New Method of Calculating ^{SR}CM Chirality Measure

Przemyslaw Szurmak$^{(\boxtimes)}$ and Jan Mulawka

Artificial Intelligence Division, Warsaw University of Technology, Warsaw, Poland
przemyslaw@szurmak.pl, jmulawka@elka.pw.edu.pl
http://szurmak.pl

Abstract. Bioinformatics plays an important role in natural sciences. One of its branches – Computer-Aided Drug Design (CADD) – gives practical insights for designing and discovering of novel – better and safer – drugs. The CADD encompasses many different techniques like docking, virtual screening and quantitative structure-activity relationships (QSAR). The latter deals with building equations relating drug activities and their structures represented by variables called molecular descriptors. An important and promising type of such descriptors are Sinister-Rectus Chirality Measures (SRCM). However, the only so far available software for ^{SR}CM calculation is very slow, and this impedes wider application of ^{SR}CM by QSAR community. Therefore, an attempt to develop a novel algorithm for calculation of ^{SR}CM (using Genetic Algorithm) and to implement it in an efficient and modern computer program was made. The result of these efforts is Chirmes. Performed tests have shown that Chirmes gives correct results of ^{SR}CM calculations and performs way faster than the so far available software does. The paper describes first chemical and computational background behind the tackled problem. Then details of the implementation are presented, along with the test results and future prospects.

Keywords: Computer-aided drug design · QSAR · SRCM · Genetic algorithms · Cuda

1 Introduction

Although the XX century witnessed unprecedented development of medicine and chemical pharmacology, there are still many diseases for which safe and efficient therapies are lacking. The efforts for finding them are the domain of an inter-disciplinary research field called medicinal chemistry. In recent years, computer-aided drug design (CADD) has become an integral part of this research. CADD encompasses a number of methodologies that allow medicinal chemists to model the behaviour of drugs and their molecular targets. Thus it facilitates rational directing of expensive laboratory work. One of CADD techniques is Quantitative Structure-Activity Relationship (QSAR) analysis. The aim of QSAR study is to find a mathematical (quantitative) relationship between chemical structure and the medicinal activity. To this aim, equations are built that are of general structure:

$$activity = f(structure) \qquad (1)$$

Here, the structure is expressed as molecular descriptors, that is variables that describe a certain aspect of molecular structure., e.g. number of atoms of a given kind, number of flexible bonds, energy of molecular orbitals etc. [7] The choice of descriptors and the way they are related to activity may be knowledge-based or supported by special methods [2]. The obtained equations - if they are of good statistical quality - may be used to explain the observed experimental findings and to predict the behaviour of novel - yet unsynthesized molecules. In such a case, QSAR allows to save money and time and to increase the rate of drug discovery. Many drug molecules bear a special property - chirality. This means that they are not superposable on their mirror image. The chirality of drug molecules is an important and well-known problem in medicinal chemistry since some chiral molecules exhibit different activity and/or toxicity depending on which mirror image (left-handed or right-handed enantiomer) they are. There appeared an idea to use chirality - described quantitatively by Sinister-Rectus Chirality Measures - for QSAR modelling. The descriptor has been successfully applied several times [5], but a lack of a fast tool to compute the descriptor seriously hampered the use of this variable in QSAR modelling. The aim of our research was to fill this gap by designing and implementing a novel method for calculating Sinister-Rectus Chirality Measures.

This paper is organized as follows. Section 2 describes theory behind chirality measures. Section 3 discusses implementation details of the developed solution. Sections 4 presents the results of tests made using Chirmes, with comparison to Chimea. Finally, Sect. 5 presents our summary.

2   Theory

2.1   Chirality Measures

Chirality is a property of a three-dimensional shape of a molecule. The degree to which a drug is efficient is strongly connected to spatial fit of a drug molecule and its molecular target. Thus chirality can be used to describe structure of molecule. Chirality measures may be of use here. They are variables that describe how much chiral a molecule is, or in other words (according to the IUPAC definition [1]): how much non-superposable on its mirror image it is.

Out of many known chirality measures, this paper is focused on Sinister-Rectus (^{SR}CM) chirality measures [3,4]. They are calculated as follows: a mirror image of an analysed molecule is generated and then superimposed onto original molecule structure so that the superposition is optimal. The goodness of fit is rated based on normalized cartesian sum of distances between corresponding atoms in original and mirror molecules weighted depending on chosen property space. Mathematical equation describing ^{SR}CM reads:

$$^{SR}CM(A) = \frac{1}{a}min(\sum_{i=1}^{n} w_i d_i) \tag{2}$$

where a is normalization component (molecular mass), d_i – distance, w_i – weight (most often, property assigned to selected atom like electrical charge or atom mass), n – number of atoms in molecule.

