop is highlighted in yellow color.


Structure Window
Display Object Measure Transform Build Modify Annotations Animate
(All)
1LYZ
1LYZ_AllPr
1LYZ_Aspl
1LYZ_Sno
(Sel)

3e: Illustration shows the molecule 1LYZ Lysozyme with brown colored spheres, which represents the probe and active site points in the molecule - Binding site detection by PASS

4.3 Simulations

The 'Simulations' module essentially simulates the behaviour of a molecule, in terms of its three dimensional structure. The different sub modules covered are, Forcefield, Energy Minimization, Molecular Dynamics, Monte Carlo simulations and Electrostatics. The molecular simulation of a system can conceptually be broken into three components: (a)

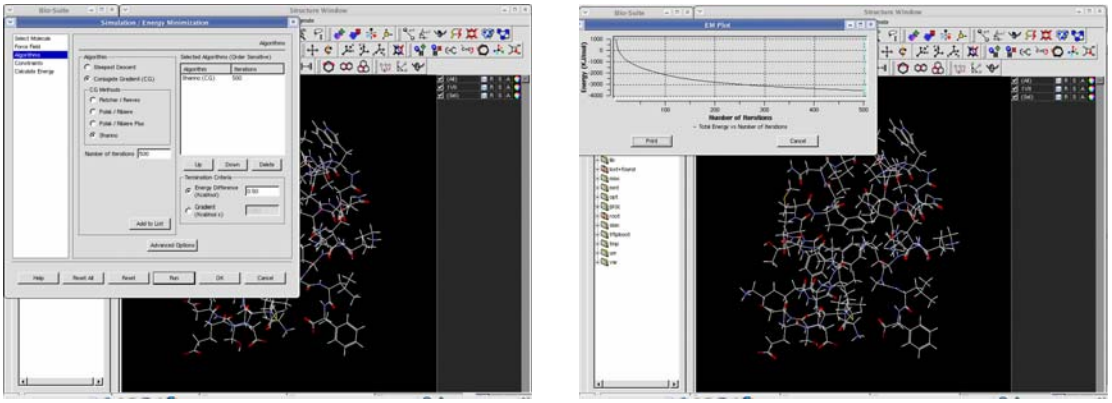
Generating a computational description of a biological/chemical system typically in terms of atoms, molecules and associated force field parameters, (b) The numerical solution of the equations which govern their evolution and (c) The application of statistical mechanics to relate the behaviour of a few individual atoms/molecules to the collective behaviour of the very many. BioSuite is compatible both with the AMBER and the CHARMM force fields for macromolecules (proteins, nucleic acids and carbohydrates) and uses GAFF for small molecules (for eg., natural substrates, drugs and drug-like substances). For each of the force fields, both treatments of the type of dielectric: either constant or distant dependent, are provided.

Several algorithms for first-order unconstrained energy minimization are contained in this module, providing a wide range of line search options. Thus, the coordinates of the molecular system can be adjusted so as to lower its energy, relative to the starting conformation, by using one of the following minimizers: Steepest Descent Algorithm, Conjugate Gradient Methods, Fletcher-Reeves Algorithm, Polak-Ribiere Algorithm, Polak-Ribiere Plus Algorithm and Shanno's Algorithm.

Further, to carry out molecular dynamics (MD) simulations, BioSuite provides NVE (Micro-canonical), NVT (Canonical), and NPT (Isobaric-Isothermal) ensembles for MD Simulations with the choice of using Velocity-Verlet or Leapfrog integrator. BioSuite also provides options for using SHAKE and RATTLE constraints.

MD being a deterministic approach, where the state of the system at any future time can be predicted from its current state, the tools provided in the suite can be used for solving Newton's equations of motion for a given initial conformation, to study how the system evolves over time. Several intuitive and user-friendly tools are provided to analyse the resulting trajectories or time series of conformations. For example, plots at various energy levels along with the temperature, can be obtained. Plots generated with defined parameters show the structure and position at various energy levels, both of them present in two adjacent panels that can help to view the position of the molecule at a given temperature. The Monte Carlo method that generates configurations randomly and uses a special set of criteria to decide whether or not to accept each new configuration, is also part of this module.

In the electrostatics sub-module, BioSuite provides a solution for the Linear Poisson-Boltzmann Equation, to enable modeling of contributions of solvent, counterions and protein charges to electrostatic fields in molecules. Four choices for boundary conditions namely, zero, partial coulombic, full coulombic and focusing, are provided. For charge distribution, there are two options, trilinear and uniform. BioSuite has a very fast SOR solver, which utilizes spectral radius calculations to speed up convergence. Some of the results obtained from simulations on example proteins are shown in Figure 4.



Algorithm: ● Conjugate Gradient (CG)

CG Methods: ● Shanno

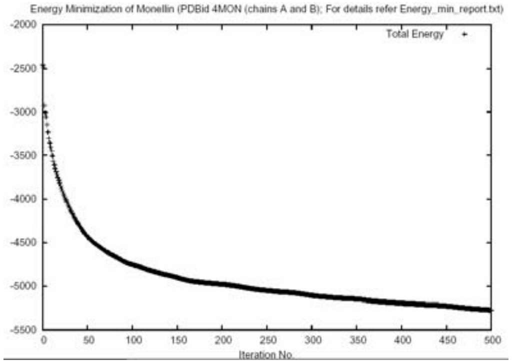
Number of Iterations: 500

Selected Algorithms (Order Sensitive)

Algorithm	Iterations
Shanno (CG)	500

Termination Criteria: ● Energy Difference (Kcal/mol) 0.50

Figures 4a :Snapshot of the various algorithms that can be used to perform Energy Minimization and **4b**: the graph of Energy versus Number of Iterations, obtained on running Energy Minimization interactively, that helps in determining convergence criteria.



