Multi-Objective Optimisation of the Protein-Ligand Docking Problem in Drug Discovery

A. Oduguwa
Cranfield University
Cranfield, Beds
MK43 0AL, UK
+44 (0) 1234 754073
{a.oduguwa,

A. Tiwari
Cranfield University
Cranfield, Beds
MK43 0AL, UK
+44 (0) 1234 754250
a.tiwari,

S. Fiorentino
Cranfield University
Cranfield, Beds
MK43 0AL, UK
+44 (0) 1234 750111
s.fiorentino,

R. Roy
Cranfield University
Cranfield, Beds
MK43 0AL, UK
+44 (0) 1234 754073
r.roy}@cranfield.ac.uk

ABSTRACT

The pharmaceutical industry is facing an ever-increasing demand to discover novel drugs that are more effective and safer than existing ones. The industry faces huge problem in improving its drug discovery and development processes since formerly used methods have shown their limits. Additionally, tests for safety of drugs are performed at the later end of the drug discovery pipeline instead of earlier. Therefore, the industry is looking for predictive tools that would be useful in testing the behaviour of a drug candidate earlier on in the pipeline before performing the large scale clinical tests. This paper explores the application of evolutionary multi-objective optimisation techniques for achieving such predictive work in protein-ligand docking. The paper reviews the literature of multi-objective optimisation and the drug discovery process and proposes a framework as a predictive tool to calculate good docking configuration for a given target protein and its binding compound. Finally existing models for drug evaluation are used for framework validation.

Categories and Subject Descriptors

J [**Computer Applications**]: J.3 [Life and Medical Sciences] - *biology and genetics, health, medical information systems*.

General Terms

Experimentation

Keywords

Multi-objective Optimisation, Drug Discovery, Evolutionary Computing, Protein-Ligand Docking

1. INTRODUCTION

The discovery and development of new drugs for treatment and prevention of diseases is a huge endeavour. Several decades ago, newer drugs introduced to the market and the pharmaceutical industry enjoyed double digit growth rate and remarkable stability. Today, this is no longer the case. Only a small fraction of the drug discovery (DD) projects undertaken eventually lead to successful medicines. Such programmes can take 12 – 15 years and involve an average investment of $800 million dollars [6].

With the advent of new computational technologies, one of the main problems of DD is finding suitable drug compounds against a disease target. The rapid generation of lead compounds is a major hurdle in the design of therapeutics. An area of particular interest is the structure-based virtual screening techniques widely used in many DD efforts. One key methodology - molecular docking was pioneered in the early 1980s, and remains a highly active area of research [9]. This refers to computational predictions of the structures of ligand-protein complexes from the conformations of the ligand and protein molecules. This method has received increasing interest due to the availability of high-resolution structures of proteins and the automation of docking processes via a computer-based simulation. With these structures, computer-based methods can be used to identify or design ligands that possess good structural and chemical complementarities to various sites of the enzyme. It is however important to note that designing a drug based on the knowledge of the target protein structure as determined by current experimental techniques is prone to error. Two main reasons responsible for this are inaccuracies in the energy models used to score potential ligand/receptor complexes and the inability of the current method to account for the conformational changes that occur during the binding process for both the ligand and the protein. Although this problem has been partially resolved by introducing flexibility in ligands, however predicting the protein arrangement is still a very complex problem which has not been fully solved.

A lot of effort is being spent on reducing the protein-ligand docking problem to a single-objective optimisation problem by aggregation of all the energy terms. However, the weighting of the individual energy terms requires knowledge of the search space. This paper addresses this by presenting a multi-objective optimisation (MOO) approach that simultaneously minimises the different energy terms to generate Pareto solutions corresponding to the optimal protein-ligand docking configurations. A framework is developed as a predictive tool to calculate good docking configuration from a given target protein and its binding compound. This framework was then tested with three existing evolutionary multi-objective optimisation techniques and validated with three known complexes downloaded from the public domain.

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-186-4/06/0007…$5.00.

2. MULTI-OBJECTIVE OPTIMISATION

Real world problems are often characterised by several conflicting objectives [14]. Multi-objective optimisation (MOO) extends optimisation theory by permitting these objectives to be optimised simultaneously. The goal is to find a set of values for the design variables that simultaneously optimise several objective functions. The solutions are often referred to as Pareto optima, vector maxima, efficient points, or non-dominated solutions.

Currently, over 30 mathematical programming techniques exist that are designed for MOO. However, most of these techniques generate elements of the Pareto optimal set one at a time [2]. And so they are usually very sensitive to the shape of the Pareto front which means, for example, that they do not work when the Pareto front is concave or when the front is disconnected. Evolutionary computing (EC) is particularly suitable to solve MOO [2, 5]. It deals simultaneously with a set of possible solutions, the so-called population; it enables the generation of different members of the Pareto front at the same time. This allows the user to find several members of the Pareto optimal set in a single run of the algorithm. Moreover, EC algorithms are less affected by the shapes or continuity that the Pareto front can present. This means that contrary to other methods they can work on discontinuous or concave Pareto fronts without much problem [2]. Such flexibility is achieved due to the stochastic nature of these algorithms and the nature of the data manipulation operators used.