A most important issue during ^{SR}CM equation solving is to find optimal superposition that would minimize the ^{SR}CM value. For achiral molecules their mirror images should be superposed ideally on original structures ($^{SR}CM = 0$), in case of chiral molecules the situation is more complicated though. Such molecule has infinite number of possible bad superpositions. If we also take into consideration that a typical molecule can contain up to several hundreds of atoms it can be clearly seen that a problem domain in such a case is enormous thus choosing a proper algorithm and implementing it in an efficient way plays a key role for solving the problem.

^{SR}CM chirality measures have been already used in a real-world scientific research, including studies on chiral heterofullerenes [4], modelling activity of steroids binding to sex-hormone binding globuline [3] and in analysis of Vibrational Circular Dichroism spectra [6]. Other chirality measures were also used to model activity of pain relief drugs, acetylcholinesterase inhibitors, behaviour of amino acids in plate chromatography, or for explaining of catalytic activity. Such numerous and versatile examples of applications show potential and necessity of developing software to calculate such descriptors in an efficient way which will help spreading usage of chirality measures in bioinformatics (especially in computer aided drug design) and in general – computational chemistry.

3   Implementation

3.1   General Application Structure

The main goal of the presented research was to develop an effective algorithm to calculate ^{SR}CM chirality measure and to implement such an algorithm as a working desktop application. The algorithmic problem is related to optimization problem during calculation of ^{SR}CM. After generating mirror image of molecules, the best superposition of mirror and base molecule needs to be found so that the value of (2) defining ^{SR}CM measure is the lowest.

During the problem discussion, the authors decided to apply genetic algorithms (GA). Main reason behind usage of GA was ease of implementation and well known as very universal and flexible method for problem solving, also in computer drug design. The detailed description is given in sections: Sect. 3.2 and 3.3.

One of the main targets for created software was performance. Because of that, the whole application was developed using C++ in its most recent specification - C++14.

Also, experiments were made to find out if implementation on GPU will gain performance increase. To allow simple integration of GPU working code with rest

of application whole program was divided to separate modules which can work independently. Another advantage of such approach is ease of allowing application to work on different operating systems such as Apple Mac OS, Microsoft Windows and Linux.

Developed software was named Chirmes from words **Chir**ality **Me**asure**s**. Figure 1 shows modular structure of application divided by work phases.
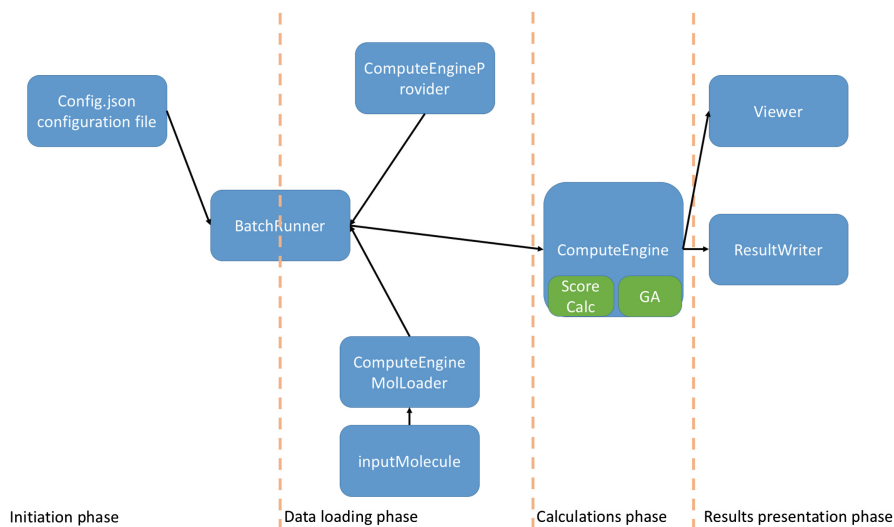

Fig. 1. Chirmes typical workflow chart with developed modules

First step of the work flow is loading configuration file which is then passed to *BatchRunner* module which. Separate class *ComputeEngineMolLoader*, loads input molecules and allocates (through helper *ComputeEngineProvider* object) chosen version of main calculation unit called *ComputeEngine* (implementation of *Abstract Factory* pattern). During loading of each molecule (from standard chemistry description file formats, more than 100 are supported) mirror images are created. Instance of *ComputeEngine* using implemented genetic algorithm (shown in Fig. 2) tries to find optimal superposition of molecule and its mirror image and then calculates ^{SR}CM value (presented in Fig. 3). Description of both processes can be found in Sects. 3.2 and 3.3.

