Sanjeevini: A comprehensive active site directed lead design software

B Jayaram*, N Latha, Tarun Jain, Pankaj Sharma, A Gandhimathi & Vidhu S Pandey

Department of Chemistry & Supercomputing Facility for Bioinformatics & Computational Biology,
Indian Institute of Technology, Hauz Khas, New Delhi 110 016, India
Email: bjayaram@chemistry.iitd.ac.in

Received 12 September 2004; revised 4 December 2005

Sanjeevini – a comprehensive active site directed lead compound design software, based on the on-going research in our laboratory, is described here. The computational pathway integrates several protocols proceeding from the design of chemical templates to lead-like molecules, given the three dimensional structure of the target protein and a definition of its active site. A conscious attempt has been made to handle the target biomolecule and the candidate drug molecules at the atomic level retaining system independence while providing access for systematic improvements at the force field level. Concerns related to geometry of the molecules, partial atomic charges, docking of candidates in the active site, flexibility and solvent effects are accounted for at the current state-of-the-art. To ensure theoretical rigor, binding free energy estimates are developed for candidate molecules with the target protein within the framework of statistical mechanics. We present herein, the technical and scientific features of *Sanjeevini*, its validation and scope for further improvement. Some modules of *Sanjeevini* have been made accessible at http://www.scfbio-iitd.res.in/drugdes/sanjeevini.html.

Lead molecule discovery and development is an expensive and lengthy process. For the pharmaceutical industry, the number of years to bring a drug from discovery to market is approximately 14 years, costing up to US$880 million per individual drug[1]. Given the vast size of organic chemical space ($>10^{18}$ compounds)[2], drug discovery cannot be reduced to a simple "synthesize and test" drudgery. There is an urgent need particularly for life threatening diseases to identify and/or design lead-like molecules from the vast expanse of what could be synthesized. There are approximately 6000 drugs in the market (CMC Database 94.1)[2] directed to approximately 1000 targets[3]. It is anticipated that application of bioinformatics and structure based drug design would significantly result in an increase of the therapeutic molecular targets to as many as 10,000 within the next few years[4]. The reader is adverted to some excellent reviews and studies published during the last decade on computer-aided drug design[5-12] addressing fragment based libraries for *de novo* design, QSAR, docking and ranking candidates, etc. and demonstrating the utility of *in silico* methods in reducing the time and cost involved in drug discovery.

The development of a comprehensive lead design software christened *Sanjeevini*, its current features and its validation, merits and scope for further research is described here. Perspective on the feasibility of automating lead design endeavors in the near future is also presented.

Theoretical

The computational pathway for *Sanjeevini*

The *Sanjeevini* software has been developed as a computational pathway paving the way expressly towards automating lead design (Fig. 1), making any number of known or new candidate molecules out of a small but versatile set of building blocks called templates, screening them for drug-likeness, optimizing their geometry, determining partial atomic charges and assigning other force field parameters, docking the candidates in the active site of a given biological target, estimating the interaction/binding energy, performing molecular dynamics simulations with explicit solvent and salt on the biomolecular target, the candidate and the complex followed by a rigorous analysis of the binding free energy for further optimization. Presently, we have coupled *Sanjeevini* with AMBER[13] and GAMESS[14] for molecular mechanics and quantum mechanics calculations, respectively. There are a total of seven modules, which make *Sanjeevini* a complete drug design software (Table 1). The source codes for all the modules are written in FORTRAN, C and C++ computer languages with numerous interfacial UNIX based shell scripts, which make all the modules work like a pipeline, such that the output of the previous step becomes the input for the next step. The modules under *Sanjeevini* can also be used independent of the pathway. The programs have been compiled and tested on LINUX and SOLARIS platforms.

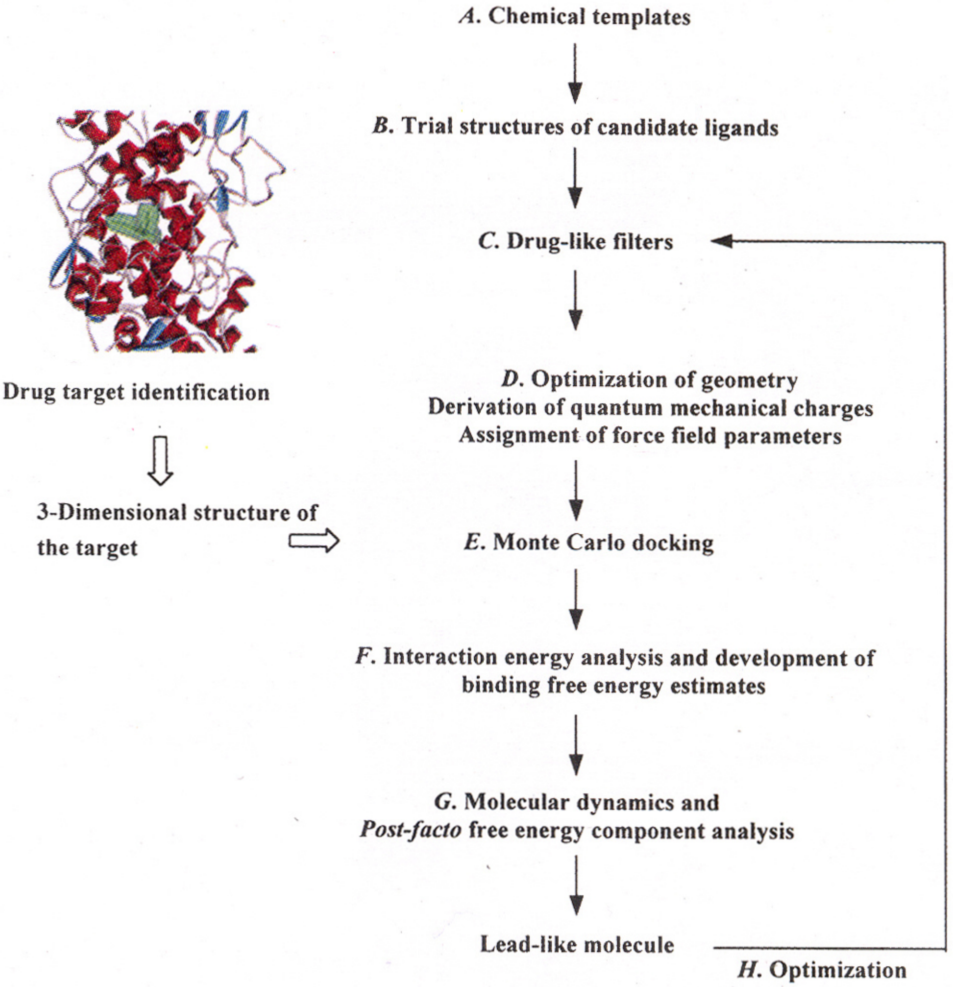

A. Chemical templates
B. Trial structures of candidate ligands
C. Drug-like filters
Drug target identification
D. Optimization of geometry
Derivation of quantum mechanical charges
Assignment of force field parameters
3-Dimensional structure of the target
E. Monte Carlo docking
F. Interaction energy analysis and development of binding free energy estimates
G. Molecular dynamics and *Post-facto* free energy component analysis
Lead-like molecule
H. Optimization

Fig. 1 — The *Sanjeevini* pathway for active site directed lead compound design *in silico*

Description of the modules and their performance

Module 1: Template library

Chemical templates are conceived as building blocks/structural frameworks for assembly and generation of new molecules. A necessary but not sufficient condition for creating a library of templates is that the structural and functional space of all known drugs be sampled, i.e. the set of templates has to be complete and non-redundant. Several alternative proposals exist in the literature for creating fragment/substructure based libraries[15,16] which typically involve some correlation with drug-likeness or biological activity to a target.