The challenge facing most solution methods is to ensure convergence of well-dispersed solutions close to the true optimal front. Three (NSGA-2, PAES, SPEA) state-of-the-art evolutionary MOO techniques are applied and compared in this study. These are discussed in detail in [2].

3. THE DRUG DISCOVERY PROCESS

The drug discovery (DD) is a process of developing drugs for the safe and effective treatment of a disease. The process identifies, evaluates, and optimises compounds and molecules with desired biological activity against a specified disease target or function [13, 21].

The DD process involves four major steps: target identification, target validation, lead identification and optimisation. This is depicted in figure 1. The process starts with identification of a disease target which originates from the discovery of a gene or from the elucidation of the molecular mechanism of a genetic defect (target identification). Actual methods of target identification aim at identifying genes and proteins related to diseases, and understanding how they differ between a healthy body and a diseased one. Once a potential target has been identified, two objectives have to be satisfied. First is to verify that the target is directly involved in a disease process (target validation). The second objective is to identify easily dockable areas on the target so that it is easily docked by a potential drug [11, 12, 15]. The next step is identifying potential drug compounds that could be optimised into drugs (lead identification). This involves searching a large compound database for new chemical entities that show positive impact on the target. The leads with positive response in the screening process are selected and optimised as potential drug candidates (lead optimisation). The result is a small number of compounds that proceed to clinical trials for development.

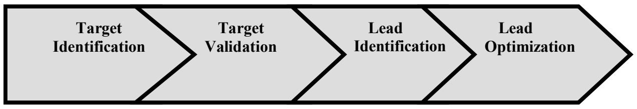


Figure 1. The drug discovery process.

The main focus of this paper is in the lead identification step where potential drug compounds are identified. Two approaches compete to achieve this goal: the random and the rational approach. Over the past decade the random approach was widely used in the industry, where high throughput screening (HTS) technology enables laboratories to synthesise, test and compare more than 10,000 compounds per day [6, 10, 17, 19]. Despite the huge number of compounds created, the number of new drugs marketed yearly did not increase after this technology was introduced. Therefore, the industry is now looking at rational and computer aided drug design, the so called ‘in silico’ tests. Indeed these recent years, many computer methods have been developed to test virtual compounds against a computer representation of the target in the docking process. This approach has presented many advantages; the main one is that it does not require the compound to be physically available. This helps in reducing cost and does not limit research to the compounds available within the company library. Furthermore, with the ever increasing power of computer processors, the throughput of these techniques has recently overtaken the one of high throughput screening techniques. However, rational drug design approach requires the following [6]:

- A quality structural model of the target
- Suitable model of interactions between proteins and their ligands
- Cross functional reiterative process to enable further investigations of promising compounds

The scope of this research work is however limited to the protein-ligand docking protocol used in virtual screening of compounds.

4. PROTEIN-LIGAND DOCKING

Three-dimensional molecular structure is one of the foundations of structure-based drug design. Often, data are available for the shape of a protein and a drug compound separately, but not for the two together. Docking is the process by which these two molecules fit together in 3D space. The docking process is a process of predicting the conformation of a ligand and its orientation within a targeted binding site [9]. It involves searching the possible binding configurations to identify potential docking states. It is generally devised as a multi-step process in which each step introduces one or more degrees of complexity. And despite improvements in computational power, docking remains a very challenging problem. Even with the fastest computers, many docking problems are still intractable. Exhaustive and systematic search methods are not always feasible even for the simplest of docking problems. Thus, this shows that docking is a search and optimisation problem, which necessitates a way of ranking potential configurations.

Two essential parts of docking are a scoring function and an efficient algorithm for searching conformation space. A good scoring function should be able to distinguish a correct binding mode from other putative modes. It is used to rank the bindings correctly. A conformational search method, docks conformationally flexible ligands by employing a simulation or optimisation method to search through the space of ligand–receptor configurations. Generally, these optimisation methods have the potential to identify a greater number and variety of known ligands. More recently, evolutionary algorithms (EA) have become a popular choice in molecule docking applications and performed better than algorithms in some applications. An EA is a generally adaptable concept for problem solving, especially well suited for solving difficult optimisation problems. It is based on ideas borrowed from genetics and natural selection. EA has been used to solve problems involving large search spaces, where traditional optimisation methods are less efficient. Two main subgroups of search methods under the heading of EA are: (i) genetic algorithm (GA) and (ii) evolutionary programming (GP) and evolution strategy (ES). Many EA-based methods [1, 3, 4, 20] have been developed to evaluate the docking ability of chemical compounds to a given protein. They differ in the nature of the chemical compound they try to dock to. While some approaches adopt ligand docking to proteins i.e. small chemical compounds that fit on the local areas of the protein, others emphasise docking full proteins together [7, 16]. However, even if the computational power required for ligand-protein docking and protein-protein docking differs, the methodology is similar.