3.2   Genetic Algorithm

Solving optimization problem (superposition of molecule and its mirror image) during calculation of ^{SR}CM is the most important part of Chirmes application. After loading molecule and its mirror image, a genetic algorithm is used, through series of rotation and translation applied to molecules to find combinations giving lowest value of chirality measure.
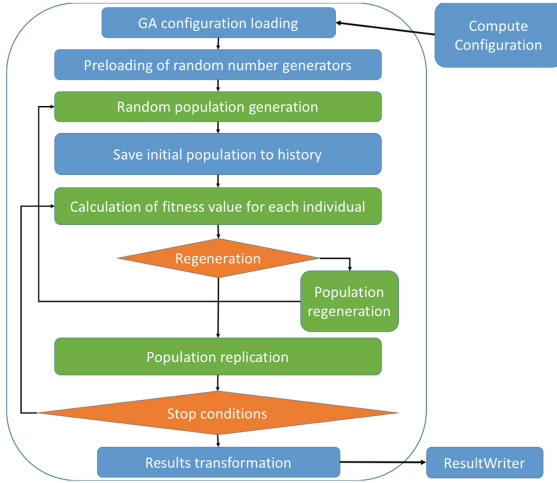
Fig. 2. Implemented genetic algorithm scheme

The application implements genetic algorithm in a typical form, overview of which is presented in Fig. 2. In order to tailor it better for the given problem, small improvements were made compared to the original idea of genetic algorithm. Gene coding is not binary but using floating point numbers[1] because they are mapping much better spatial coordinates for rotation, translation and, at the same time, providing better precision for such problem. Also, usage of binary coding would impose usage of much more complicated crossover and mutation operators to ensure correct solution domain (not every numeric value has meaning for spatial rotation or coordinate).

In order to achieve best performance, each gene (x, y, z for rotation and for translation) is coded using 32 bit floating point number instead of double precision. Main benefit of such approach is possibility to use vectorization support (SSE/AVX) provided by *Eigen* library which allows to make two times more calculations using single precision numbers than with double precision numbers at the same time. To prepare transformation matrix that converts genotype into phenotype following equation is used (3):

$$matrix = (translation * translationFromOrigin$$
$$* rotation * translationToOrigin) \tag{3}$$

Another modification of original genetic algorithm was implemented in process of random population generation. Right from the start values generated using *Mersenne Twister 19937* algorithm[2] are limited to only those having physical meaning – for example, translation components should not make

---

[1] Usage of real numbers was inspired by [8].
[2] Period of algorithm is $2^{19937} - 1$ which is more than enough for generating random population. Moreover, it is well known and optimized.

absolute distance between corresponding atoms greater than distance between geometrical center of molecule and its mirror image.

After generation of random population *Compute Engine* calculates value of chirality measure for each individual Fig. 3 (described in Sect. 3.3).

Next step is conditional population regeneration which is also another modification comparing to original genetic algorithm. This operator was introduced to respond more effectively in case of poor improvement of best individual score comparing to previous algorithm iterations. When such situation is discovered operator takes best individual from current population and, depending on chosen algorithm settings and current population situation, puts it into new population, created randomly from scratch without memory about previous iterations.

The operations mentioned above are important, however the essence of GA are three following steps: 1. Selection 2. Crossover 3. Mutation. They are responsible for exchange of genetic information between individuals which is why algorithm is able to find proper solution.

Out of many known selection methods, tournament selection was chosen as the one that is efficient enough and allows easily to control selective pressure through size of tournament. High value of this parameter decreases diversity of population[3].

In Chirmes two methods of crossover were implemented, one with fixed exchange point (between genes in genotype) and another one with random exchange point. After that, with random probability of occurence, mutation operator is used (independently for each child from crossover). During mutation, the application determines first whether mutate rotation or translation and then, which component should be changed – x, y or z. After randomly selecting mutation type new value is assigned, either by adding some random value to rotation or by generating new translation value in domain prepared in random population generation phase.

These operations are repeated for each individual from population to get completely new that replaces existing one for next algorithm iteration.