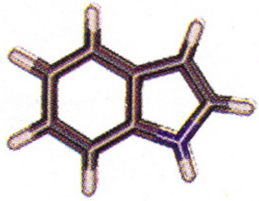
Table 1 — Main modules currently present in *Sanjeevini*

Template library
Molecule generator
Molecular descriptors/drug-like filters
Molecular docking
Structural analysis of the receptor-candidate complex
Energy analysis of the receptor-candidate complex
Binding affinity analysis

We have built a sub-structure based template library for the design of novel compounds and created a set of 160 templates[11], classified into three groups: rings, side chains and linkers. The molecular structural formulas of the templates have also been described already[11]. For each template the structure was initially geometry optimized using AM1[17], followed by HF/6-31G*/RESP[14,18] calculations to derive the partial atomic charges. Non-bonded Lennard-Jones (12,6) parameters are assigned in a force field compatible manner[19]. Each template file, illustrated in Table 2, apart from containing the

Table 2 — Template file for indole (template code i02). The molecular structure file organized in the protein data bank (RCSB) format[20]

1	2	3	4	5	6	7	8	9	10	11	12	13
ATOM	1	C1	i02	1	-1.037	-1.538	0.000	0	CA	1.9080	0.0860	-0.2194
ATOM	2	C2	i02	1	0.208	-2.147	0.000	0	CA	1.9080	0.0860	-0.1605
ATOM	3	C3	i02	1	1.399	-1.397	0.000	0	CA	1.9080	0.0860	-0.2049
ATOM	4	C4	i02	1	1.377	-0.011	0.000	0	CA	1.9080	0.0860	-0.1994
ATOM	5	C5	i02	1	0.145	0.644	0.000	0	CA	1.9080	0.0860	0.1409
ATOM	6	C6	i02	1	-1.076	-0.135	0.000	0	CA	1.9080	0.0860	0.0895
ATOM	7	N7	i02	1	-2.153	0.754	0.000	0	N3	1.8240	0.1700	-0.2875
ATOM	8	C8	i02	1	-1.647	2.058	0.000	0	CM	1.9080	0.0860	-0.1766
ATOM	9	C9	i02	1	-0.256	2.033	0.000	0	CM	1.9080	0.0860	-0.3063
ATOM	10	H10	i02	1	-1.959	-2.133	0.000	1	HA	1.4590	0.0150	0.1554
ATOM	11	H11	i02	1	0.272	-3.246	0.000	1	HA	1.4590	0.0150	0.1487
ATOM	12	H12	i02	1	2.362	-1.928	0.000	1	HA	1.4590	0.0150	0.1513
ATOM	13	H13	i02	1	2.309	0.571	0.000	1	HA	1.4590	0.0150	0.1572
ATOM	14	H14	i02	1	-3.104	0.503	0.000	1	H	0.6000	0.0157	0.3298
ATOM	15	H15	i02	1	-2.326	2.913	0.000	1	H1	1.3870	0.0157	0.1997
ATOM	16	H16	i02	1	0.413	2.887	0.000	1	HC	1.4870	0.0157	0.1823
CONECT	1	2	6	10								
CONECT	2	1	3	11								
CONECT	3	2	4	12								
CONECT	4	3	5	13								
CONECT	5	4	6	9								
CONECT	6	5	1	7								
CONECT	7	6	8	14								
CONECT	8	7	9	15								
CONECT	9	8	5	16								
CONECT	10	1										
CONECT	11	2										
CONECT	12	3										
CONECT	13	4										
CONECT	14	7										
CONECT	15	8										
CONECT	16	9										
END												
REMARK:												
001002004	0	0	0									
001006004	0	0	0									
001010001	0	0	0									
002003004	0	0	0									
002011001	0	0	0									
003004004	0	0	0									
003012001	0	0	0									
004005004	0	0	0									
004013001	0	0	0									
005006004	0	0	0									
005009001	0	0	0									
006007001	0	0	0									
007008001	0	0	0									
007014001	0	0	0									
008009002	0	0	0									
008015001	0	0	0									
009016001	0	0	0									



Cartesian (X, Y & Z) coordinates (columns 6, 7 and 8) also carries atomic locations for further connections and extensions of the molecular framework (column 9), atom types (column 10), force field parameters (columns 11 and 12), partial atomic charges (column 13), connect records and mole information about all the atoms, making each template chemically complete and ready for molecule generator module.

Module 2: Molecule generator

As a step towards *de novo* lead design, candidates are generated from chemical templates introduced in the previous step. Recovery of the known drugs as well as novelty of the new candidates generated, besides their activity, are the prime considerations at this stage. The molecule generator requires parameter file for bond length, bond angle and bond dihedral parameters[19]. User has to select input files of two templates from the template library with their linkage positions specified. At the junction point of linkage, a hydrogen atom is lost from each template and a single bond is formed. The generated candidate molecule could act as a template for larger molecules in the next generation. The dihedrals around the junction are kept at 180°. The program can generate user specified molecules or it can run in a high