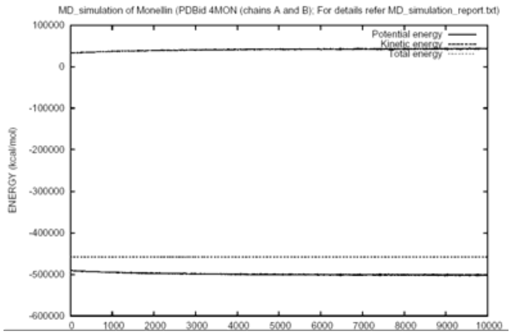
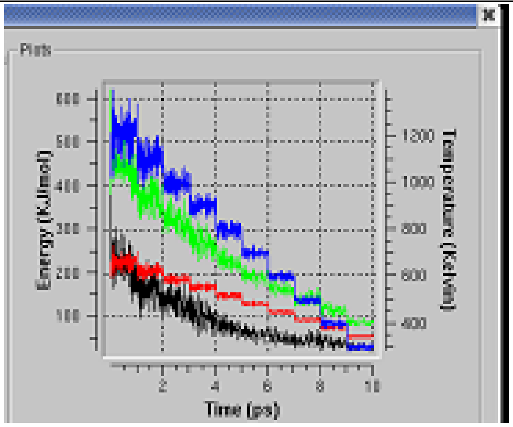
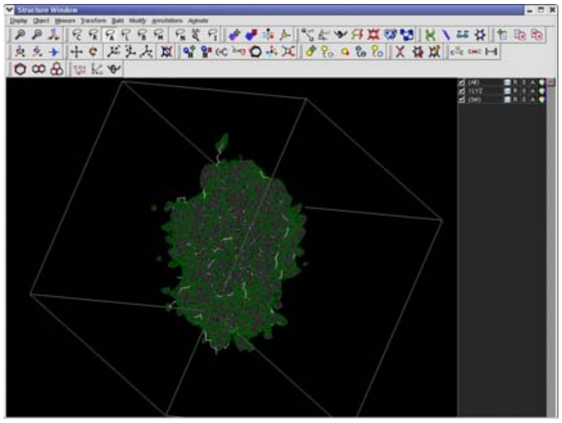


Figure 4c: Hydrogen atoms were added to the system (PDB:1MON). Then energy minimization was done for 500 iterations. Default parameters were followed. A plot of the total minimized energy is shown.

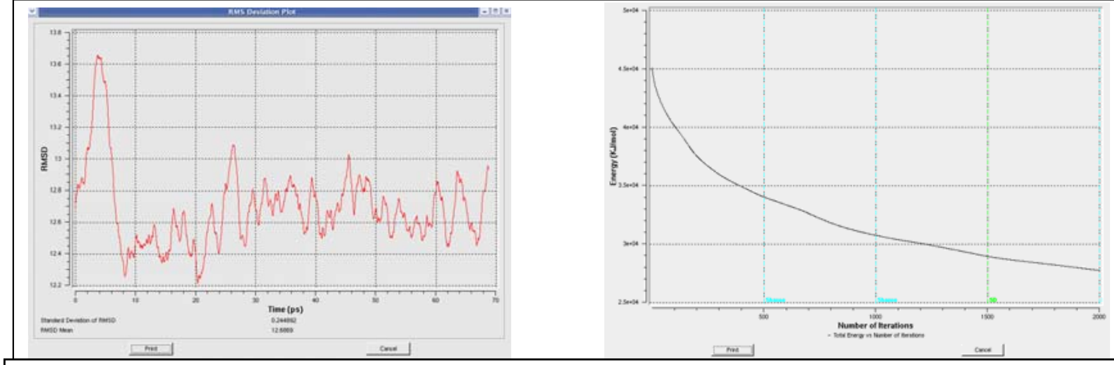
4d: The energy minimized system given in 4a was then solvated with 2358 water molecules by choosing a radius of 5 angstrom. Then MD simulation was carried out for 10000 iterations using default parameters. A plot of potential energy, kinetic energy and total energy is shown.



4e: An example of MD-analysis, Variation in Kinetic Energy, Potential Energy, Total Energy, Temperature during simulation.

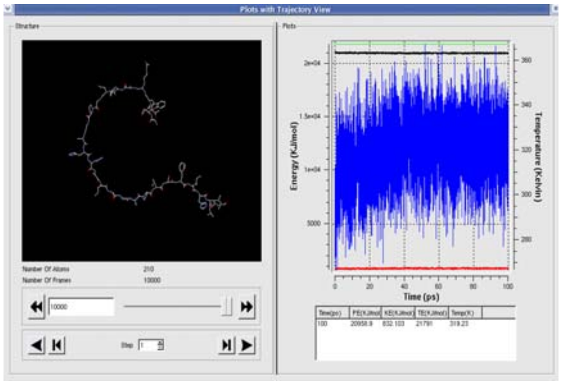


4f: Illustration showing Isopotential surface around the Lysozyme molecule for given set of potential charges (pink and green color represent the charges) -Electrostatic fields



Standard Deviation of RMSD 0.244862
RMSD Mean 12.6869

Gastrin, a 28 amino acid peptide was subjected to a molecular Dynamics simulation and analysed using BioSuite. Hydrogens were added to the initial model and the electrostatic charges using the ‘Electrostatics’ module, followed by an energy minimization of the peptide, using the Steepest descent algorithm followed by a conjugate gradient (Fletcher/Ribiere/Ribiere Plus/Shanno) algorithm. MD simulation was done in vacuum using the CHARMM force field and with periodic boundary conditions for 100 ps with an initial 10ps of equilibration. The time step of integration was 1fs and non-bonded update was done every 20fs. Figure **4e** shows a trajectory of RMSD of all Cα atoms of gastrin, **4g**: Plot of Total energy of gastrin as a function of number of iterations during the Energy minimization of gastrin



Number Of Atoms 210
Number Of Frames 10000

10000

Time(ps)	PE(KJ/mol)	KE(KJ/mol)	TE(KJ/mol)	Temp(K)
100	20958.9	832.103	21791	319.23

Figure **4h** gives the snapshot of the window of trajectory analysis showing the animation of structure as well as the evolution of energy with time

4.4 Drug design

The Drug Design module provides the following functionalities:(a) Prediction of biological activities of unknown chemical entities using QSAR, (b) Identification of Pharmacophores in biologically active molecules, (c) Superimposition of a set of molecules in 3D space by Alignment, (d) Identification of the ligand poses in 3D space when it binds to a target using Docking. Using the functionalities provided in the Drug Design module, one can identify lead-like molecules from a set of molecules, redesign them and predict their activities. Thus, lead optimization can be achieved iteratively. If the target structure is known, then the lead optimization can be done using the structure based method, such as by docking.