3.3   Chirality Measure Calculations

Fitness function (chriality measure) in Chirmes is defined by Eq. (2). In the first step, new temporal individual is created by multiplying transformation matrix (inherited from genotype) with atoms positions matrix. Having that the application enters the most computationally intensive part – calculation of chirality measure. It is presented in Fig. 3. Initially, distance matrix between corresponding atoms from mirror image and original molecule is calculated. Currently, the application uses only distances in geometrical space without taking into consideration chemical or physical properties of molecule – it is planned to be implemented in future versions. In next step iterative search of smallest distance sum

---

[3] Strong selective pressure make algorithm to convergence to early, from the other side, weak pressure searching less effective. It is important to find correct bias between too strong and too weak pressure [8].
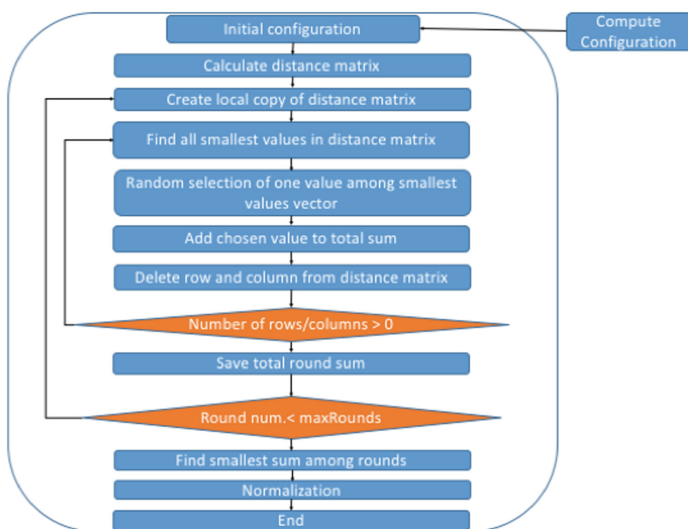
Fig. 3. Algorithm for fitness function calculating – chirality measure

between atoms is being performed. Each iteration consists of following parts: 1. finding smallest values in local distance matrix copy, 2. random selection from one of them. Because several equally small values can be found whole process needs to be repeated multiple times, 3. chosen distance is added to general sum of distances. Row and column where this distance was located in matrix are deleted from local copy, 4. whole process is continued until local matrix will be empty. In last step the algorithm finds smallest value among all calculated in all rounds and its normalization through division by number of molecules in analysed molecule.

4   Results

4.1   Achiral Molecules Test

The basic test for the novel method of calculating chirality measures is to check if ^{SR}CM values calculated for achiral molecules are zero. These molecules are identical with their mirror image thus in most optimal superposition atoms of input compound and their mirrored version are on the exactly same positions so final value of optimization method will be null.

In order to verify if Chirmes fulfill this requirement, a test with four achiral compounds with different size was performed, also in comparison to CHIMEA application. Results are shown in Table 1.

It can be seen that Chirmes finds values very close to ideal 0.0000. Deviation from this value is relatively small, moreover it is known that usually chiral compounds has ^{SR}CM measure in a range of 0.100–0.200. Error is even more

Table 1. ^{SR}CM values calculated for achiral molecules by Chirmes and Chimea

n	Molecule	Atoms no	^{SR}CM value	Calculation time [ms]	^{SR}CM value, CHIMEA	Calculation time, CHIMEA[s]
1	Fullerene	60	0.0067	132884	0.0007	3480
2	Tetraphenyl	44	0.0018	49360	0.0018	600
3	Butane	14	0.0043	4479	0.0005	120
4	Etane	8	0.0020	1819	0.0017	60

negligible when compounds are presented visually using built in visualisation module - in all four cases atoms were perfectly superimposed on their mirror images. What is even more important, achieving similar level of accuracy with CHIMEA takes more than an hour while using Chirmes it took about three minutes.

4.2   Chiral Molecules Test

Further, a comparison between values of ^{SR}CM calculated by CHIMEA and Chirmes for eleven chiral compounds was made. Results are presented by Table 2.

Table 2. Results gathered for ^{SR}CM values from CHIMEA and Chirmes

n	Molecule	CHIMEA	Chirmes	Deviation [%]
1	5-androstenediol	0.6777	0.7051	4
2	3β-androstanediol	0.6962	0.7756	11
3	5.6-didehydroizoandrosterone	0.7281	0.7944	9
4	5α-dihydrotestosterone	0.7512	0.7828	4
5	3α-androstanediol	0.7689	0.8139	6
6	androsterone	0.8154	0.8978	10
7	epitestosterone	0.8188	0.8372	2
8	testosterone	0.8435	0.8727	3
9	4-androstenedione	0.8464	0.9758	15
10	5β-dihydrotestosterone	0.848	1.0453	23
11	4-androstenediol	0.8568	0.8686	1

Percentage deviation of values calculated by Chirmes is in range of 1 to 23 percent. However results from CHIMEA and Chirmes are correlated and correlation rate is $R = 0.86$. Therefore, despite quite high error for some compounds

made by Chirmes, results are correct for QSAR results application. In QSAR most important problem is to find good mapping of interrelationships in set of many compounds, not only about absolute values.