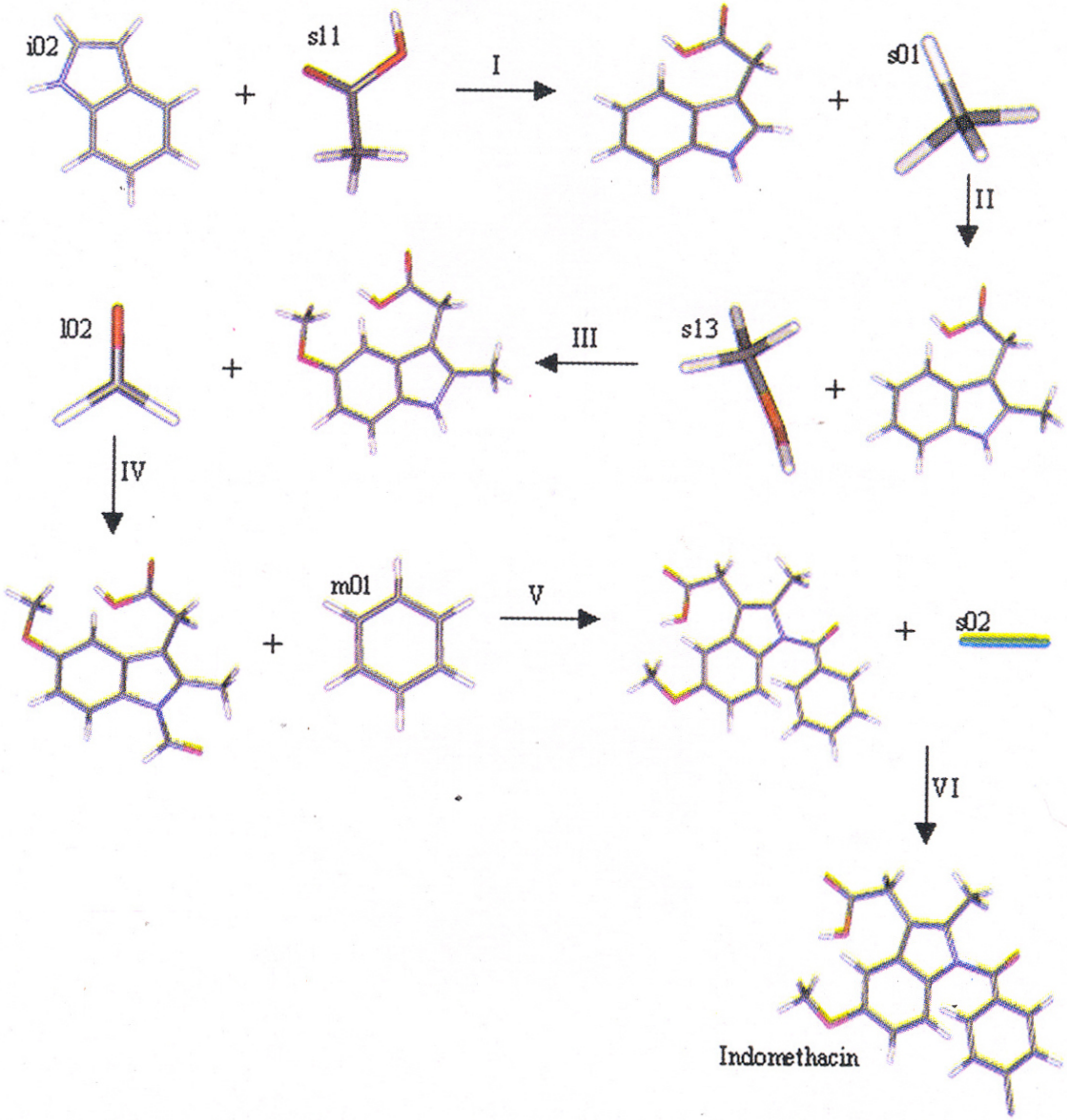
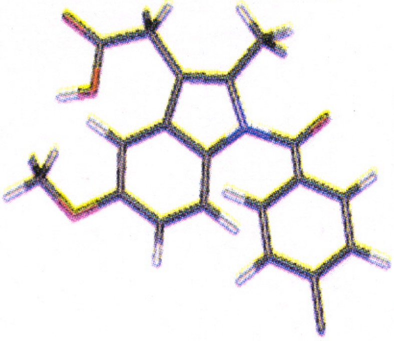


Fig. 2 — Generation of Indomethacin using template library[11] (Module 1) and molecule generator (Module 2). Here indomethacin is synthesized *in silico* in 6 steps. Templates i02 and s11 are linked in step I to form an intermediate, which is then linked with the template s01 in step II. The intermediate formed is now linked with template s13 in step III to form an intermediate, which is then linked with l02 to form an intermediate in step IV. In step V, the intermediate formed in step IV is linked with m01 in step V and the resultant intermediate is finally linked with s02 in step VI to form indomethacin.

Table 3 — Indomethacin molecule file generated by the molecule generator module

```
ATOM      1 C1   DRG     1      23.123  23.209  14.940 0 CA  1.9080  0.0860 -0.1989
ATOM      2 C2   DRG     1      21.973  22.594  15.435 0 CA  1.9080  0.0860 -0.1963
ATOM      3 C3   DRG     1      22.053  21.643  16.466 0 CA  1.9080  0.0860  0.2296
ATOM      4 C4   DRG     1      23.261  21.401  17.135 0 CA  1.9080  0.0860 -0.1536
ATOM      5 C5   DRG     1      24.411  22.001  16.648 0 CA  1.9080  0.0860 -0.0688
ATOM      6 C6   DRG     1      24.361  22.851  15.479 0 CA  1.9080  0.0860 -0.0233
ATOM      7 N7   DRG     1      25.692  23.196  15.137 0 NC  1.8240  0.1700 -0.0379
ATOM      8 C8   DRG     1      26.554  22.619  16.112 0 CA  1.9080  0.0860  0.0417
ATOM      9 C9   DRG     1      25.803  21.918  17.053 0 CM  1.9080  0.0860 -0.0786
ATOM     10 H10  DRG     1      23.048  23.958  14.139 0 HA  1.4590  0.0150  0.1633
ATOM     11 H11  DRG     1      20.996  22.875  15.010 0 HA  1.4590  0.0150  0.1424
ATOM     12 H12  DRG     1      23.287  20.715  17.994 0 HA  1.4590  0.0150  0.1495
ATOM     13 O13  DRG     1      21.037  20.797  16.863 0 OS  1.6837  0.1700 -0.3215
ATOM     14 C14  DRG     1      19.694  21.205  16.672 0 CT  1.9080  0.1094 -0.0680
ATOM     15 H15  DRG     1      19.402  21.931  17.471 0 HC  1.4870  0.0157  0.0805
ATOM     16 H16  DRG     1      19.120  20.247  16.784 0 HC  1.4870  0.0157  0.0965
ATOM     17 H17  DRG     1      19.518  21.640  15.660 0 HC  1.4870  0.0157  0.0791
ATOM     18 C18  DRG     1      26.271  21.288  18.295 0 CT  1.9080  0.1094 -0.0603
ATOM     19 C19  DRG     1      26.233  19.790  18.281 0  C  1.9080  0.0860  0.7494
ATOM     20 O20  DRG     1      25.428  18.995  17.795 0  O  1.6610  0.2100 -0.5886
ATOM     21 O21  DRG     1      27.263  19.240  18.989 0 OH  1.7210  0.2104 -0.6970
ATOM     22 H22  DRG     1      25.616  21.613  19.152 0 HC  1.4870  0.0157  0.0863
ATOM     23 H23  DRG     1      27.322  21.610  18.536 0 HC  1.4870  0.0157  0.0789
ATOM     24 H24  DRG     1      27.176  18.274  18.989 0 HO  0.0000  0.0000  0.4707
ATOM     25 C25  DRG     1      28.026  22.723  16.058 0 CT  1.9080  0.1094 -0.2449
ATOM     26 H26  DRG     1      28.354  23.739  16.392 0 HC  1.4870  0.0157  0.1069
ATOM     27 H27  DRG     1      28.411  22.567  15.018 0 HC  1.4870  0.0157  0.1044
ATOM     28 H28  DRG     1      28.485  21.947  16.722 0 HC  1.4870  0.0157  0.0969
ATOM     29 C29  DRG     1      26.168  23.978  14.062 0  C  1.9080  0.0860  0.5454
ATOM     30 O30  DRG     1      27.194  24.658  14.246 0  O  1.6610  0.2100 -0.5200
ATOM     31 C31  DRG     1      25.516  23.949  12.723 0 CA  1.9080  0.0860 -0.0130
ATOM     32 C32  DRG     1      24.581  22.989  12.326 0 CA  1.9080  0.0860 -0.1840
ATOM     33 C33  DRG     1      24.047  23.017  11.038 0 CA  1.9080  0.0860 -0.0638
ATOM     34 C34  DRG     1      24.474  23.981  10.124 0 CA  1.9080  0.0860 -0.0098
ATOM     35 C35  DRG     1      25.443  24.923  10.495 0 CA  1.9080  0.0860 -0.0556
ATOM     36 C36  DRG     1      25.953  24.908  11.791 0 CA  1.9080  0.0860 -0.1550
ATOM     37 H37  DRG     1      24.261  22.195  13.018 0 HA  1.4590  0.0150  0.1719
ATOM     38 H38  DRG     1      23.287  22.270  10.750 0 HA  1.4590  0.0150  0.1315
ATOM     39 H39  DRG     1      25.808  25.667   9.771 0 HA  1.4590  0.0150  0.1298
ATOM     40 H40  DRG     1      26.718  25.638  12.096 0 HA  1.4590  0.0150  0.1610
ATOM     41 C141 DRG     1      23.823  24.007   8.558 0 CL  1.9480  0.2650 -0.0766
CONECT    1    2    6   10
CONECT    2    1    3   11
CONECT    3    2    4   13
CONECT    4    3    5   12
CONECT    5    4    6    9
CONECT    6    5    1    7
CONECT    7    6    8   29
CONECT    8    7    9   25
CONECT    9    8    5   18
CONECT   10    1
CONECT   11    2
CONECT   12    4
CONECT   13   14    3
CONECT   14   13   15   16   17
CONECT   15   14
CONECT   16   14
CONECT   17   14
CONECT   18   19    9   22   23
CONECT   19   18   20   21
CONECT   20   19
CONECT   21   19   24
CONECT   22   18
CONECT   23   18
CONECT   24   21
CONECT   25    8   26   27   28
CONECT   26   25
Contd.
```

```
Contd.
CONECT   27   25
CONECT   28   25
CONECT   29   30    7   31
CONECT   30   29
CONECT   31   32   36   29
CONECT   32   31   33   37
CONECT   33   32   34   38
CONECT   34   33   35   41
CONECT   35   34   36   39
CONECT   36   35   31   40
CONECT   37   32
CONECT   38   33
CONECT   39   35
CONECT   40   36
CONECT   41   34
END
001002004   0      0   0
001006004   0      0   0
001010001   0      0   0
002003004   0      0   0
002011001   0      0   0
003004004   0      0   0
003013001   0      0   0
004005004   0      0   0
004012001   0      0   0
005006004   0      0   0
005009001   0      0   0
006007001   0      0   0
007008001   0      0   0
007029001   0      0   0
008009002   0      0   0
008025001   0      0   0
009018001   0      0   0
013014001   0      0   0
014015001   0      0   0
014016001   0      0   0
014017001   0      0   0
018019001   0      0   0
018022001   0      0   0
018023001   0      0   0
019020002   0      0   0
019021001   0      0   0
021024001   0      0   0
025026001   0      0   0
025027001   0      0   0
025028001   0      0   0
029030002   0      0   0
029031001   0      0   0
031032004   0      0   0
031036004   0      0   0
032033004   0      0   0
032037001   0      0   0
033034004   0      0   0
033038001   0      0   0
034035004   0      0   0
034041001   0      0   0
035036004   0      0   0
035039001   0      0   0
036040001   0      0   0
```