MIAX [3] and GOLD [8] are examples of GA-based approaches for docking found in the literature. MIAX, a protein-protein docking algorithm, achieves flexibility through allowing conformation flips of the amino acid side chains. Thus the algorithm codes the torsional angles of the side chains of the amino acids. However, all side chains do not have the same flexibility potential; many of them have limited flexibility. Only the side chains that have high flexibility potential are coded in the chromosomes. The determination of side chain flexibility is performed thanks to statistical study carried out over the Brookhaven Protein Data Bank about the frequency and variability of the torsional angle of the 20 amino acid side chains. Hence the chromosome is coded as follows: amino acids are sequentially stored with their phi and psi angle and then the possible torsional angles of their side chain. The operation is repeated for the second molecule then the relative coordinates and angles of one molecule against the other are stored. The search is then performed by a single objective GA by evaluating the energy of the complex defined as [3]:

$$\Delta G = E_{hy} + E_{hb} + E_{elec} + E_{tor} + E_{desol}$$

Where E_{hy} is the hydrophobic interaction energy mostly the Van der Waal's energy, E_{hb} the hydrogen bond energy, E_{elec} the electrostatic energy, E_{tor} the internal torsion energy and E_{desol} the desolvation energy.

GOLD contrary to MIAX is a ligand-protein docking algorithm. It enables ligand flexibility and a partially protein flexibility, that is to say the protein is only flexible in the neighbourhood of the binding sites identified. The chromosome representation shares ideas with MIAX representation: it is composed of 2 binary strings; one for the protein, the other for the ligand, each byte of the string encoding a single rotatable bond. The fitness function contains 6 steps:

- Generation of the conformation of the ligand and the protein active site
- Placing of the ligand within the active site using a least square fitting procedure
- Calculation of hydrogen bonding energy
- Calculation of pair wise energy for the interaction between both molecules
- Calculation of the ligand internal energy
- Summation of the energy terms

Existing algorithms, including the ones discussed earlier show that a lot of effort is being spent on reducing the protein-ligand docking problem to a single-objective optimisation problem by aggregation of all the energy terms as the fitness function. This creates a number of problems. The weighting of the individual energy terms requires knowledge of the search space. Since this is a real world problem, the search space is unknown due to the complex interaction of the design variables. Estimating the weighting can distort the topology of the search space which may suggest deception for the search algorithm. Also, analysis of the energy model used for computational biology reveals the multi-objective nature of molecular behaviour prediction. This is evident in current energy field models such as CHARMM, MM3 or AMBER [18]. Indeed these models are in constant evolution, since any new molecule discovered is put against its model prediction, then the model is changed to fit the existing molecule. The advantage of the multi-objective approach in this case is to provide an energy model that relies only on equations and no tweaking of any kind so that no continuous 'update' of the model is required. Hence, this study proposes a MOO approach which simultaneously minimises the conflicting energy terms to generate Pareto solutions corresponding to optimal protein-ligand docking configurations.

5. APPLICATION OF MULTI-OBJECTIVE OPTIMISATION IN DOCKING

A framework is implemented in this study as a predictive tool to calculate 'good' docking configurations for a given target protein and its binding compound. This section presents a description of the framework.

5.1 Framework Development

The aim of the calculation is to optimise 3 objectives which include: internal energy of the compound, the protein-compound couple's Van der Waal's and electrostatic energy of interaction, and the shape complementarities. Optimisation is performed through the use of a user defined evolutionary technique chosen from the PAES, SPEA and NSGA-II algorithms. The framework is implemented in MS Windows environment using C++ programming language. It uses a single input '.ini' file that contains all the parameters it needs. This file links to three files that contain the information on the molecules to be analysed. It performs a series of calculations and outputs its results in the form of several files, one per solution. These files are text files in a standard format used for protein description. A description of the inputs/outputs, the encoding mechanism adopted for the chromosome structure and the energy terms used to model the docking problem within the framework are described as follows.

5.1.1 Inputs and Outputs

The framework uses three main input files. These three files are:

- The protein target input file – This file contains information about the target protein and was downloaded from the RCSB (Research Collaboratory for Structural Bioinformatics) Protein Data Bank. This is a free online database that is most widely used in Bioinformatics study.
- The protein target pockets input file – This file contains information about the location of the protein 'pockets'. These are the different binding sites that exist at the surface of the protein. The file is downloaded from the CastP web service.
- The docking compound input file - This file is also downloaded from the RCSB website. It contains information about the docking compound.

The software saves the output generated in a single file per solution. These solutions are possible docking configurations of the protein-compound couple containing atoms' coordinates and information in PDB (protein data bank) format.

