**Ligand GA Parameters, Files**

This document explains the different parameters and path variables required to use Ligand GA. These have all been defaulted except for the path variables. All of the parameters enclosed in the Ligand\_GA\_Global\_Config file, which is the input to Ligand\_GA.

1) Path names to external software,

- CCDC GOLD executable path,

path/CCDC/Discovery\_2021/bin/gold\_auto

- MN&AM Corina Classic executable path,

path/corina/corina

- MGLTools prepare\_ligand4.py executable path,

path//MGLTools-1.5.6/MGLToolsPckgs/AutoDockTools/Utilities24/prepare\_ligand4.py

pdbqt files are a generalized pdb file that contains dihedral angle information

- Optionally, GOLD can be substituted with AutoDock Vina, an open source software for docking and binding affinity calculation of molecules/stereoisomers.

path/autodock\_vina\_1\_1\_2\_linux\_x86/bin/

2) Ligand\_GA has the file/filename inputs:

- config\_file\_path

The gold.txt has the information necessary to dock the ligand to the protein in GOLD. It can be prepared using CCDC Hermes and also has the paths to the location of the protein pdb file, ligand pdb file, output directory and type of output files. The file has the parameters used in the GOLD genetic algorithm and also specific molecular information that GOLD may use if not defaulted. The file is defaulted with name,

gold.conf

- InitialPopFile

The InitialPopFile is the file that has the initial set, i.e. molecules, used by Ligand GA. This is a one column file of SMILES expressions.

- OutFileName

This file is a .mat file that contains the molecule expressions of the population and their GOLD docking scores at each iteration. This file has non-isomeric SMILES expressions, and if stereoisomer information is required then the molecule has to be re-expanded using Corina into all of these and then re-docked for the individual docking score.

The post-processing function is Ligand\_GA\_Output\_Function. This script will find the unique non-isomeric molecules in the .mat data file and generate a sorted list according to docking score. There could be duplicates of the non-isomeric smiles expressions in the .dat file with different GOLD PLP score due to the GOLD genetic algorithm being non-deterministic. The sorted list will take the duplicated molecule in the .dat file with the highest score.

- ligand\_dir

This is the path/dir of the working directory of Ligand GA. The software will create as many sub-directories as there are in the initial population and use these for large scale calculations.

3) Parameters used in Ligand GA. There are 2 types: genetic algorithm parameters and molecular parameters used in the molecular modifications. The first set is contained in the Ligand GA.m script are used in calling the genetic algorithm within the Matlab global optimization toolbox. The second set are in the file Ligand\_GA\_Global.m and become global variables when Ligand\_GA is executed. Ligand\_GA\_Global.m is the configuration file for Ligand\_GA.

*GA parameters*

These are the parameters for the operation of a Matlab GA in the global optimization toolbox with special crossover and mutation functions. These are defaults and the population size variable is calculated from the InitialPopFile input initial population file which consists of a column of SMILES strings. 12 parameters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PopulationSize | pop\_size |  | TolCon | .01 |
| InitialPopulation | Initial\_pop |  | CrossoverFunction |  |
| Generations | 10000 |  | MutationFunction |  |
| TolFun | .02 |  | OutputFcns |  |
| EliteCount | .1\*pop\_size |  | PlotFcn |  |
| StallGenLimit | 100 |  | Display | Iter |
| crossover\_minus | .9 |  | mutation\_minus | .9 |
| crossover\_increment | .1 |  | mutation\_increment | .1 |
| crossover\_plus | .9 |  | mutation\_plus | .9 |

If the software is to be used without an open Matlab session as in batch, then comment out the display function line in the GA options definition.

In the use of a genetic algorithm, tuning of its GA parameters should be preformed to find the best set. In the problem of molecular construction the crossover and mutation rates are both set to be high, .9 . There is a for loop statement to scan over these parameters but the default is one run with .9 for both. Generations=10000 is an infinite run and only one crossover and mutation rate will be used.

The use of Ligand GA will add the molecules at each iteration to the .dat file if it already exists. The program will first check if the ..dat output file exists and pause if the user wants to delete it before continuing. Else deleting, the software will continue adding. The docking calculation is complicated and it is suggested to use multiple runs and multiple starting points to improve the molecular inhibitor search. This adding is also useful to restart a run with a new InitialPop file containing the output(s) of a previous run.

PopulationSize is determined from the length of the InitialPop file.

There are stopping criteria used in a genetic algorithm, which are conditions used to halt the execution of the GA. In an optimization problem, the solutions are found non-deterministically, and if one is found that is accurate to a required precision then the algorithm will finish. In the molecular problem, due to the complicated nature of the docking process and the landscape of the objective function, these stopping criteria are not as important as the initial population and the execution of multiple runs.

The number of Generations used in the search is a maximum number of iterations of a call to the GA in Matlab. For an indefinite run, this should be set very large, e.g. 10,000, which will make the GA run indefinitely.

EliteCount preserves a fraction of the population from one generation to the next. The default is 10%, but this is a changeable parameter.