performance computing environment to generate millions of molecules using the template library in a combinatorial fashion scripted through UNIX shells. The generated candidate molecules carry information on: (i) Cartesian coordinates of all atoms, (ii) non-bonded parameters, (iii) bond connectivity, and, (iv) sites for further linkage.

Generation of indomethacin, an NSAID, is illustrated in Fig. 2 and the corresponding molecular file is shown in Table 3. The source code has been written in FORTRAN and tested on LINUX and SOLARIS platforms. Using the template library and molecule generator program, we are attempting to build a database of synthesizable compounds. The molecules so generated could be used for any biological target after screening them through drug-like filters discussed in the next section.

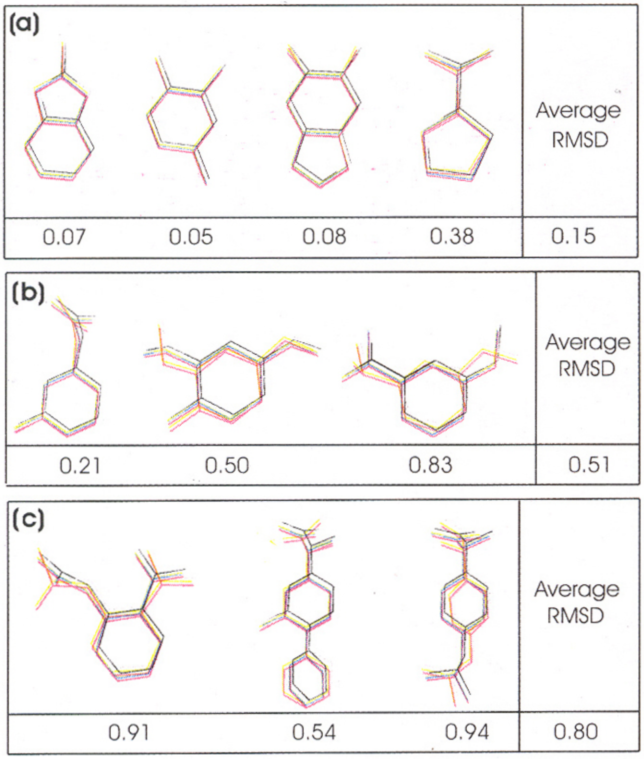


Fig. 3 — Molecular structural formulas of some small molecules generated using molecule generator (Module 2), AM1 geometry optimized (red) and compared with their crystal structures from RCSB (blue), shown along with the calculated Root Mean Square Deviations. The number of rotatable bonds are: (a) 0 to 1, (b) 2 to 3 and (c) > 3.

The output of the molecular generator module, viz., the structure of the molecule *inter alia* is compared with experimental data where available (Fig. 3). The average root mean square deviations of the molecules synthesized *in silico* range from 0.15 Å to 0.80 Å depending on the number of rotatable bonds in relation to crystal structures providing a validation of the *ab initio* molecule generation protocol adopted.

Module 3: Molecular descriptors and drug-like filters

A successful lead discovery strategy must ensure bioavailability from the very start in generating leads while eliminating wrong candidates from consideration. In *Sanjeevini,* we have introduced some empirical computational filters based on 'drug-like' properties/molecular descriptors of known drugs[21-23]. These include: (i) molecular weight, (ii) molecular volume, (iii) number of hydrogen bond donors and acceptors, (iv) log *P*, (v) molar refractivity (MR), and (vi) rotatable bonds. These molecular descriptors could act as computational filters based upon their accepted limits (Table 4) to screen the candidate compounds (Table 5). Introduction of filters facilitates computational tractability by restricting the chemical space for potential candidates, saving much time and cost in new lead discovery. However, there are drugs within the therapeutic classes of antibiotics,

Table 4 — Molecular descriptors/drug like filters adopted in *Sanjeevini*

Lipinski's Rule of Five[24]	Acceptable limits
Molecular weight	≤ 500
Number of hydrogen bond acceptors	≤ 10
Number of hydrogen bond donors	≤ 5
log *P*[25]	≤ 5
Additional filters	
Molar refractivity (MR)[25]	≤ 140
Number of rotatable bonds[26]	≤ 10

