Systems Engineering

Computational Visualization Method for Drug Discovery

Executive Summary

Student: Shy Alon

ID: 038505665

Supervisor: Dr. Abraham Yosipof

Signature: _____

Submission: 19.09.2017

AFEKA - Tel-Aviv Academic College of Engineering

EXECUTIVE SUMARY

Department of Systems Engineering

Master of Science

Computational Visualization Method for Drug Discovery

by Shy Alon

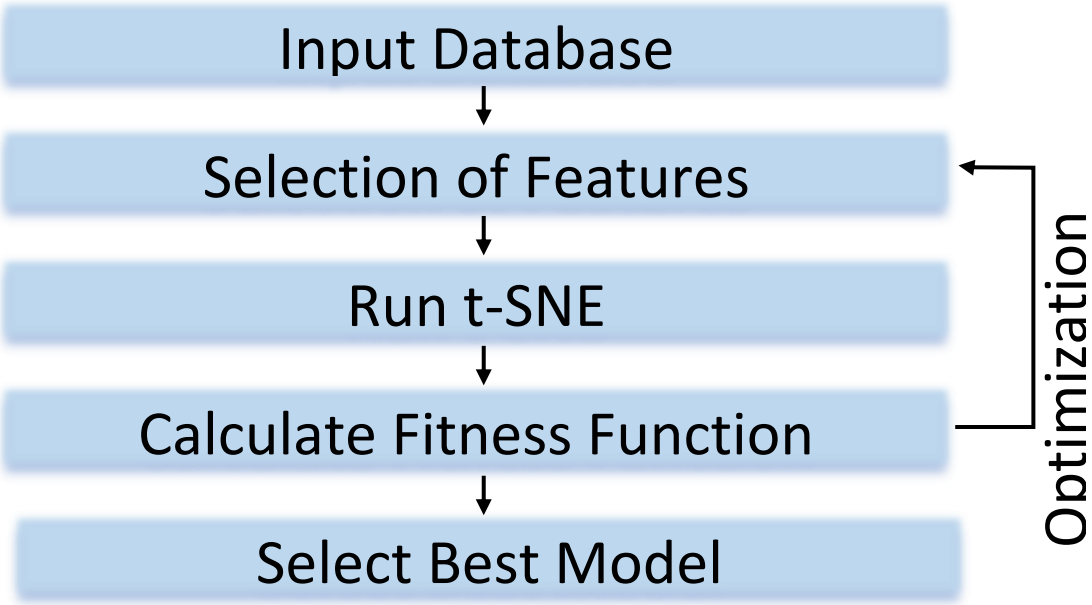
The cost of the process of researching and developing new medicines is currently estimated by the pharmaceutical companies to be five and a half billion USD and is constantly on the rise.

The first phase of said process – the pre-clinical phase - is generally considered to be so risky and unprofitable that the pharmaceutical industry has abandoned it completely to NGOs and academic institutions which pursue the discovery of new drugs for motives other than profit.

The hypothesis of this paper is that a cost effective computational method could be utilized to significantly reduce the costs of the pre-clinical phase to the degree it would be cost effective. It is demonstrated on G-protein-coupled receptors (GPCRs) datasets which, while being only around 3% of known molecular targets, represent as much as 45% of current drug targets and thus have an excellent potential for drug discovery.

The paper uses a database of over seven thousand compounds having over one hundred and thirty features. The computational method used in this paper uses genetic algorithm to identifie the optimal feature sets to be used in a dimensionality reduction process - t-distributed stochastic neighbor embedding (t-SNE). T-SNE has been selected because is prioritizes short distances over large distances and therefore serves better as basis for nearest neighbor based classification method.

The criteria for success of the algorithm is the nearest neighbor criteria – meaning how likely is the nearest neighbor of a positively tagged compound to have the same tag.



The results of the algorithm (as seen in the table below) have proven the hypothesis to be correct and that a computational process could identify high potential candidate and reduce the number of laboratory tests by more than 90 percent.

In the table below Q-1 means the probability that an unknown compound which nearest neighbor is a compound activating a certain agent (for example Dopamine) will activate the same compound. Since the average probability is 97.7% it means that using this algorithm on a similar database will result in a very high degree of certainty regarding the compound's activity before laboratory test.

Table 1 Result Summary

	Q-1	***Q-3***	***Q-5***	***Trust***
Dopamine	0.9650	0.9650	0.9610	0.0870
Adrenoceptors	0.9830	0.9850	0.9840	0.0770
Histamine	0.9670	0.9630	0.9640	0.0780
Muscarinic	0.9930	0.9930	0.9930	0.0760
Serotonin	0.9380	0.9340	0.9210	0.0900
Average	*0.9692*	*0.9680*	*0.9646*	*0.0816*

Even though complexity requirements make the algorithm unsuitable to Big Data scale databases is has been shown to be very effective when used in smaller scale databases (like the 7000 by 130 database used in this paper).