The framework uses five C++ classes: Atom, Gene, Individual, Population and Solution. Figure 2 presents a summary of the data structure of the five classes. Within the input files, both molecules present their own axis systems. These axis systems are Cartesian systems, which is the most convenient and classical used for common object manipulations in 3 dimensions. For design purposes, these two axis systems are kept within the framework data structure.

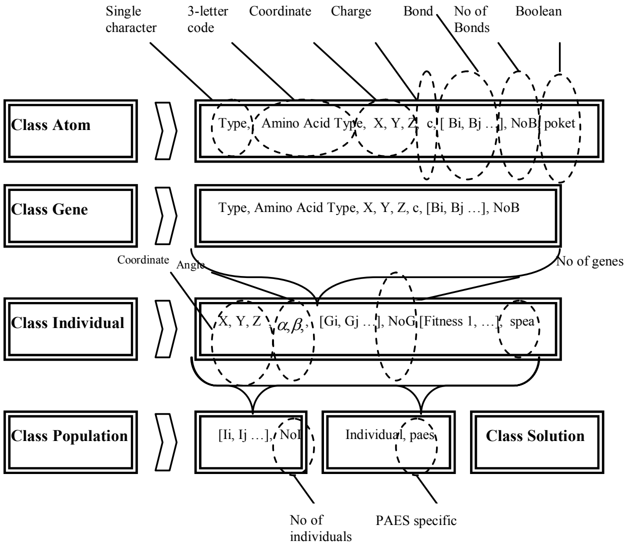


Figure 2. Summary of the class data structure.

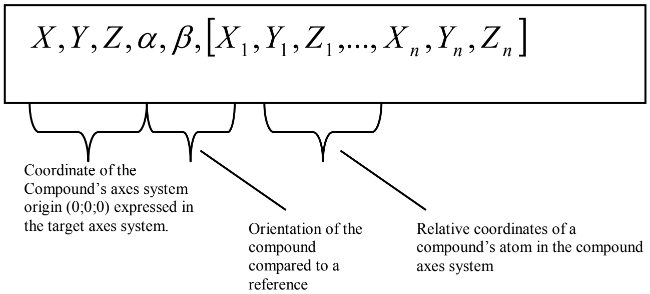

X, Y, Z, α, β, [X1, Y1, Z1, ..., Xn, Yn, Zn]
Coordinate of the Compound's axes system origin (0;0;0) expressed in the target axes system.
Orientation of the compound compared to a reference
Relative coordinates of a compound's atom in the compound axes system

Figure 3. Chromosome description.

5.1.2 The Chromosome Structure

The parameters described were used to design the chromosome structure (i.e. used by the algorithm to optimise the problem). The structure of the chromosome (Figure 3) consists of: 3 coordinates of the chromosome in the target axes system, 2 angles of the chromosome as compared to the reference compound, a set of the atoms' 3 coordinates (described by the 'Gene' objects) in its own axes system.

5.1.3 Energy Functions

In order to evaluate the structure of a molecule, the common way to achieve this is to calculate its potential energy. Due to protein size, the best way to calculate such energy is to look at quantum mechanics interactions within the considered protein, but such methodology is heavily computationally complex to be practical to model large systems. So as a compromise classical physics is used to come up with potential energy functions. These functions return a value for energy based on the conformation of the molecule. They provide information on what conformations of the molecule are better or worse. The lower the energy value, the better the protein structure. The main fields of energy used are bond, angle, torsion, Van der Waal's interaction, electrostatic and non-bonded energies [18]. In this framework, these energy functions are used to evaluate how good a conformation is. An outline of these energy terms is given as follows:

1. Bond strength energy – This corresponds to the stretching and compressing of the length of a bond. The simplest form is a quadratic equation:

$$E_{bonds} = K_b(r - r_0)^2$$

……………………….. …… **Equation 1**

With K_b is an empirically determined constant, r is the current bond length, and r_0 is the equilibrium bond length.

2. Bond angle energy – This corresponds to angle changes between bonds. Similar to bond length, the angles have an ideal angle value, and deviation from that creates energy. This is modelled by a simple quadratic equation:

$$E_{angles}=K_{\theta}(\theta-\theta_0)^2[1+7.10^{-8}(\theta-\theta_0)^4]$$

………..**Equation 2**

where K_{θ} is a constant value and θ_0 is the equilibrium angle.

3. Bond torsion energy: Torsions are created by three bonds, the middle one is rotatable. Usually it is described by a Fourier series expansion. The simplest being a single term:

$$E_{tor} = K_{tor}(1 \pm \cos(n\omega))$$

………………. ……...**Equation 3**

Where K_{tor} is a constant value, *n* the periodicity and *w* the angle.