Table 5 — Molecular descriptors of Indomethacin as reported by *Sanjeevini*

Drug	Molecular weight	Molecular volume	No. of H-bond donors	No. of H-bond acceptors	log *P*	Molar refractivity (MR)	No. of rotatable bonds
Indomethacin*	357 daltons	162.9 $Å^3$	1	5	3.2	95.9	4

*Molecular structural formula of indomethacin is shown in Table 3 as an inset.

antifungals, vitamins and cardiac glycosides which do not fall in the acceptable limits of these drug-like filters[24]. *Sanjeevini* provides the option to bypass the filters selectively if desired. Molecular weight is the sum of all the atoms in a molecule and molecular volume is computed via a grid-based method. Sum of OH and NH groups is counted towards number of H-bond donors and sum of the number of N and O is counted towards number of H-bond acceptors. Calculation of log *P* and molar refractivity (MR) are based upon Crippen's parameters[25]. Rotatable bond is defined as any single bond, not in a ring, bound to a non-terminal heavy (i.e. non-hydrogen) atom. An illustrative calculation is reported in Table 6. All the properties have been verified to yield values as reported in the literature. The source codes for these molecular descriptors/drug-like filters are written in FORTRAN, C and C++ and utilize the CONNECT and MOLE information of the input molecular file (from Module 2). Work is in progress to develop chiral centre and ring strain filters, ADMET filter and *in silico* checks on the ease of synthesizability and shelf life.

Table 6 — Experimental[27] (E) and calculated (C) log *P* values for some NSAIDs

Drug	(E) log *P*	(C) log *P*
Diclofenac	4.4	3.328
Flurbiprofen	4.16	2.535
Ibuprofen	3.51	2.036
Indomethacin	3.08	3.182
Ketoprofen	3.12	2.069
Mefenamic acid	5.12	3.707
Naproxen	3.34	2.003
Piroxicam	1.8	1.629

To explore the conformational space of the candidate molecule and to arrive at its equilibrium geometry in its free and bound states, a link between structure and energy needs to be established. This necessitates geometry optimization, assignment/derivation of partial atomic charges as well as other force field parameters to the candidate molecule. The filtered candidate molecules are geometry optimized by semi-empirical AM1 calculations[17] using GAMESS quantum mechanics package[14]. Partial atomic charges of each candidate are re-derived by HF/6-31G*/RESP procedure[14,18] using GAMESS[14] and AMBER[13] molecular modeling package. The bond length, bond angle, dihedral parameters and the Lennard-Jones (12, 6) parameters are assigned in a force field compatible manner[19]. We are simultaneously working towards developing rules for charge transferability and incorporating an energy minimizer in Module 2 to speed up the process.

Module 4: Molecular docking

That drug activity is obtained through the molecular binding of one molecule (the ligand) to the active site of another molecule (the receptor), which in a majority of cases is a protein was proposed as a concept as early as in 1870 by Langley[28]. In their binding conformations, the molecules exhibit geometric and chemical complementarity, both of which are essential for successful drug activity. Computer-aided methods involve two steps at this stage: docking and scoring. Docking is aimed at predicting the binding modes of a ligand in the active site of a molecular target by generating multiple configurations possible of the ligand. Scoring produces an estimate of the binding affinity between the target and the ligand for each generated configuration. Computational strategies for docking to study the formation of stable intermolecular complexes had been the subject of intense research since the days the 3D structures of the targets became available and the issues encountered in designing docking algorithms were thoroughly reviewed[29-34]. The most systematic approach is to search through all binding orientations of all conformations of the ligand and receptor. Two major classes of automated searching are geometric methods that match ligand and receptor site descriptors and energy-driven searching based on molecular dynamics (MD) and Monte Carlo (MC) simulations. MC approach involves a random perturbation of the ligand in the active site enabling the algorithm to escape from getting trapped in local minima. Some of the softwares that employ MC methods in different forms include MCDOCK, PRODOCK and PRO_LEADS, ICM, SCVMC.

In *Sanjeevini*, we have designed a docking algorithm that is based upon a Monte Carlo sampling procedure in the active site in the 6D space of the candidate in the coordinate framework of the target biomolecule. For an enhanced sampling of the space, about 100 random rotations are generated per translation resulting in a rigorous search around the point of translation. The scoring function used is an empirical potential energy function which considers electrostatic and van der Waals interactions at the atomic level between the ligand and the target molecule and includes solvent implicitly in the electrostatics and in a hydrophobic term[35,36]. All atom energy calculations are performed for scoring the different configurations of the ligand in the active site of the protein during the Monte Carlo run. The algorithm implements a parallel search approach, making it faster and provides scope for a detailed exploration in an acceptable time frame negotiating an acceptable level of tradeoff between rigor and intensity of the docking run. The docking protocol as a novel idea divides the job among multiple processors, accumulates the results as independent intermediate solutions coming from each processor, minimizes each one of them and picks up the protein-candidate complex with the best score (lowest interaction energy). Figure 4 shows one such result in which indomethacin (blue) was docked in the active site of cyclooxygenase-2 (COX-2). The RMSD obtained between the docked and the crystal structure was 0.2 Å providing a proof of concept of the docking methodology.

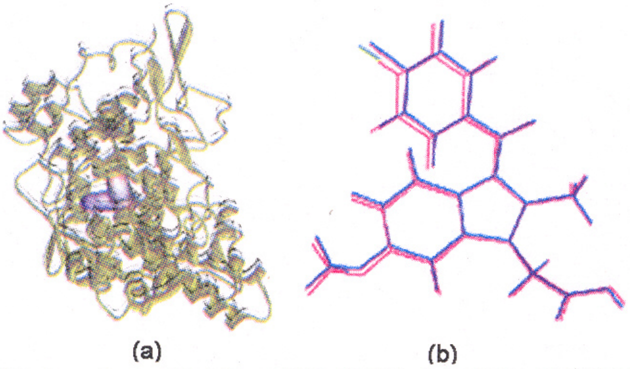


Fig. 4 — A representation of the docked and energy minimized structure of: (a) COX-2–indomethacin complex from Module 4; and, (b) Result of candidate indomethacin (blue) docked in COX-2 shown along with the crystal structure (4cox.pdb[37]) in the same coordinate frame.

Module 5: Structural analysis of the receptor-candidate complex

Considering the complexity of the search for and in the active site and the importance attached to a good scoring function, a mechanism to validate the final docked structure with the crystal structure is essential. At the same time for a better understanding of the interactions contributing to the activity of a particular molecule in its bound state, an in-depth analysis of the receptor candidate complex is also essential. With the availability of protein-ligand databases[38,39], it is now possible to test docking protocols on different systems contributing to a more rigorous screening of the proposed protocols and in turn improving the internal functioning of the algorithms. One such criterion is to find the RMSD between the candidate predicted after docking exercise and the corresponding crystal structure. *Sanjeevini* provides the user with an option to check the RMSD between the proposed docked candidate and the experimental structure provided. Figure 4b is generated from one of the routines in this module. Apart from handling ligand-receptor complexes, *Sanjeevini's* Module 5 has additional capabilities enabling users in superimposition of two stand alone molecules and calculating RMSD between them using single value decomposition procedure. The same module extends to accommodate structural analysis of a given receptor-candidate complex. Some of the questions that can be answered by this module are: (i) whether the hydrogen bond donors and acceptors are aligned appropriately, (ii) whether there is any hydrogen bond formation between the candidate and the protein, (iii) whether the charged residues are aligned appropriately, (iv) how much percentage of the drug is clashing (as the case might be) with the protein, (v) list of clashing points both on the protein and the candidate, etc. The answers can be elicited from the module independent of the pathway.