The process of aligning a set of molecules in three dimensional space, to find the superimposable regions of a group of molecules or to estimate molecular similarity can be performed by using either the ‘Field Fitting’ or the ‘RMS Fitting’ approaches. The Field fitting is done by aligning molecules using their electrostatic potentials and steric shapes, starting from their atomic coordinates and charges computed from Gaussian functions, while the ‘RMS Fitting’ is done by minimizing the distances between specified atoms in the molecules. Flexible superposition can also be achieved by allowing rotations about single bonds.

For deriving and matching ‘3D-Pharmacophores’, the following features are extracted/used: (a) Hydrogen Bond Donor (b) Hydrogen Bond acceptor, (c) Aliphatic hydrophobic group, (d) Aromatic ring, (e) Negatively charged group and (f) Positively charged group. Identification of pharmacophores is done by using configurations of features common to a set of molecules. The pharmacophoric configurations are identified by a pruned exhaustive search, starting with small sets of features and extending them until no larger common configuration exists.

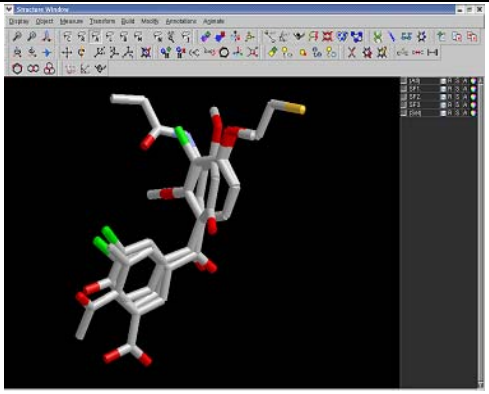
To carry out QSAR, where consistent relationships between the variations in the values of molecular properties and the biological activity for a series of compounds are sought, so that these "rules" can be used to evaluate new chemical entities, a series of widely accepted feature extraction and statistical tools are provided within BioSuite. For example, a 2D-QSAR calculation uses either one or combinations of (a) Electronic, (b) Spatial, (c) Structural, (d) Thermodynamic and (e) Topological descriptors. BioSuite has the ability to compute 89 different descriptors. a few representative descriptors from different classes e.g. Polarizability, HOMO and LUMO (electronic), Hf and Log P from (thermodynamic), log P, MR (thermodynamic), etc. and were compared with those computed from standard softwaers. using a dataset of 33 isooxazoles (figure 1) as potential thrombin receptor antagonists and in general, a high correlation (>0.9) was observed for the descriptor values, as illustrated in Figure 5a.

Creating and refining a training set required for QSAR predictions, are aided by (a) K-means, (b) K-Nearest Neighbours or (c) UPGMA hierarchical clustering algorithms. Tools are also provided for building user-defined data sets/ training sets as well as for searching chemical databases. The QSAR model can be generated using regression techniques such as Multiple Linear Regression or Partial Least Squares. If the linearly independent descriptors for the molecules have to be eliminated while generating the model, then a dimensionality reduction can be performed by using either (a) Principal Component Analysis or (b) Discriminant Analysis. Validation of the model to check the accuracy of the generated model can be performed by the K–fold cross validation technique

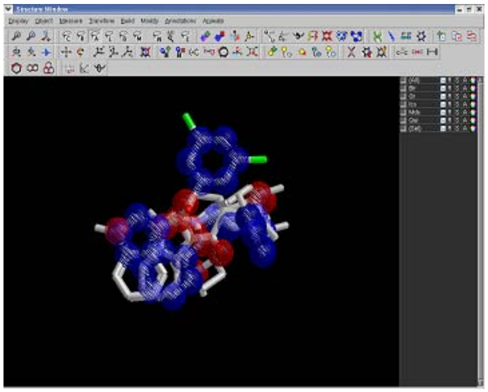
The structure based drug design sub-module, contains algorithms and utilities required for carrying out molecular docking. Using either simulated annealing or genetic algorithms (GA) based technique, the ligand conformations are searched and docked into the binding site of the macromolecule. In a simulated annealing based method, the ligand's current position, orientation and conformation are changed during each cycle, to reach the most energetically favorable conformation of the ligand bound to the target macromolecule. Thus these algorithms predict both the lowest energy conformation of the bound ligand as well as the best position and orientation for its binding to the target molecule, within the realm of the scientific capabilities of the approach.

A second popular algorithm is provided for this, the one based on genetic algorithms. The conformations of the ligand are encoded as a chromosome. The crossover and mutation operators are used to bring about random changes in the conformations of the ligand. A fitness function is defined for the calculating the energy of the conformations generated. Through a number of runs of the GA cycle, a conformation having minimum energy is obtained.

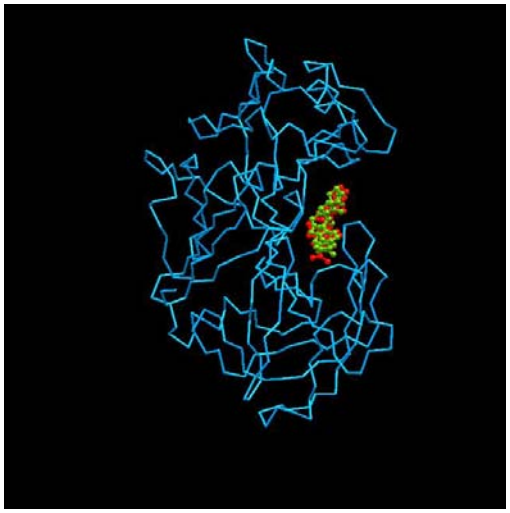
Conformation search functionality generates the conformations for an input molecule, clusters the conformations and displays energy and torsion angle values of low energy conformations. This application generates conformations using two different methods, namely Random Conformation Search and Systematic Conformation search. Random Conformation Search uses the Simulated Annealing algorithm. Option is provided to the user to select the rotatable bonds in the molecule. A few sample results from the Drug-Design modules are presented in Figure 5.



5a: Alignment of ligand molecules



5b: Pharmacophore fitting



5c: Results obtained using the Docking routine of BioSuite. A modified peptide inhibitor has been docked to find the position for the best interaction with rhizopuspepsin. The docked inhibitor shown in green ball and stick representation has the lowest energy of interaction of -14.1 KCal/mol. Inhibitor in the crystal structure of the complex (PDB: 3APR) with rhizopuspepsin is shown in red.