4. Van der Waal's energy: This is a combination of repulsive and attractive forces. The repulsive force dominates when the distance between the atoms is small enough that the electron-electron interaction is strong. The attractive force occurs further away when there are fluctuations in the charge distribution within electron clouds. Van der Waal's energy is normally modelled using the Lennard-Jones 12-6 function:

$$E_{vdw} = K_{ij}\left[\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}}\right]$$

…………………………**Equation 4**

The "A" and "B" parameters control the depth and position of the potential energy for a given pair of non-bonded interacting atoms. Indeed, "A" determines the degree of "stickiness" of the Van der Waal's attraction and "B" determines the degree of "hardness" of the atoms. "r" is the distance between atoms.

5. Electrostatic energy – Electrostatic interactions are usually based on Coulomb's Law, assuming that atoms are point charges located at the atom's centre.

$$E_{elec} = \frac{q_i q_j}{Dr_{ij}}$$

………………………………. ……..**Equation 5**

Where $q_i q_j$ are the charges of the atom, r_{ij} the distance between the two atoms, and *D* the effective dielectric constant.

5.2 Step-by-Step Description of Framework

A brief description of the key steps followed in this framework is as follows and a flowchart of these steps is presented in Figure 4:

Step 1: Control of Input Data

This step verifies the input data to avoid problems during the computation; it especially checks if the input files exist and are valid. When all parameters have been checked, the framework starts coding the molecule information in the data structure.

Step 2: Saving the Molecules

The target protein is coded as a set of 'Atom' objects. The framework then reads in the target protein file, extract the atom coordinates, atom type and amino acid type of each atom found as an 'Atom' object. The charges of the different atoms are calculated according to the amino acid that the atom is associated with. Then the protein pockets input file is read, to identify which atoms of the set are parts of the pocket. The compound structure is then saved.

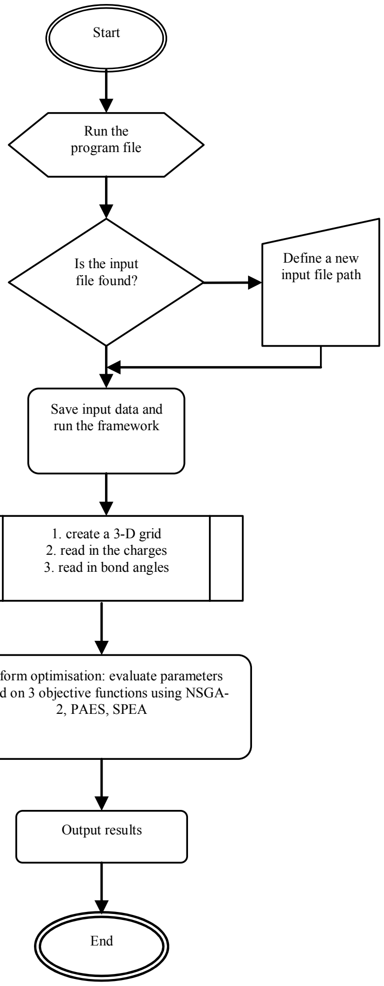

Start
Run the program file
Is the input file found?
Define a new input file path
Save input data and run the framework
1. create a 3-D grid
2. read in the charges
3. read in bond angles
Perform optimisation: evaluate parameters based on 3 objective functions using NSGA-2, PAES, SPEA
Output results
End

Figure 4. Flow chart of framework.

Step 3: Creation of a Grid

Due to the number of atoms in the data, and the molecular energy equation that needs to be applied to every atom, computational expense soon becomes an issue. As many equations require distance calculation, it is of prior interest to find an efficient and effective way to handle this. Since the contributions to the energy equations are always decreasing while distance increases, it is required that a given atom easily identifies its nearest neighbours without calculating every single atom-atom distance. A 3-D grid is created to address this. It enables each atom to look for neighbouring atoms first in its grid location and then move away from this location to look step-by-step in other locations. This process is extended until a preset distance limit (for which the energy contribution is considered negligible) is reached. Once the grid is created, each atom is given a location in the grid.

Such a 'griding' is also applied to the compound, the main difference being the fact that it has to be done each time a new solution is proposed by the framework. This grid is illustrated in Figure 5.

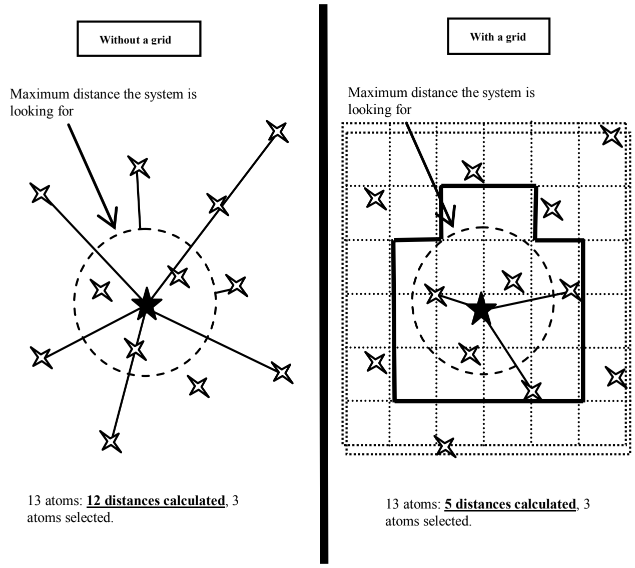

Without a grid
Maximum distance the system is looking for
13 atoms: 12 distances calculated, 3 atoms selected.
With a grid
Maximum distance the system is looking for
13 atoms: 5 distances calculated, 3 atoms selected.