Module 6: Energy analysis of the receptor-candidate complex

Estimating binding free energies accurately is a time consuming process. The need for a fast, yet accurate, scoring function for docking studies has led to a number of different scoring functions that can be divided into three main classes, namely first-principles methods, empirical methods and knowledge-based methods[31]. We have incorporated an empirical interaction energy module in *Sanjeevini*. The input structure can be a protein-ligand X-ray crystal structure from RCSB[20] or a docked complex from Module 4 of *Sanjeevini* with hydrogens added and preferably energy minimized. Interaction energy is computed as a pair wise sum of three components between the protein and the ligand atoms: electrostatics, van der Waals and hydrophobicity. A sigmoidal dielectric function is adopted for electrostatics and hydrophobicity is computed via Gurney function. The performance of this energy function has been examined previously[35,36]. Figure 5 shows a residue-wise decomposition of the interaction energy between COX-2 and indomethacin. The drug carries a carboxylate group and a series of hydrophobic moieties. Figure 5 clearly depicts the reported preferred interactions which make indomethacin a good drug[37]. ARG121 forms a strong hydrogen bond interaction with the anionic groups of indomethacin. TYR356, VAL524 and ALA528 form a hydrophobic pocket in the active site and the large hydrophobic moiety of the indomethacin fits very well in this region showing large favorable contributions towards interaction energy. Amino acid residues help the target in folding and binding, e.g. they play a role in both the structure and function of the target. A residue, which is identified as proximal to the ligand in a structural analysis, may or may not contribute favorably to binding/interaction. This module helps in highlighting all the favorable interactions, which can be utilized effectively in the optimization of the candidate.

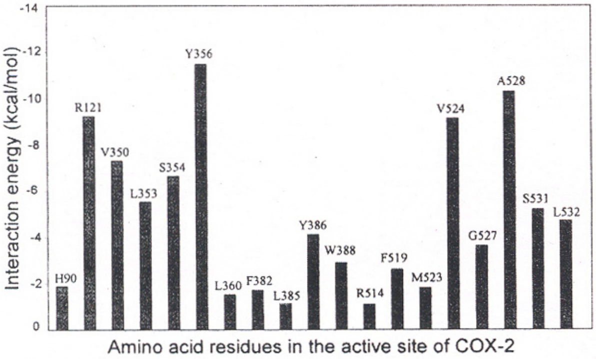


Fig. 5 — Residue-wise energy analysis of the interaction between COX-2 and indomethacin. Single letter codes are used for amino acids. Residues are numbered as in Ref. 37.

The source code for this Module has been written in C and tested on LINUX and SOLARIS platforms. Further calibration of the interaction energy function against experimental binding free energies using the available binding data of protein-ligand complexes has resulted in a correlation coefficient of $r = 0.92$ (Ref. 52).

Module 7: Binding affinity analysis

Computation of absolute binding free energies from atomic level descriptions of the systems is a formidable task[40-50]. In a phenomenological view, the net binding free energy may be considered to be a sum of the free energy changes due to the following contributions: (i) van der Waals interactions between the protein and the inhibitor indicating the influence of shape complementarities and packing effects; (ii) net electrostatics which includes interactions between partial or full charges, hydrogen bonds and electrostatics of desolvation upon binding and added salt effects; (iii) cavitation effects, which account for

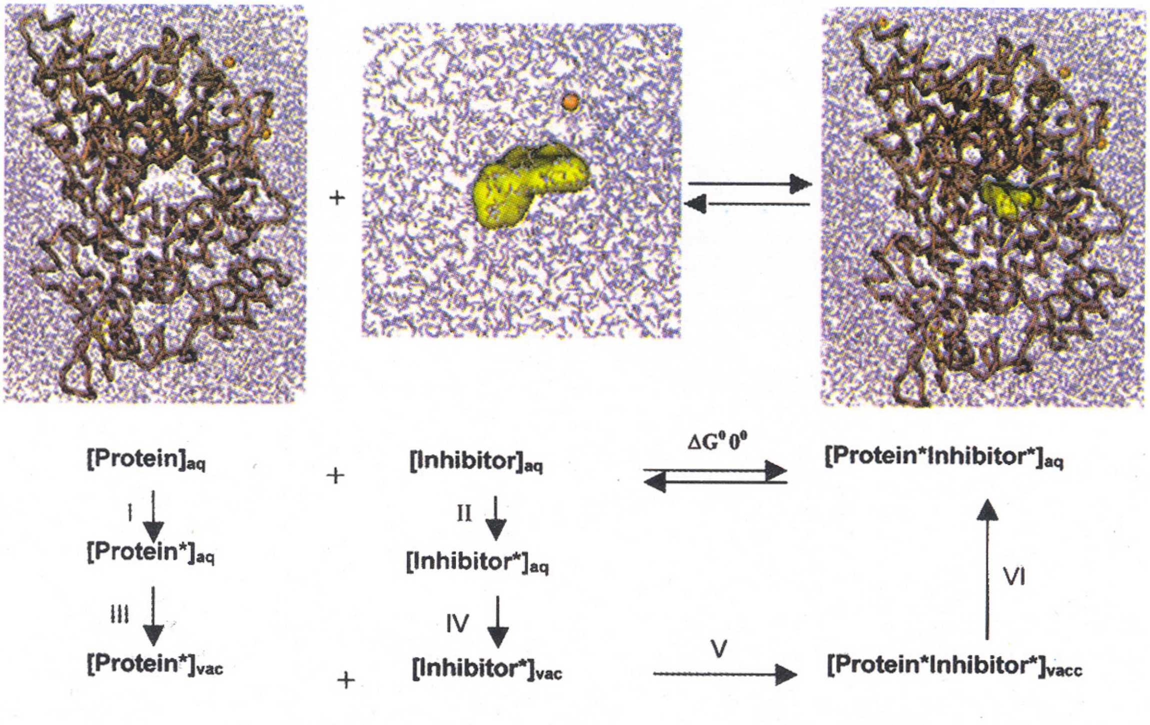


Fig. 6 — Thermodynamic cycle adopted for the computation of absolute binding free energies (ΔG°) for protein-ligand binding.