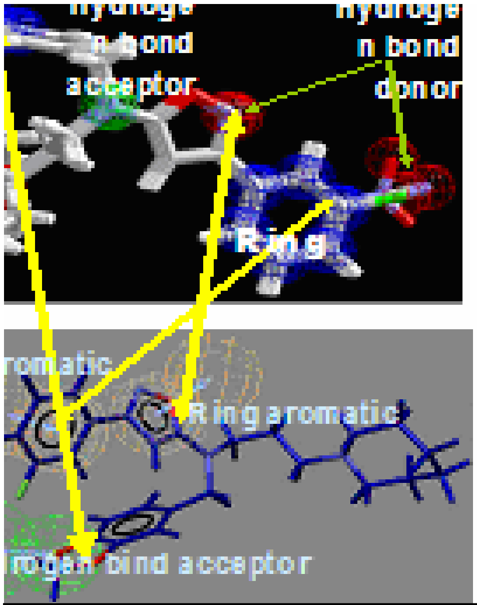


Figure 5e: A comparison of the common chemical features identified by Biosuite and Catalyst

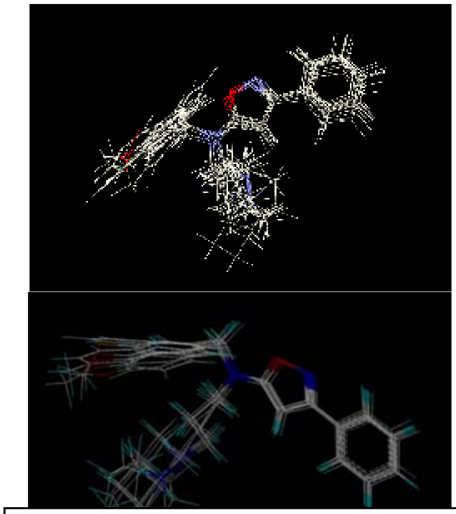


Figure 5f: Evaluation of field fit alignment: A visual inspection of the alignments produced by both Biosuite and Sybyl shows that they generate comparable alignment. Molecular similarity between a pair of molecules is calculated by using the Gaussian function in BioSuite.

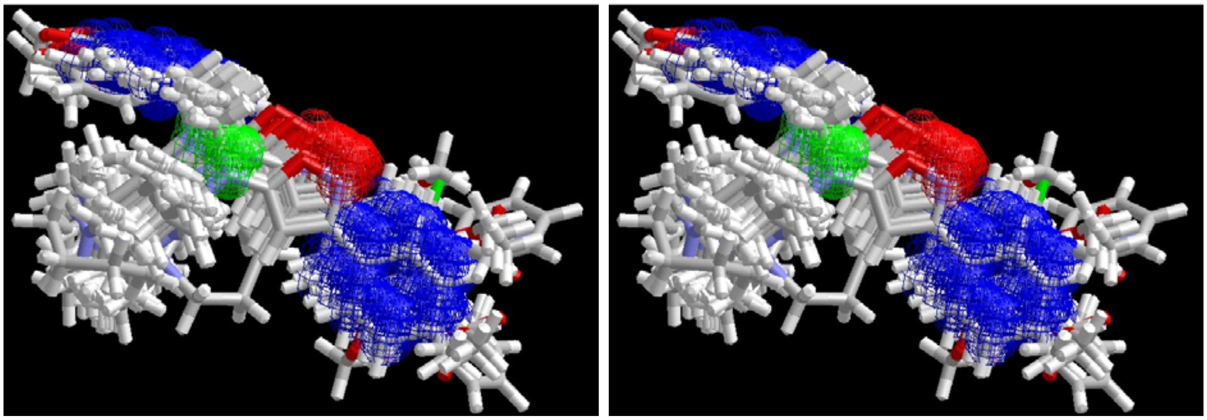


Figure 5g: Alignments produced by BioSuite derived pharmacophore model

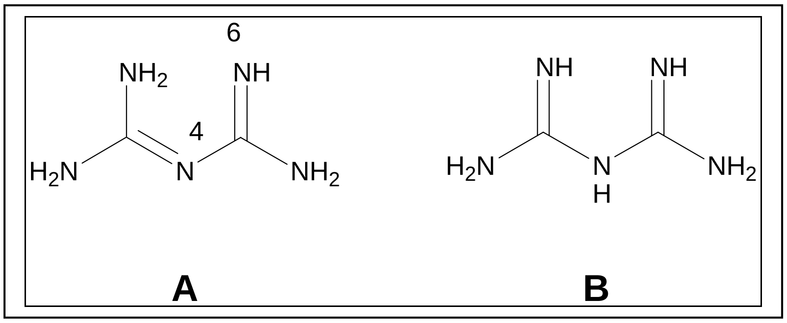


Figure 5h: Predicting the energetically favourable tautomeric form (A) and hence the conformation of metformin, an antidiabetic drug, as compared to the alternate form (B), consistent with experimental and quantum chemical observations.

5. Performance evaluation

Evaluation has been an integral part of the entire development process. To start with, the choice of modules and the choice of algorithms themselves were evaluated, both at TCS and by the academic partners. The pseudo-codes and the SRS documents were then verified, followed by verification of the software codes by the TCS team. The scientific performance of the algorithms, at various stages (versions 0.3, 0.7, 1.0a and 1.0) were evaluated independently by the academic partners at their institutions and any bugs reported or improvements suggested, were subsequently considered and implemented into the suite, where appropriate. The outputs of each program were compared with those of other established academic codes/commercial packages, to verify the scientific performance. They were also compared with the latest implementations of the chosen algorithms in the public domain, where available. The performance has been found to be comparable in all cases. While the utilities of many of the individual programs have been enhanced while implementing in BioSuite, the scientific capabilities and limitations of each of the programs are bounded by those of the corresponding original algorithms cited in Table 1.

An example of the manner in which the scientific performance was evaluated, is cited below. For testing the drug design module, 42 thymidine monophosphate kinase inhibitors were taken and minimization performed using both AMBER and CHARMM force fields with the conjugate gradient algorithm method. Conformational searches were tested with both systematic and randomized search methods. Alignments were satisfactory and we obtained low RMSD values for similar molecules, comparable to those obtained in Cerius^2. The time for computation was found to be good and comparable to other competitor software. The docking procedure is simple and user-friendly.