Figure 5. Comparison of distance calculation with and without grid.

Step 4: Charge and Bond Angle Initialisation

This is the last step before invoking the evolutionary method. The target molecule and the compound have been saved, the coordinate information has also been saved. Then the charges and the bond angles between atoms are modified for every single amino acid. It is then extracted and read into the framework to perform energy calculations.

Step 5: Optimisation Process

Through the use of evolutionary method, the docking compound is tweaked. The compound is moved around the protein, to the internal position of its atoms to find a docking configuration (Figures 6 and 7).

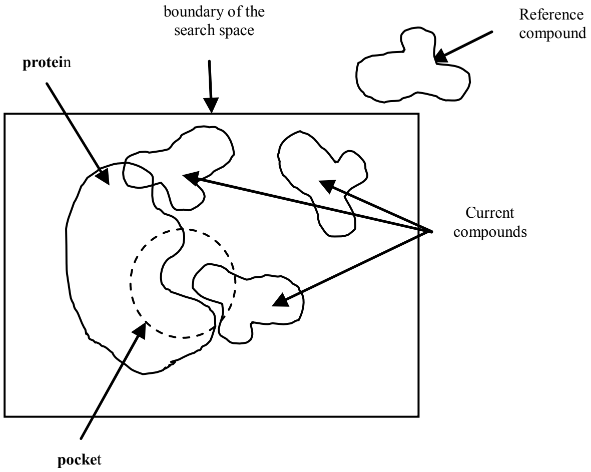


Figure 6. Evaluation of parameters.

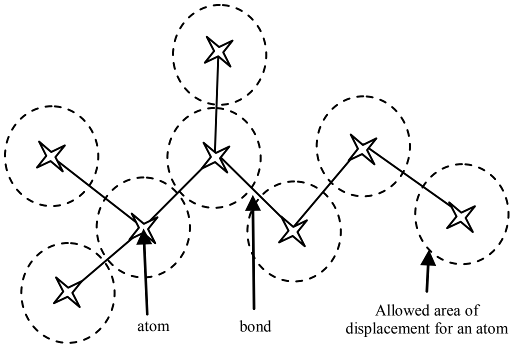


Figure 7. Reference compound and possible movements of its atoms.

Step 6: Evaluation

The framework is based on 3 objective functions that evaluate: (i) internal energy of the compound, (ii) interaction energy between the compound and the target molecule, and (iii) shape complementarities. These functions have to be minimised. The internal energy is made of 5 different energy terms; the bond strength energy, the bond angle energy, the bond torsion energy, the Van der Waal's energy and the electrostatic energy (equations 1 – 5). The first 3 energies are calculated for every single bond in the compound so that the resulting energies are sums of the respective values from the equations. The 2 last energies are calculated for every couple of atoms of the compound within a user set maximum distance. Again the resulting energies are sums of all these single energies. The interaction energy is made of 2 energy terms; Van der Waal's energy and electrostatic energy. Again these energies are calculated for couple of atoms within a user set maximum distance, but this time this couple is made up of one atom from the compound and one atom from the protein pocket.

5.3 Validation of Framework

The framework is tested for accuracy of its results. This is performed by testing the framework with complexes whose docking configurations are known and recorded in the public domain. The framework was tested using three evolutionary algorithms – PAES, SPEA and NSGA-II. The complexes used for this validation are:

- Ovomucoid (2ovo) docked to α chymotripsin (5cha) – 1CHO
- Human pancreatic trypsin inhibitor (1hpt) docked to α-chymotrypsinogen (1chg) – 1CGI
- Bovine pancreatic trypsin inhibitor (4pti) docked to trypsin (2ptn) - 2PTC

These complexes have been chosen because of their wide use in the literature. The tests have all been performed on a Pentium IV 2.66GHz computer. The size of population was set to 100 and the number of generations to 500. This corresponds to 50,000 fitness calculations. The solutions generated by the evolutionary optimisation algorithms are compared to the real complex using a RMSD (root mean square distance) calculation that is performed by the Qmol software that was downloaded from its authors' website (www.mbg.cornell.edu/Shalloway_Lab_QMOL.cfm).