Crossover and mutation rates are scanned over if required. This is very time consuming, but is an option. The default is to perform a Ligand\_GA calculation at fixed rates of .9

*Molecular parameters*

These molecular parameters are used to guide the chemical side of the Ligand GA evolution and are not internal GA parameters. These parameters are used to guide the evolution of the molecules and also specify the atomic content. There are 42 Ligand GA parameters, in addition to the parameters used in GOLD and Corina, which are not given.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| atom\_type | ‘CNOPSF’ |  | INTERCHANGE\_DIHEDRAL |  |
| atom\_val | [4,3,2,3,2,1] |  | SINGLE\_DOUBLE\_BOND |  |
| unnorm\_percent\_atom | ones(1,6) |  | DOUBLE\_SINGLE\_BOND |  |
| max\_tries | 20 |  | SINGLE\_TRIPLE\_BOND |  |
| CHANGE\_ATOM |  |  | TRIPLE\_SINGLE\_BOND |  |
| ADD\_ATOM |  |  | DOUBLE\_TRIPLE\_BOND |  |
| DELETE\_ATOM |  |  | TRIPLE\_DOUBLE\_BOND |  |
| ADD\_BRANCH |  |  | unnorm\_mut\_probability |  |
| DELETE\_BRANCH |  |  | min\_heavy\_atoms | 10 |
| CLOSE\_RING |  |  | inter\_bond\_distance | (next) |
| OPEN\_RING |  |  | cutoff\_bond\_percentage | 1.10 |
| OPEN\_BOND |  |  | max\_stereoisomers | 512 |

*Atom\_type* is a char array with the atoms used in Ligand GA. Any set of atoms that can be used in Corina and GOLD can be included in this array. Both of these external softwares are involved and the types of atoms, ions, metallic elements is very broad. The next version of Ligand GA will include the use of pseudo-atoms. An example of a pseudo-atom is -OP(=O)O- with valence 2. The pseudo-atom is treated as a 5 character name, e.g. X0111, and is treated as an atom in the molecular modifications. In the use of Corina or GOLD, the pseudo-atom is expanded into its full SMILES expression. Pseudo-atoms can be residue building blocks such as GlcNAc, Glc, …, sulfates, ions, …, even small substructural units.

*Unnorm\_percent\_atom* is the ‘abundance’ of the atoms in atom\_type. In CHANGE\_ATOM and ADD\_ATOM, an atom type must be selected. The random choice is guided by the probabilities in percent\_atom. For example, if in Table 2, the percent\_atom(5)=0 then no sulfurs will be selected to include in the molecular modifications.

*Max\_tries* is a limit on the attempts made in implementing a molecular modification. For example, in changing an atom, if an F is selected and there are no available atoms to replace due to valence of 1 then the software will stop trying at 20 attempts. Another example would be to implement a change single to double bond in the molecule and there being no possibility. In the next version of Ligand GA, this will be computationally improved by examining the molecule first to determine if the mutation is possible and if so, where in the molecule for a small boost in efficiency.

*Unnorm\_mut\_probability* guides the choice of mutation function if a mutation is wanted. The default is uniform, but by no means would this be the correct choice for specific small molecule goals. In the example directory of the download there is a biased Ligand\_GA\_Global.m configuration file for producing modified Aspirins. There are several configuration files and outputs, but in this biased example, the unnormalized probability array is [3,1,2,1,1,3,1,1,.02,.02,.02,.02], which guides the molecular evolution away from adding flexible chains and branches, and minimizes the number of double and triple bonds, which are less common generally. These parameters are important in the design of molecules.

*Min\_atoms* is the strict minimum number of atoms allowed in a molecule. Any molecule with less than this are not permitted in the population. The implementation is made within the mutation and crossover functions.

*Inter\_bond\_distance* is used in the clash detection process. All small molecules cannot have inter-distances between a pair of non-bonded less than these distances x cutoff\_bond\_percentage, and the default of the latter is 1.10 . The coordinates of the mol2 files are checked in a mutation or crossover inside the CLASH\_CHECK.m function and if there is a distance clash then the molecule is not accepted in the population. These inter\_bond\_distance distances are the single-bond covalently bonded atoms, which is the longest distance out of a bond (e.g., single, double, triple). Inter-atomic single bond lengths are a 2d array inter\_bond\_distance,

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CC | 1.53 |  | NN | 1.45 |  | OO | 1.49 |  | PP | 2.18 |
| CN | 1.49 |  | NO | 1.46 |  | OP | 1.61 |  | PS | 2.12 |
| CO | 1.43 |  | NP | 1.68 |  | OS | 1.52 |  | PF | 1.67 |
| CP | 1.82 |  | NS | 1.77 |  | OF | 1.21 |  |  |  |
| CS | 1.81 |  | NF | 1.30 |  |  |  |  | SS | 2.05 |
| CF | 1.40 |  |  |  |  |  |  |  | SF | 1.61 |

*Max\_stereoisomers* is used to limit the number of stereoisomers generated by Coraina. =512, so the default limit is 9 chiral centers or cis-/trans- bonds.