6. Prominent features of the package

For the most part, the existing software packages evolved out of academia, and were implementations of algorithms developed at different places and different times by different persons. As such, often there is no single "superstructure" into which the algorithms fit seamlessly. To overcome these issues, BioSuite has been written in a modular fashion, which would permit the easy implementation of new algorithms as and when they are discovered. The unique partnership of the industry with academia harnesses the strengths of both communities, thus leading to a superior product both scientifically as well as according to software engineering standards. Some of the unique features of BioSuite are,

(a) It is comprehensive, contains programs for carrying out sequence, whole genome and structure analysis, drug design, all under a common framework.
(b) The software runs on simple personal computers on a Linux platform.
(c) Domain identification and domain searching tools also available
(d) Transmembrane beta strand prediction, enhanced capability in building molecules in terms of the number of secondary structure templates available
(e) Enhanced capability in building larger carbohydrate structures
(f) Code written fresh with CMMi-5 standards and consistency in coding methods to incorporate versatility in each program making up the entire suite, keeping in view of the genome-scale operations in bioinformatics.

7. Roadmap for the future

Going forward, several features are planned to be added to BioSuite to make it an even more useful platform for scientific research. Some developments in the pipeline are described below:

ADME: The Absorption, Distribution, Metabolism and Excretion profile (ADME) of a drug is an important determinant of its therapeutic efficacy. Accurately modelling the ADME properties of a candidate drug molecule is a necessary step to increase the chances that it will eventually become a successful drug. In the recent past, models have been developed for estimating various ADME related properties such as blood-brain barrier penetration (Narayanan *et al.* 2005), human intestinal absorption (Zhao *et al*., 2001 and Feher *et al*., 2002), binding affinity to Human Serum Albumin (Colmenarejo *et al*., 2001) and $CaCO_2$ cell permeability. These will be integrated into the existing QSAR module of BioSuite.

Flexible Docking: Docking, in BioSuite 1.0, explores the energetically optimal fit of a flexible small molecule with a rigid protein molecule. In subsequent releases, an improved version of the docking algorithm will be implemented that allows restricted flexibility in the protein molecule as well. This has been shown to be useful in improving the accuracy in prediction of the optimal binding conformation.

De novo Drug Design: An important requirement for drug design is the ability to generate novel molecules that bind to a known active site. Implementation of an algorithm is underway for the generation of novel binding candidates using a strategy of fragment docking followed by elaboration of selected fragments.

tRNA Identification: A procedure for identifying tRNA genes in a genome will be included in the next version of BioSuite. The program identifies tRNAs based on the recognition of two intragenic control regions known as A and B boxes, a highly conserved part of B box, a transcription termination signal, and the evaluation of the spacing between these elements (Pavesi *et al.*, 1994, Laslett *et al.*, 2004 and Hentschel, 2001).

Improved Whole Genome Comparison:_MUMmer is an open source software package for the rapid alignment of very large DNA and amino acid sequences. A newer version of the MUMmer package has been integrated in BioSuite to find maximal unique matches between two genomes. The MUMmer output can also be viewed in the dot-plot format.

Improved Graphics: Several techniques are being implemented to enhance the quality of the 3D graphics display in BioSuite while speeding up the display.

Scripting Interface: While BioSuite provides a number of features and a vast array of functionality, users might want to implement their own procedures and programs. For this purpose, a scripting interface that exposes the functionality in BioSuite will be provided so that users can create their own workflows, develop and test new ideas and automate several tasks.

Sketcher: The next version of Bio-Suite will include a 2D sketcher for drawing molecules in a manner that chemists are familiar with and to automatically generate 3D structures for the molecules.

A high-performance version called Bio-Cluster for some of the memory intensive applications is also planned.

8. Availability-contact person(s) for BioSuite and websites

BioSuite Web-site: http://www.atc.tcs.co.in/BioSuite/

Contact: Dr. Sharmila Mande
Head, Life Sciences R&D Division
Advanced Technology Centre
Tata Consultancy Services
Deccan Park
#1 Software Units Layout
Hyderabad – 500 081
E-Mail: sharmila@atc.tcs.co.in
Tel: +91 40 5567 3541(D) / 5567 2000(B)
Fax No: +91 40 5567 2222

Sales Contact:
In-Charge, BioSuite sales team,
Life Sciences R&D Division
Advanced Technology Centre
Tata Consultancy Services
Deccan Park
#1 Software Units Layout
Hyderabad – 500 081
E-mail: biosuite_sales@atc.tcs.co.in
Tel.: +91 5567 3576(D) / 5567 2000(B)
Fax No: +91 5567 2222

9. Hardware requirements and Documentation:

The minimum hardware requirements for BioSuite are as follows: Intel compatible x86 Processor, 1.5 GHz, 256 MB RAM, 3 GB Free Hard Disk Space, Display capable of 1280 X 1024 pixel resolution, High end graphics card with 3D support for better viewing, Red-Hat Linux 8.0 or 9.0 or Fedora-Core 1/ 2 operating systems. BioSuite comes with its own set of documentation. The entire package is well documented and comes with easy to use tutorials, which reduce the learning curve and increase efficiency.

10. Summary

BioSuite, a comprehensive software package dealing with Bioinformatics and computational biology tools has emerged as a result of the CSIR sponsored (NMITLI) industry-academia collaboration. The industry Tata Consultancy Services, has undertaken the coding responsibility with several academic partners playing the advisory role. The capabilities of the different modules of BioSuite are presented in this paper. The package contains algorithms that comprehensively cover several aspects of computational biology through sequence and structural analysis of biological macromolecules. It also contains computational tools that enable application of bioinformatics and chemoinformatics analysis to aid drug discovery at various stages. Further enhancements to the software are also planned by means of incoporating newer algorithms to provide additional capabilities. It is expected that the package will be used by a large community of research institutions, pharmaceutical companies and universities for research, development and teaching purposes. This project also demonstrates the merits of collaboration between the industry and the academia that has led to harnessing the strengths of both strong fundamental domain knowledge as well as that of professional software development. This project can also be viewed as a stepping stone in the area of commercial bioinformatics software development in the country, which could lead to the genesis of more such ventures, taking India to a leadership position in the area.