6. RESULTS AND DISCUSSION

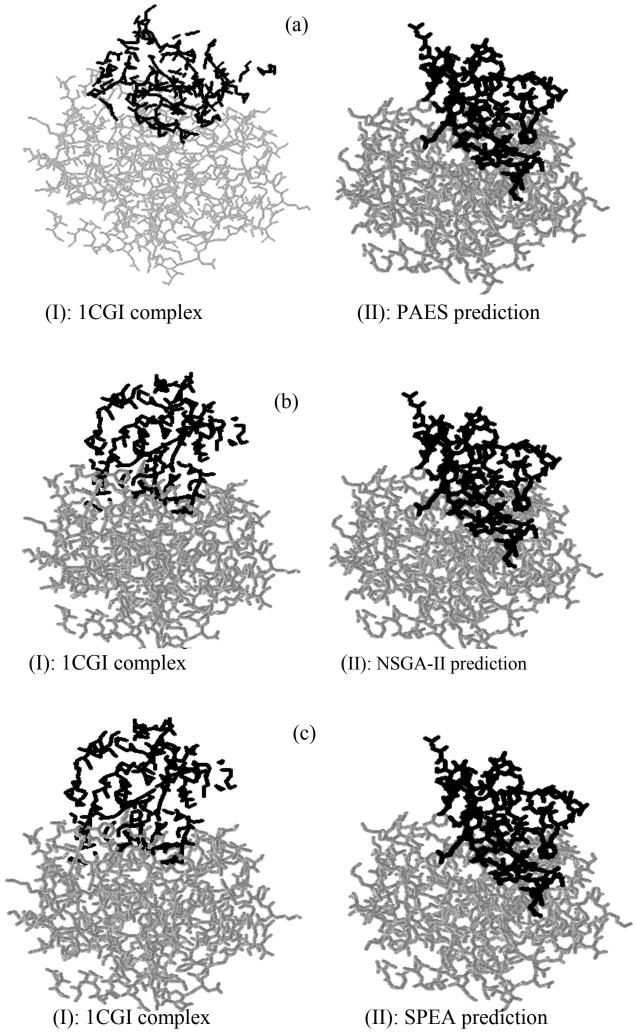
This section presents the results generated by the tests performed. Table 1 reports the RMSD values that compare the real complex with the complex predicted by the evolutionary optimisation algorithms. The smaller the RMSD value the higher the level of similarity between the real complex and the predicted complex. Table 1 gives the 3 best and 3 worst results obtained from the 9 runs performed.

Table 1: RMSD values

ET	PAES	SPEA	NSGA-II	PAES	SPEA	NSGA--II	PAES	SPEA	NSGA-II
Comp-lexes	1CGI			1CHO			2PTC		
3 best	8.089	8.125	7.952	10.234	12.31	11.456	9.571	10.245	9.845
	8.231	8.457	8.124	10.427	12.461	11.587	10.542	10.692	10.002
	8.943	9.023	8.571	10.568	12.463	11.869	10.852	10.845	10.425
3 worst	36.544	37.426	34.568	17.847	17.956	18.484	37.248	37.148	38.201
	37.236	39.523	35.623	18.215	18.524	18.878	37.869	37.956	38.204
	38.195	40.339	37.845	18.329	18.968	19.503	38.918	38.632	38.542
Average	18.396	19.527	17.854	15.256	16.538	15.683	20.467	21.527	20.546
Time (min)	581	520	534	602	637	728	607	586	617

Table 1 shows that the 'best' solutions predicted by each evolutionary algorithm (i.e. the ones with the smallest RMSD from the real complexes) tend to be similar. This is depicted by similar RMSD values corresponding to the 'best' solutions predicted by the three algorithms. However, PAES shows superior performance on two of the complexes (1CHO and 2PTC) and NSGA-II shows superior performance on one complex (1CHI). Same trends are shown by the average RMSD values from the 9 runs performed. The conclusion from these analyses is that the most accurate and robust performance on this problem is shown by PAES and NSGA-II. SPEA shows inferior performance on all the three complexes. As far as computational time is concerned it is interesting to notice that the three evolutionary techniques have the same computational expense since the total computation has less than 5% difference of the total times for the PAES, SPEA and NSGA-II. Indeed for this particular problem most of the time is spent to calculate the fitness functions. Therefore, the different data manipulations performed by the evolutionary techniques do not impact the computation time. Since published work does not provide information about the computation time of the docking algorithms, it has not been possible to compare the computation time required by the proposed framework to the previous work.

Figures 8a to 8c compare the real 1CGI complex and the best predicted complexes from the three MOO evolutionary algorithms. These 3-dimensional views of the complexes produced by the framework support the observation made earlier that the 'best' solutions predicted by each evolutionary algorithm tend to be similar. However, the differences between the real complex and the predicted complexes suggest that although the framework actually docks the compound to the target, but the final conformation differs from the real one since the binding site is not correctly defined. This may be due to limited accuracy of the pocket information provided by the CastP web service. It should be noted here that previous research efforts are predominantly manual. They are heavily based on trial-and-error guided by the experience of the biologists.