Table 7 — Free energy based component-wise analysis of the binding of COX-2 with aspirin (in kcal/mol)

Energy components	Analysis after energy minimization	Post facto analysis of molecular dynamics (2 nanosec) trajectories
van der Waals	- 21.3	-20.8
Net electrostatics	-13.3	-8.6
Cavitation	-3.4	-3.6
Entropy	22.5	23.9
Adaptation	0	3.7
Net binding free energy*	-15.5	-5.4
Experimental binding free energy	-5.9	

*The computed absolute binding free energies with current state-of-the-art methodology carry an uncertainty of the order of ±2 kcal/mol.

change in size and shape of solvent cavity on complexation giving rise to water reorganization, a component of which, originating from nonpolar sources, is the hydrophobic effect. Here the non-electrostatics of desolvation of both polar and nonpolar atoms is accounted for in the cavitation term; (iv) the structural deformation expense (i.e. the intramolecular contributions due to structural variations upon binding); (v) translational, rotational and vibrational, configurational entropy losses. We use a generalized master equation[51] in an energy component framework of MMGBSA method[40] following a thermodynamic cycle (Fig. 6) linking the initial unbound states to the bound state to obtain semi-quantitative estimates of the standard binding free energies. The computed binding free energies include enthalpies, entropies and solvation effects. An illustrative calculation is shown in Table 7. Such a comprehensive free energy component analysis is helpful in understanding the effect of optimizations to be made to the ligand. The qualitative estimates of binding free energies provide a quick means to check whether or not to pursue with a candidate molecule. We have previously demonstrated based on binding free energy calculations that, *Sanjeevini* could sort drugs from non-drugs for COX-2[11]. The source code to compute the components as well as the net binding free energies are written in FORTRAN and have been tested on SOLARIS platform.

To account for flexibility of the candidate and the target and to deal with solvent and salt effects in binding in a rigorous way, all atom molecular mechanical simulations form the current choice. Molecular dynamics simulations can be configured on the bound (protein-candidate molecule docked complex) and the unbound species (free protein and free ligand) with explicit solvent and small ions under ambient temperature and pressure conditions.

To obtain converged net binding free energies with the above protocols, averages over at least 100 structures or more of the free protein, the drug and the complex from molecular dynamics simulations (typically of length 2 nanoseconds or longer) with explicit solvent have to be developed. In addition, a force field compatible continuum solvent model for estimating the electrostatic component of solvation free energy of each structure is required. Furthermore, trajectory analysis programs for calculating intermolecular as well as intramolecular interaction energies, besides programs for calculating accessible areas, translational, rotational and vibrational/configurational entropies are necessary. In relation to the proposed pathway, this is one of the most compute-intensive steps and practical only for a select few promising candidates, which could be identified in the previous step via qualitative estimates of free energies on the docked complexes with a multitude of ligands. Table 6 shows a calculation of the free energy of binding between COX-2 and aspirin before and after molecular dynamics simulations. Configurational averaging implemented via molecular dynamics results in semi-quantitative estimates of the free energies besides accounting for flexibility of the target and the ligand as well as explicit solvent effects. Results on the binding energetics of aspirin to COX-2 indicate the significance of the molecular dynamics study, without which the predictions could go wrong. Molecular dynamics enable a reliable quantification of the structural deformation expense due to binding. Particularly important are the explicit solvent effects in the net electrostatics, which in the case illustrated, turn out to be unfavorable due to desolvation expense. It is interesting to note that ligand design based on electrostatic complementarity may not always be productive and may indeed lead to unpleasant surprises. Binding free energy based analysis of the structures generated from molecular simulations can provide enhanced insights into factors favoring strong binding.

Table 8 — CPU times* for various modules in *Sanjeevini*

Module	CPU times	
	Ultra SparcIII	PIV
Template library	Pre-generated database	
Molecule generator	0m0.024s	0m0.002s
Molecular descriptors/drug-like filters	0m0.084s	0m0.016s
Molecular weight	0m0.008s	0m0.001s
Molecular volume	0m0.020s	0m0.006s
Hydrogen bond donors and acceptors	0m0.016s	0m0.002s
log *P*	0m0.014s	0m0.001s
Molar refractivity	0m0.014s	0m0.001s
Rotatable bonds	0m0.012s	0m0.005s
Molecular docking (@ Nine processors)	21m15.338s	17m40.997s
Structural analysis of the receptor-candidate complex	0m0.779s	0m0.450s
Clash identification	0m0.573s	0m0.434s
RMSD calculation	0m0.070s	0m0.006s
Charge alignment identification	0m0.068s	0m0.005s
Donor/acceptor alignment identification	0m0.068s	0m0.005s
Energy analysis of the receptor-candidate complex	0m7.621s	0m3.775s
Binding affinity analysis	4m90.254s	

The time factors are given in minutes (m) and seconds (s). CPU times for all the modules are for single processor, except for Molecular docking (Module 4) which is implemented in parallel mode over nine processors. GAMESS[14] and AMBER[13] for quantum mechanical and molecular mechanics calculations respectively have been implemented. CPU time for AM1 geometry optimization is 2m7.000s and for HF/6-31G/RESP calculations it is 74m2.000s. For energy minimization it is 16m13.507s and for a 2 nanosecond molecular dynamics simulation on COX-2 aspirin binding comprising 22,442 atoms, the CPU time is 210 days.

Discussion

Development of *Sanjeevini* is strongly motivated by the following questions. If a small molecule to act as a lead could be synthesized *in silico*, how good are the geometries vis-à-vis crystal structures? Where do the molecules designed bind on the target, in what conformation and with what affinity? How to ensure that the *in silico* combinatorial attempts at generating lead-like molecules bear fruition, i.e. they possess proper ADMET profiles and target specificities? How to proceed with an automated optimization of a candidate so that it becomes lead-like? If lead design projects can accommodate these concerns, they will have delivered more than what can be expected in the context of drug discovery today. *Sanjeevini* is an earnest attempt to address these questions and to bring to bear an array of physico-chemical concepts in a biomolecular framework to facilitate lead design with desired affinity and specificity. Table 8 indicates the CPU times involved. The most time consuming steps are contained in Modules 4 and 7. Implementations in a high performance cluster environment of course ensure computational tractability.

The following improvements are envisioned in the subsequent versions of *Sanjeevini* to curtail the computational times and to improve its potential for lead molecule design: (i) incorporation of a preprocessor in Module 2 (Molecule generator) to attempt a directed synthesis *in silico* of candidates avoiding generation of millions of improbable candidates and to ensure a higher success rate; (ii) development of transferability rules for assigning partial atomic charges for organic molecules obviating quantum calculations for each candidate; (iii) a more efficient sampling of conformational space of the candidate molecules during docking in Module 4 via a restricted activation of conformational degrees of freedom around each rotatable bond of the candidates; (iv) improvements to the scoring function (Module 6) for better correlations with IC_{50} values; (v) an artificial intelligence based processor which utilizes the structural and energy information from Modules 6 and 7 for informed decision making on functional group mutations/optimization to be attempted on candidates. Some of these steps will hopefully bring the protocols to the desired time lines of one molecule per processor per minute. The more computationally demanding molecular dynamics simulations with explicit solvent and *post facto* free energy analyses can be reserved for propitious cases. As of now the pathway proposed (Fig. 1) is robust in itself but can accommodate further improvements.

The worldwide efforts on genomics and proteomics have given a significant boost to both experimental and computational methods to march towards personalized medicine. Mapping of the metabolic pathways for a choice of the target biomolecule to ensure minimal side effects, usage of bioinformatics tools for modeling structures of the target followed by molecular dynamics refinement of the target structures coupled with advances in computer-aided structure based drug design strategies assure us that "Gene to Drug *in silico*" is a realizable dream in the foreseeable future.