11. References

1. Acton FS (1990). Numerical methods that work. Corrected edition (Washington: Mathematical Association of America), 55: 454-458.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local

alignment search tool. Journal of Molecular Biology, 215: 403 - 410.

3. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Meller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.
4. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T. A uniform system for microRNA annotation RNA, 2003, 9(3), 277-279.
5. Andersen H (1983). Journal of Computational Physics, 52: 24-34.
6. Andersen HC (1980). Journal of Chemical Physics, 72: 2384-2393.
7. Arthur, L. D., Kasif, S., Fleschmann, R. D., Peterson, J., White, O. and Salzberg, S. L. (1999). Alignment of whole genomes. Nucleic Acid Research 27(11): 2369-2376.
8. Bailey, T. L. and Elkan, C. (1995). Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. Machine Learning Journal, 21: 51 – 83.
9. Bailey, T. L. and Gribskov, M. (1996). The megaprior heuristic for discovering protein sequence patterns, Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California. pp. 15-24.
10. Bailey, T. M. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics, 14: 48-54.
11. Baker, E. N., and Hubbard, R. E. (1984). Hydrogen Bonding in Globular Proteins, Progress in Biophysics and Molecular Biology, 44: 97-179.
12. Bansal M, Bhattacharyya D and Ravi B (1995) NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. Comput Appl Biosci., 11(3):281-287.
13. Barnum, D., Greene, J., Smellie, A., and Sprangue, P. (1996). Identification of Common Functional Configurations Among Molecules. J.Chem.Inf.Comput.Sci.36: 563-571.
14. Barry, D. and Hartigan, J. A. (1987). Asynchronous distance between homologous DNA sequences. *Biometrics* **43**: 261-276
15. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984). Molecular Dynamics with coupling to an external bath. Journal of Chemical Physics, 81(8): 3684-3690.
16. Bhattacharya, R., Samanta, U., and Chakrabarti, P. (2002). Aromatic-aromatic interactions in and around alpha helices, Protein Engineering, 15, 2: 91-100.

17. Borodovsky, M. and McIninch, J. (1993). GENMARK: Parallel gene for both DNA strands. Computers chem. 17(2): 123 – 133.

18. Bottcher, C. J. F. (1973). Title: Theory of Electric Polarization, Elsevier Press, Amsterdam

19. Brady, G. P., and Stouten, F. W. P. (2000). Fast Prediction and Visualization of Protein Binding Pockets with PASS, Journal of Computer-Aided Molecular Design 14: 383-401.

20. Branden.C & Tooze.J (1991). Introduction to Protein Structure. Garland Publishing inc, New York and London.

21. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S and Karplus M (1983). CHARMM: A program for Macromolecular Energy Minimization, and Dynamics Calculations. J. Comp. Chem., 4(2): 187-217.

22. Chou, P. Y. and Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid Sequence. Advances in Enzymology 47: 45 – 148.

23. Connolly, M. L (1983B). Solvent-accessible surfaces of proteins and nucleic acids,Science, 221: 709.

24. Connolly, M. L. (1983A). Analytical Molecular Surface Calculation Journal of Applied Crystallography, 16: 548.

25. Connolly, M. L. (1985). Computation of Molecular Volume. Journal of the American Chemical Society, 107: 1118-1124

26. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson Jr. DM, Spellmeyer DC, Fox T, Caldwell JW and Kollman PA (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc., 117: 5179-5197.

27. Cserzo, M., Wallin, E., Simon, I., von Heijne, G. and Elofsson, A. (1997). Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the Dense Alignment Surface method; Protein Engineering, 10 (6): 673-676.

28. Dan Gusfield. (1997). Algorithms on Strings, Trees, and Sequences. Cambridge University Press, Cambridge, UK. Pp23-29.

29. Dayhoff, M. O. (1978). Survey of new data and computer methods of analysis. In M. O. Dayhoff ed., Atlas of Protein Sequence and Structure, vol. 5, supp. 3, National Biomedical Research Foundation, Silver Springs, Maryland. pp. 29.

30. Debye, P. and Huckel, E. (1923). *Physik. Z.* **24**: 185.

31. Delcher, A. L., Harmon, D., Kasif, S., White, O. and Salzberg, S. L. (1999). Improved

microbial gene identification with GLIMMER. Nucleic Acids Research 27(23): 4636-4641.

32. Dodd, I. B. and Egan, J. B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Research, 18: 5019 - 5026.

33. Dunbrack, Jr. and M. Karplus. "Backbone-dependent Rotamer Library for Proteins: Application to Side-chain prediction." J. Mol. Biol., 230, 543-574 (1993).

34. Durbin, R. Eddy, S. Krogh, A. and Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, UK.

35. Eck. R. V. and Dayhoff, M. O. (1966). Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Springs, Maryland.

36. Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics, 14: 755–763.

37. Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature, 299(5881): 371-374.

38. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. and Scharf, M. (1995). Double Cubic Lattice Method: Efficient Approaches to Numerical Integration of Surface Area and Volume to Dot Surface Contouring of Molecular Assemblies, Journal of Computational Chemistry, 16: 273-284.

39. Enright, A. J. and Ouzounis, C. A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16: 451-457.

40. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90.

41. Corpet, F. (1988), "MultAlin: Multiple sequnce alignment with hierarchical clustering", Nucl. Acids Res. 16 (22) 10881-10890

42. Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783-791.

43. Felsenstein, J. (1981). Evolutionary Trees from DNA sequences: a maximum likelihood approach *Journal of Molecular Evolution* **17**: 368- 376.

44. Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. Nucleic Acids Research 10(17): 5303 – 5318.

45. Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**: 279-284.

46. Fitch, W. M. (1971). Towards defining the course of evolution: Minimum change for a specific tree topology. Systematic Zoology 20: 406-416.

47. Fletcher R and Reeves C (1964). Function minimization by conjugate gradients. Computational Journal, 7: 149-154.

48. Fogolari, F., Zuccato, P., Esposito, G. and Viglino, P. (1999). *Biophysics Journal* **76**(1): 1-16.

49. Frenkel D and Smit B (2002). In: Understanding Molecular Simulations – From Algorithms to Applications. Academic Press, New York

50. Garnier. J., Gibrat, J. F. and Robson, B. (1996). GOR secondary structure prediction method version IV. Methods in Enzymology (Ed.) R. F. Doolittle. 266: 540 – 553.