*ADME restriction parameters*

The restriction of molecules to satisfy ADME heuristic requirements requires parameters. These are factors used in Lipinski’s Rule of 5 and subsequent rule of 3 (RO3). There are 17 parameters. The masses are in amu.

*Max\_mass\_ADME* is the maximum mass of a molecule.

*Max\_donor\_ADME* is the maximum number of hydrogen bond donors.

*Max\_acceptor\_ADME* is the maximum number of hydrogen bond acceptors.

*Max\_rotatable\_dihedrals* is the maximum of rotatable dihedral angles in the ligand.

There are 2 types of implementation: soft or hard. Soft includes a penalty by the amount of violation of each of these constraints in the fitness function. The penalty parameters are listed also. A hard implementation penalizes the molecule by the number of total violations, regardless of how much, times the hard penalty. The ADME\_Penalty\_Type is specified by ‘SoftLipinski’, ‘HardLipinski’, or anything else, e.g. ‘None’.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ADME\_Penalty\_Type | None SoftLipinski, HardLipinski |  |  |  |
| max\_mass\_ADME | 500 |  | penalty\_mass | 4 |
| max\_donor\_ADME | 5 |  | penalty\_donor | 10 |
| max\_acceptor\_ADME | 10 |  | penalty\_acceptor | 10 |
| max\_rotatable\_dihedrals | 7 |  | penalty\_dihedrals | 10 |
| atomic\_daltons |  |  | hard\_violation\_penalty | 20 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| C | N | O | P | S | F |
| 12.011 | 14.007 | 15.999 | 30.974 | 32.065 | 18.998 |

*Software components*

These are the head program Ligand GA function and the various components used in its execution.

|  |  |  |
| --- | --- | --- |
| Ligand\_GA | Ligand\_GA\_Global | Ligand\_GA\_Output\_Function |
| Ligand\_GA\_Crossover | Ligand\_GA\_Mutation | Ligand\_GA\_Load\_Population |
| MolecularStructure | CLASH\_CHECK | AtomCharVal |
| Ligand\_GA\_Fitne…ion\_GOLD\_ADME | RING\_RENUMBER\_CHECK | Ligand\_GA\_Pop\_Save |
| Ligand\_GA\_Fitne…ion\_Vina |  | Load\_Population |

A brief description of the functions of the programs are:

*Global\_Ligand\_GA* : input file with the file/filenames and parameters used in the different functions

*Ligand\_GA* : main call to the GA, takes the Configuration file as input

*Ligand\_GA\_Mutation* : the custom mutation top function

*Ligand\_GA\_Crossover* : the custom crossover function

*Ligand\_GA\_Fitness\_Function\_GOLD\_ADME* : the CCDC GOLD docking fitness function

*Ligand\_GA\_Pop\_Save* : function to save the output of the population docking at each iteration

Mutation operations in Ligand\_GA\_Mutation used on the chromosomes (i.e. molecules), called from

CHANGE\_ATOM

ADD\_ATOM

DELETE\_ATOM

ADD\_BRANCH

DELETE\_BRANCH

CLOSE\_RING

OPEN\_RING

OPEN\_BOND (not ring point)

SINGLE\_DOUBLE\_BOND

DOUBLE\_SINGLE\_BOND

SINGLE\_TRIPLE\_BOND

TRIPLE\_SINGLE\_BOND

DOUBLE\_TRIPLE\_BOND

TRIPLE\_DOUBLE\_BOND

crossover operation called from Ligand\_GA\_Crossover:

INTERCHANGE\_SEGMENT

ring number checking and rewriting if necessary:

RENUMBER\_RING\_CHECK

*MolecularStructure* : used to get geometric information from the SMILES expression

this is used in the mutation and crossover functions

*AtomCharVal* : gives the numeric valence of the atom, e.g. 'C'->4

*CLASH\_CHECK* : used to check if there is a geometric clash of nonbonded atoms in the molecular modification. The 1.1 (default) is used to determine if a pair of atoms is within 1.1 x single\_pair\_distance. If the pair is not in a CONECT record and within inter\_bond\_distance\*cutoff\_bond\_percentage then there is a clash and the molecular modification is not allowed.

The maximum ring number is 9 in this Ligand GA version.

The docking jobs are sent processes that will use any number of processors as permitted in the cluster. The number of processes at each iteration can be large due to the population size and number of stereoisomers. The number of stereoisomers can grow as Ligand GA runs due to the growth in chiral centers and cis-/trans- double bonds increases, if no constraints are applied to these.

Note: Only one Ligand\_GA instance should be running for a given ligand\_dir, with multiple processes. The directories molecule and molecule(i) are being used and there could be overwriting while a process is using the file. If more than one instance of Ligand\_GA is wanted, then the ligand\_dir should be different.

(Ligand\_GA\_Molecule\_Fitness\_Stereoisomer : stereoisomers are generated during the run of the program, but its information is not kept in the output. this function makes a list of stereoisomers, e.g. stereoisomer number and GOLD score. This is in the next version.)