Conclusions

A pathway has been explored for making lead-like molecules to any biomolecular target *in silico* all the way from building blocks with the eventual goal of automating the process. *Sanjeevini* — an active site directed lead design software, is the result of the ongoing research in our laboratory. The protocols proposed and illustrated, combine in a natural way basic concepts in chemical bonding (generation of candidate molecules from templates), quantum mechanics (geometry optimization and charge derivation), classical mechanics (molecular mechanics and dynamics), statistical mechanics (configurational/ Boltzmann averaging) and thermodynamics (standard free energies of complex formation). The software can be fine-tuned at each stage to improve accuracies. The successes seen are indicators of the current state of computational chemistry and molecular theoretical biophysics in service of biology.

Acknowledgement

Funding from the Department of Biotechnology (DBT) is gratefully acknowledged.

References

1 *PAREXEL's Pharmaceutical R&D Statistical Sourcebook*, edited by M P Mathieu, (Blackwell, UK) 2001, p. 96.

2 Walters P W, Stahl M T & Murcko M A, *Drug Discov Today*, 3 (1988) 160.

3 *Comprehensive Medicinal Chemistry Database* (Molecular Design Limited, San Leandro, CA), http://www.mdli.com.

4 Dean P M, Zanders E D & Bailey D S, *Trends Biotech*, 19 (2001) 288.

5 Ghose A K, Viswanadhan V N & Wendoloski J J, *J Comb Chem*, 1 (1999) 55.

6 Ramstrom O & Lehn J M, *Nature Rev Drug Discov*, 1 (2002) 26.

7 Agrafiotis D K, Lobanov V S & Salemme V R, *Nature Rev Drug Discov*, 1 (2002) 337.

8 Klebe G J, *Mol Med*, 7 (2000) 141.

9 Kuntz I D, *Science*, 257 (1992) 1078.

10 Marrone T J, Briggs J M & McCammon J A, *Annu Rev Pharmacol Toxicol*, 37 (1997) 71.

11 Latha N, Jain T, Sharma P & Jayaram B, *J Biomol Sruct Dyn*, 21 (2004) 6.

12 Latha N & Jayaram B, *Drug Design Rev-Online*, 2 (2005) (in press).

13 Pearlman D A, Case D A, Caldwell J W, Ross W S, Cheatham III T E, DeBolt S, Ferguson D, Seibel G & Kollman P, *Comput Phys Commun*, 91 (1995) 1.

14 Schmidt M W, Baldridge K K, Boatz J A, Elbert S T, Gordon M S, Jensen J S, Koseki S, Nguyen K A, Su S, Windus T L, Dupuis M & Montgomery J A, *GAMESS J Comp Chem*, 14 (1993) 1347.

15 Bemis G W & Murcko M A, *J Med Chem*, 39 (1996) 2887.

16 Merlot C, Domine D, Cleva C & Church D J, *Drug Discov Today*, 8 (2003) 594.

17 Dewar M J S, Zoebisch E G, Healyand E F & Stewart J J P, *J Am Chem Soc*, 107 (1985) 3902.

18 Bayly C I, Cieplak P, Cornell W D & Kollman P A, *J Phys Chem*, 97 (1993) 10269.

19 Cornell W D, Cieplak P, Bayly C I, Gould I R, Merz K M, Ferguson D M, Spellmeyer D C, Fox T, Caldwell J W & Kollman P A, *J Am Chem Soc*, 117 (1995) 5179.

20 Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N & Bourne P E, *Nucl Acids Res*, 28 (2000) 235.

21 Walters P W & Murcko M A, *Adv Drug Del Rev*, 54 (2002) 255.

22 Clark D E & Pickett S D, *Drug Discov Today*, 5 (2000) 49.

23 Ajay, Walters W P & Murcko M A, *J Med Chem*, 41 (1998) 3314.

24 Lipinsky C A, Lombardo F, Dominy B W & Feeney P J, *Adv Drug Del Rev*, 23 (1997) 3.

25 Wildman S A & Crippen G M, *J Chem Inf Comput Sci*, 39 (1999) 868.

26 Veber D F, Johnson S R, Cheng H Y, Smith B R, Ward K W & Kopple K D, *J Med Chem*, 45 (2002) 2615.

27 Hadgraft J, Plessis, J D & Goosen C, *Int J Pharm*, 207 (2000) 31.

28 Gringauz A, *Introduction to Medicinal Chemistry* (Wiley – VCH, New York) 1997, p 32.

29 DesJarlais R L, Sheridan R P, Seibel G L, Dixon S & Kuntz I D, *J Med Chem*, 31 (1998) 722.

30 Morris G M, Goodsell D S, Halliday R S, Huey R, Hart W E, Belew R K & Olson A, *J Comput Chem*, 19 (1998) 1639.

31 Brooijmans N & Kuntz I D, *Annu Rev Biophys Biomol Struct*, 32 (2003) 335.

32 Wong C F & McCammon J A, *Annu Rev Parmacol Toxicol*, 43 (2003) 31.

33 Teague S J, *Nature Reviews Drug Discovery*, 2 (2003) 527.

34 Oshiro C M & Kuntz I D, *J Comput-Aided Mol Design*, 9 (1995) 113.

35 Arora N & Jayaram B, *J Phys Chem B*, 102 (1998) 6139.

36 Arora N & Jayaram B, *J Comput Chem*, 18 (1997) 1245.

37 Kurumbail R G, Stevans A M, Gierse J K, McDonald J J, Stegeman A R, Pak J Y, Gildehaus D, Iyashiro J M, Penning T D, Siebert K, Isakson P C & Stallings W C, *Nature*, 384 (1996) 644.

38 Roche O, Kiyama R & Brooks C L III, *J Med Chem*, 44 (2001) 3592.

39 Puvanendrampillai D & Mitchell J B O, *Bioinformatics*, 19 (2003) 1856.

40 Kalra P, Reddy V & Jayaram B, *J Med Chem*, 44 (2001) 4325 and references cited therein.

41 Simonson T, Archontis G & Karplus M, *Acc Chem Res*, 35 (2002) 430.

42 Soliva R, Almansa C, Kalko S G, Luque F J & Orozco M J, *J Med Chem*, 46 (2003) 1372.

43 Beveridge D L & DiCapua F M, *Annu Rev Biophys Biophys Chem*, 18 (1989) 431.

44 Noskov S Y & Lim C, *Biophys J*, 81 (2001) 737.

45 Kalra P, Das A, Dixit S B & Jayaram B, *Indian J Chem*, 39A (2000) 262.

46 Kalra P, Das A & Jayaram B, *Appl Biochem Biotech*, 96 (2001) 93.

47 Madhu Sudhan M S, Vishveshwara S, Das A, Kalra P & Jayaram B, *Indian J Biochem Biophy*, 38 (2001) 27.

48 Shaikh S A, Ahmed S R & Jayaram B, *Arch Biochem Biophy*, 429 (2004) 81.

49 Jayaram B, McConnell K J, Dixit S B, Das A & Beveridge D L, *J Comput Chem*, 23 (2002) 1.

50 Jayaram B, McConnell K J, Dixit S B & Beveridge D L, *J Comput Phys*, 151 (1999) 333.

51 Ajay & Murcko M A, *J Med Chem*, 38 (1995) 4953.

52 Jain T & Jayaram B, FEBS Letters, 579 (2005) 6659.