51. Garnier. J., Osguthorpe, D. J. and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology 120 (1): 97 - 120. (GOR I).

52. Gerstein, M., Tsai, J., Levitt, M. (1995). Volume of Atoms in Protein Surface, Journal of Molecular Biology, 249: 955-966.

53. Gibrat, J. F., Garnier, J. and Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. Journal of Molecular Biology, 198(3): 425-443.(GOR III).

54. Gilbert JC and Nocedal J (1992). Global convergence properties of conjugate gradient methods for optimization. SIAM Journal on Optimization, 2; 21-42.

55. Good A.C., Hodgkin, E.E. and Richards, W.G. (1992). Utilization of Gaussian functions for the rapid evaluation of molecular similarity. J. Chem. Inf. Comput. Sci. 32: 188.

56. Goodman, J.M. (1998). Chemical Applications of Molecular Modelling. The Royal Society of Chemistry, London, pp. 61-69.

57. Goodsell, D. S. and Olson, A.J. (1990). "Automated Docking of Substrates to Proteins by Simulated Annealing”. Proteins: Str. Func. and Genet., 8:195-202.

58. Gotoh, O. (1982). An Improved Algorithm for Matching Biological Sequences. Journal of Molecular Biology, 162: 705-708.

59. Griffiths-Jones S. The miRNA Registry NAR, 2004, 32, Database Issue, D109-D111.

60. Gromiha, M. M., Majumdar, R. and Ponnuswamy, P. K. (1997). Identification of membrane spanning beta strands in bacterial porins. Protein Engineering, 10: 497-500.

61. Gund, P (1977). Prog Mol Subcell Biol. 5: 117-143.

62. Guner, O.F. (2000), Pharmacophore perception, development and use in drug design. The International University Line, pp. 24-35.

63. Gute, B.D. and Basak, S.C. (2001). Molecular similarity-based estimation of properties: a comparison of three structure spaces. Journal of Molecular Graphics and Modelling, 20:95–109.

64. Hansch, C. and Leo, A. (1995). Exploring QSAR – Fundamentals and Applications in Chemistry and Biology. ACS Professional Reference Book, American Chemical Society, Washington D.C: 1-19, 69-85, 97-118 and 513-535.

65. Hendy, M. D. and Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences* **59**: 277-290 .

66. Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of National Academy of Science* **89**: 10915 – 10919.

67. Higgins, D.G., Bleasby, A.J. and Fuchs, R. (1992) CLUSTAL V: improved software for multiple sequence alignment. Computer Applications in the Biosciences (CABIOS), 8(2):189-191.

68. Hockney RW and Eastwood JW (1981). In: Computer Simulations using Particles. McGraw-Hill, New York.

69. Hopp, T. P. and Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proceedings of National Academy of Sciences, 78(6): 3824-3828.

70. Huang, X. (1992). A contig assembly program based on sensitive detection of fragment overlaps. Genomics 14(1): 18 – 25.

71. Hussain, A.S.Z. Kiran Kumar, Ch. Rajesh, C.Ksheik, S.S. and Sekar, K. Nucleic Acid Research (2003), 31, 3356-3358.

72. Hyndman, D. L. and Mitsuhashi, M. (2003). PCR Primer Design. Methods in Molecular Biology 226: 81 - 88.

73. Ivo L. Hofacker Vienna RNA secondary structure server Nucleic Acids Research, 2003, Vol. 31, No. 13 3429-3431.

74. Jaeger, J. A., Turner, D. H. and Zuker, M. (1989). Improved predictions of secondary structures for RNA. Proceedings of National Academy of Sciences, 86: 7706-7710.

75. James U. Bowie *et al* (1991). A Method to Identify Protein Sequences That Fold into a Known Three- Dimensional Structure. Science, 253:164-170.

76. Jayaram B, Sharp KA and Honig B (1989). Biopolymers, 28: 975-993.

77. Jonathan M,.Goodmann (1998). Chemical Applications of Molecular Modelling. 61-69

78. Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). *Computer Applied Bioscience* **8**:

275282

79. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology, 292: 195-202.

80. Jones, G.; Willet, P.; Glen, R. C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. J. Comput.-Aided Mol. Des. 1995, 9, 532

81. Jukes, T. H. and Cantor, C. R. (1969). Evolution of protConformation Search functionality In H N Munro ed., Mammalian Protein Metabolism, Academic Press, New York. pp. 21-132.

82. Kabsch, W. (1978). Acta crystallographica, A34: 827.

83. Kimura, M. (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**:111-120.

84. Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, Massachusetts.

85. Klapper I, Hagstrom R, Fine R, Sharp K and Honig B (1986). Proteins, 1: 47.

86. Kleffe, J., Hermann, K., Vahrson, W., Wittig, B. and Brendel, V. (1996). Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. Nucleic Acids Research, 24(23): 4709 – 4717.

87. Knuth, D. E., Morris, J. H. and Pratt V.R. (1977) Fast Pattern Matching in Strings. SIAM Journal of Computing 6, 2, 323-350.

88. Kolaskar and Tangaonkar (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Letters 276:172-174.

89. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology, 235: 1501–1531.

90. Kurogi, Y. and Guner, O.F. "Pharmacophore Modeling and Three-Dimensional Database Searching for Drug Design Using Catalyst". Curr. Med. Chem., 2001, 8(9),1035-1055

91. Kyte, J. and Doolittle, R. F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein . Journal of Molecular Biology, 157(6): 105-142.

92. Landau, G. M., Schmidt, J. P. and Sokol, D. (2001). An algorithm for approximate tandem repeats. Journal of Computational Biology 8(1): 1-18.

93. Laskowski R A, MacArthur M W, Moss D S & Thornton J M, PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst, 26, 283-291, (1993).

94. Lassmann, T and Erik L.L. Sonnhammer (2002) Quality assessment of multiple alignment programs, FEBS Letters 529 126-130

95. Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. Journal of Molecular Biology, 55:379.