As in test from previous section here also Chirmes was significantly quicker than CHIMEA. To calculate results for all eleven compounds it took only 16 min for CHIMEA comparing to less then 2 min for Chirmes which is almost ten times better.

4.3   Chirmes Usage in Drug Research

To find out about practical advantages of developed application, test verifications were made in Mossakowski Medical Research Centre Polish Academy of Sciences which are shortly described below.

Values of chirality measure for 11 steroids shown in Sect. 4.2 were used to build QSAR model describing affinity of those molecules to androgen receptor. It needs to be explained that androgen receptor is a protein which connects with testosterone and causes production and maintenance of male sexual characteristics. Moreover, it is responsible for building bones, muscles as well as muscles strength. Androgen receptor is important target for drugs assisting in muscle recovery in case of serious diseases or surgery.

Hypothesis, which was verified thanks to calculated QSAR model, says that presence, character of molecule ending elements and shape of whole molecule is most important for successful bonding of steroids with androgen receptor.

After QSAR modelling following equation describing relationship between molecule affinity ($log(RBA)$) and character of ending elements (presented as partial molecular charges $q3$ i $q17$) and general shape of molecule described by ^{SR}CM chirality measure was developed

$$log(RBA) = 4.1(\pm 3.1) + 9.2(\pm 2.5) * q3 - 7.0(\pm 2.3) * q17$$
$$-3.3(\pm 1.2) * {}^{SR}CM, r = 0.83, n = 11 \tag{4}$$

where $log(RBA)$ – relative binding affinity; $q3$ i $q17$ – electrical charges of c3 and c17 carbon atoms. It can be clearly seen that correlation ratio between test and experimental data for this equation (which is only a preliminary model) is at acceptable level ($R = 0.83$) thus it is a good base for more detailed analysis.

Again, it should be emphasized that all these ^{SR}CM values were calculated more then 10 times faster then using previous CHIMEA application which will be especially important for calculating ^{SR}CM for larger molecules. Time gain achieved by Chirmes is very promising for real-world use of the developed application in CADD scientific research. Preliminary scalability tests performed on larger molecules show a similar (or even higher) improvement, as compared to CHIMEA. Their results are beyond the scope of this paper and they will be presented in the following works.

5   Summary

The main goal of the presented research was achieved. A new method for efficient calculation of ^{SR}CM chirality measures by usage of genetic algorithms was developed and implemented as a computer application.

As it is presented in Sect. 4 usage of Chirmes gives significant performance gain (comparing to existing CHIMEA software) with necessary level of correctness. Thanks to possibility of customisation of all parameters in genetic algorithm it is possible to achieve even better results after tuning several GA parameters.

Another advantage of Chirmes is handling of more then 100 well used formats of chemistry related files and multiplatform availability.

During works on Chirmes authors also analysed possibility of porting Chirmes into highly parallel environments such as CUDA. Even though time frame for this paper was too short to prepare fully functional CUDA implementation quick proof of concept application showed potential to gain even higher performance increase then with usage of regular CPU.

The application developed in the presented research is an important step forward in bringing chirality measures to the mainstream of Computer-Aided Drug Design. Chirmes advantages shows that is a good answer for real needs of chemistry science. In biochemistry there is still a lot of space for use of computer aided computations with addition of artificial intelligence and new hardware solutions. Developed software opens new chances for further and intensive computers use in chemistry.

References

1. Blackwell Scientific Publication: Chirality. IUPAC. Compendium of Chemical Terminology (1997)
2. Gonzalez, M.P., Teran, C., Saiz-Urra, L., Teijeira, M.: Variable selection methods in QSAR: an overview. Curr. Top. Med. Chem. **8**(18), 1606–1627 (2008)
3. Jamroz, M.H., Rode, J.E., Ostrowski, S., Lipinski, P.F.J., Dobrowolski, J.C.: Chirality measures of α-amino acids. J. Chem. Inf. Model. **6**(52), 1462–1479 (2012)
4. Jamroz, M., et al.: On stability, chirality measures, and theoretical VCD spectra of the chiral C58X2 fullerenes (X = N, B). J. Phys. Chem. A **1**(116), 631–643 (2012)
5. Nguyen, L.A., He, H., Pham-Huy, C.: Chiral drugs: an overview. Int. J. Biomed. Sci. **2**(2), 85–100 (2006)
6. Lipinski, P., Dobrowolski, J.: Local chirality measures in QSPR: IR and VCD spectroscopy. RSC Adv. **87**(3), 47047–47055 (2014)
7. Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H., Timmerman, H.: Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry, Wiley (2008)
8. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, Berlin (1996)