(I): 1CGI complex (II): PAES prediction

(I): 1CGI complex (II): NSGA-II prediction

(I): 1CGI complex (II): SPEA prediction

Figure 8. Known CGI complex and MOO prediction.

7. FUTURE RESEARCH ACTIVITIES

The limitations of the current research work and corresponding future research activities are listed as follows:

- It would be interesting to evaluate how the different parameters introduced by the framework impact the final results.
- Another future direction would be to perform a detailed comparative analysis of various Pareto solutions.
- It would also be useful to extend the predictive capability of the framework to the design of new compounds for docking.

8. CONCLUSIONS

An evolutionary multi-objective optimisation approach is presented to perform prediction of docking configurations for protein-ligand couples. Three evolutionary algorithms – PAES, SPEA and NSGA-II – are applied in this paper. The tests are performed on three complexes - 1CGI, 1CHO and 2PTC - whose docking configurations are known and recorded in the public domain. The results compare the real complex with the complexes predicted by the algorithms. The differences between the real complex and the predicted complexes suggest that although the framework actually docks the compound to the target, but the final conformation differs from the real one. It is observed that the most accurate and robust performance on this problem is shown by PAES and NSGA-II. SPEA shows inferior performance on all the three complexes. However, the 'best' solutions predicted by the three evolutionary algorithms and the corresponding computational times tend to be similar. The evolutionary multi-objective optimisation approach therefore shows promising results on this problem. This work will now be extended by developing the predictive capability in protein folding and the design of new compounds for docking.

9. ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the Engineering and Physical Research Council (EPSRC) and the Centre for Decision Engineering at Cranfield University, UK.

10. REFERENCES

[1] Clark, D. E. *Evolutionary Algorithms in Molecular Design. Methods and Principles in Medicinal Chemistry.* Wiley, GmbH , 2000.

[2] Deb, K. *Multi-objective Optimisation Using Evolutionary Algorithms.* John Willey & sons, Chichester, UK, 2001.

[3] Del Carpio, C. A., and Yoshimori, A. MIAX: A system for assessment of macromolecular interaction. *Genome Informatics, 11* (2000), 205 - 214.

[4] Eisenstein, M., and Katchalski-Katzir, E. On proteins, grids, correlations and docking. *C.R. Biologies, 327* (2004), 409-420.

[5] Goldberg, D. E. *Genetic Algorithms in Search, Optmization and Machine Learning.* Addison-Wesley, Massachussetts, 1989.

[6] Hillisch, A. and Hilgenfield, R. *Modern Methods of Drug Discovery*, Springer Verlag, Germany, 2003.

[7] Jackson, R. M., and Sternberg, J. E. A continuum model for protein-protein interactions: Application to the docking problem. *Journal of molecular Biology, 250* (1995), 258-275.

[8] Jones, G., Willet, P., et al. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology, 267* (1997), 727-748.

[9] Kitchen, D. B., Decornez, H., et al. Docking and scoring in virtual screening for drug discovery: Methods and application. *Nature Review Drug Discovery, 3* (2004), 935-948.

[10] Lyne, P. D. Structure-based virtual screening: An overview. *Drug Discovery Today, 7,* 20 (2002), 1047-1055.

[11] Maggio, E. T., and Ramnarayan, K. Recent developments in computational proteomics. *Drug Discovery Today, 6,* 19 (2001), 996-1004.

[12] Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology, 42* (1970), 245-261.

[13] Oduguwa, A., Tiwari, A. and Roy, R. An overview of soft computing techniques in drug discovery. In *9th online world conference on soft computing in industrial applications*, Online, 2004.

[14] Tiwari, A., Roy, R., Jared, G. and Munaux, O. Evolutionary-based techniques for real-life optimisation: Development and testing. *Applied Soft Computing (ASC) Journal, 1,* 4F (2002), 301-329.

[15] Oshiro, C. M., Kuntz, I. D., et al. Flexible ligand docking using a genetic algorithm. *Journal of Computer-Aided Molecular Design, 9,* 1 (1995), 113-130.

[16] Oyama, T., Kitano, K., et al. Mining association rules related to protein-protein interactions. *Genome Informatics* (2000), 358-359.

[17] Pegg, S. C. H., Haresco, J. J., et al. A genetic algorithm for structure-based de novo design. *Journal of Computer-Aided Molecular Design, 15* (2001), 911-933.

[18] Schleyer, P. V. R. (ed.) *Encyclopedia of Computational Chemistry.* John Wiley, 1998.

[19] Schneider, G., Clement-Chomienne, O., et al. Virtual screening for bioactive molecules by evolutionary de novo design. *Angewandte Chemie, 39* (2000), International Edition in English, 4130-4133.

[20] Taylor, J. S. and Burnett, R. M. DARWIN: A program for docking flexible molecules. *Proteins: Structure, Function and Genetics 41* (2000), 173-191.

[21] Verkman, A. S. Drug discovery in academia. *American Journal of Physiology - Cell Physiology, 286* (2004), C465-C474.