96. Lichtarge, O, Bourne, H. R., and Cohen, F. E. (1996). An Evolutionary Trace method defines binding surfaces common to protein families. ,Journal of Molecular Biology 257: 342-358.

97. Lupas, A, Van Dyke, M. and Stock, J. (1991). Predicting coiled coils from protein sequences. Science, 252: 1162-1164.

98. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. Nature 402: 83-86.

99. Mazumdar, A., Kolaskar, A. and Donald, S. (2001). GeneOrder: Comparing the order of genes in small genomes. Bioinformatics 17: 162-166.

100. McDonald, I. K., and Thornton, J. M. (1994). Satisfying Hydrogen Bonding Potential in Proteins, Journal of Molecular Biology, 238: 777-793.

101. McGaughey, B. G., Gagne, M., and Rappe, A.K. (1998). Pi Stacking Interactions, Journal of Biological Chemistry, 273, 25: 15458-15463.

102. McLachlan A. D (1982). Acta Cryst A38: 871-873

103. McLachlan, A. D. (1979). J. Mol. Biol. 128: 49—79.

104. Metropolis, N, Rosenbluth, A.W, Rosenbluth, M.N, , Teller, A.N, and Teller, E. Equations of state calculations by fast computing machines. Journal of Chemical Physics, 21, 1087-1092 (1953)

105. Mitchell, J.B.O., and Nandi, L.C., McDonald, I. K., and Thornton, J. M. (1994). Amino/Aromatic Interactions in proteins: Is the Evidence Stacked Against Hydrogen Bonding? , Journal of Molecular Biology, 239: 315-331.

106. More’ JJ and Thuente DJ (1994). Line search algorithms with guaranteed sufficient decrease. ACM Transactions on Mathematical Software, 20(3): 286-307

107. Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998). "Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function". J. Computational Chemistry, 19: 1639-1662

108. Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology,

48: 443-453.

109. Nicholls A. and Honig B (1991). Journal of Computational Chemistry, 12(4): 435-445

110. Orengo, C.A and Taylor, W.R. (1989). Protein Structure Alignment, Journal of Molecular Biology208:1-22.

111. Orengo, C.A and Taylor, W.R. (1996). SSAP: Sequential Structure Alignment Program for Protein Structure Comparison, Methods in Enzymology 266: 617-635.

112. Persson B and Argos P (1997). Prediction of membrane protein topology utilising multiple sequence alignments. Journal of Protein Chemistry, 16: 453-457.

113. Polak B and Ribiere G (1969). Note sur la convergence des methodes de directions conjugees. Rev. Fran. Informat. Rech. Oper., 16: 35-43.

114. Prakash T, Khandelwal M, Dasgupta D, Dash D, and. Brahmachari S.K. (2004) CoPS: Comprehensive Peptide Signature Database, Bioinformatics; 20: 2886 - 2888.

115. Luthy R, Mclachlan A, Eisenberg D (1991). Proteins, 10: 229.

116. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE, 77: 257–286.

117. Ramachandran plot on the web. S.S. Sheik, P. Sundararajan, A.S.Z. Hussain and K.Sekar. BIOINFORMATICS.(2002) 18, 1548-1549.

118. Richards, F. M. (1977). Annual Review on Biophysics, Bioengineering, 6: 151.

119. Richards, F.M. (1974). Interpretation of Protein Structures, Journal of Molecular R. L.

120. Rossmann M. G and E. Arnold, International Tables for Crystallography Vol. F: Crystallography of Biological Macromolecules.

121. Saitou, N., and M. Nei. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.

122. SantaLucia, J.Jr., Allawi, H.T. and Seneviratne, P.A. (1996). Improved Nearest-Neighbor parameters for predicting DNA duplex stability. Biochemistry 35: 3555-3562.

123. Schuler, G. D. (1997). Sequence Mapping by Electronic PCR. Genome Research 7: 541-550.

124. SEM : Symmetry Equivalent Molecules - A web based GUI to generate and visualize the macromolecules

125. Shanno DF (1978). Conjugate gradient methods with inexact searches. Mathematics of

Operation Research, 3(3): 244-256

126. Shrake and Rupley (1973). Environment and exposure to solvent of protein atoms:Lysozyme and insulin, Journal of Molecular Biology, 79: 351-371.Biology, 82: 1-14.

127. Smith, T. F. and Waterman, M. S. (1981). Identification of Common Molecular Subsequences. Journal of Molecular Biology, 147: 195-197.

128. Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*. Freeman, San Francisco

129. Sober, E. (1988). Reconstructing the Past. MIT Press, Cambridge, Massachusetts.

130. Sutcliffe, M.J., Haneef, I., Carney, D., & Blundell, T.L (1987). Knowledge based modeling of homologous proteins, Part I: Three dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Engineering 1: 377-384

131. Swofford, D. L. (1993). Phylogenetic Analysis Using Parsimony (PAUP), Version 3.1.1. University of Illinois, Champaign.

132. Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996). Phylogenetic inference. In: (Eds.) Hillis, D. M., Moritz, C., Mable, B.K. Molecular Systematics, Sinauer, pp. 407-514.

133. Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**: 512-526.

134. Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Molecular Biology and Evolution* **9**: 678-687.

135. Tajima, F. and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* **1**: 269-285.

136. Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33-36.

137. Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33-36.

138. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,

positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22: 4673-4680.

139. Todeschini, R. and Consonni, V. (2000). Handbook of Molecular Descriptors, Wiley-VCH, Weinheim(Germany).

140. Watowich SJ, Meyer ES, Hagstrom R and Josephs R (1988). A stable, rapidly converging conjugate gradient method for energy minimization. Journal of Computational Chemistry, 9(6): 650-661.

141. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S Jr. and Weiner PK (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. J. Am. Chem. Soc., 106: 765-784.

142. Zhang, C., and Kim, S. (2000). Environment -dependent residue contact energies for proteins,Proceedings of the National Academy of Sciences, 97: 2550-2555

143. Zhu, H., and Rohwer, R. (1996). No free lunch for cross-validation, Neural Computation, 8:1421-1426.

144. Zuker, M., Mathews, D. H. and Turner, D. H. (1999). Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide In RNA Biochemistry and Biotechnology, (eds.) Barciszewski, J. and Clark, B. F. C. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, NL. Pp 11-43.

145. Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. Science, 244: 48-52.