

**TRIBHUVAN UNIVERSITY**

**INSTITUTE OF ENGINEERING**

**PULCHOWK CAMPUS**

**THESIS NO.: 2072-MSCSKE-663**

**A GENETIC ALGORITHM FOR PROTEIN CRYSTALLIZATION SCREENING**

**BY**

**RHITWIK TIWARI**

**A THESIS**

**SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SYSTEM AND KNOWLEDGE ENGINEERING**

**DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING**

NOVEMBER, 2017

**A Genetic Algorithm for Protein Crystallization Screening**

By

Rhitwik Tiwari

2072-MSCSKE-663

Thesis Supervisor

Sanjeeb Prasad Panday (Ph.D.)

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer System and Knowledge Engineering

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Tribhuvan University

Lalitpur, Nepal

November, 2017

**COPYRIGHT ©**

The author has agreed that the library, Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus, may make this thesis freely available for inspection. Moreover, the author has agreed that the permission for extensive copying of this thesis work for scholarly purpose may be granted by the professor(s), who supervised the thesis work recorded herein or, in their absence, by the Head of the Department, wherein this thesis was done. It is understood that the recognition will be given to the author of this thesis and to the Department of Electronics and Computer Engineering, Pulchowk Campus in any use of the material of this thesis. Copying of publication or other use of this thesis for financial gain without approval of the Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus and author’s written permission is prohibited.

Request for permission to copy or to make any use of the material in this thesis in whole or part should be addressed to:

Head

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Pulchowk, Lalitpur, Nepal

**Recommendation**

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a thesis entitled **“A Genetic Algorithm for Protein Crystallization Screening**”, submitted by **Rhitwik Tiwari** in partial fulfillment of the requirement for the award of the degree of “**Master of Science in Computer System and Knowledge Engineering**”.

……………………………………

**Supervisor: Sanjeeb Prasad Panday(Ph.D.)**

…………………………………….

**External Examiner:**

**Departmental Acceptance**

The thesis entitled **“A Genetic Algorithm for Protein Crystallization Screening”,** submitted by **Rhitwik Tiwari** in partial fulfillment of the requirement for the award of the degree of “**Master of Science in Computer System and Knowledge Engineering**” has been accepted as a bonafide record of work independently carried out by him in the department.

-----------------------------------------------------------------

**Dibakar Raj Pant(Ph.D.)**

Head of the Department

Department of Electronics and Computer Engineering, Pulchowk Campus,

Institute of Engineering,

Tribhuvan University, Nepal

ACKNOWLEDGEMENT

First and foremost, I would like to thank our Head of Department Dr. Dibakar Raj Pant for the enduring support. I would express a word of thankfulness to our Coordinator Dr. Aman Shakya for his motivation.

I would also like to thank Professor Dr. Shashidhar Ram Joshi, Professor Dr. Subarna Shakya and Dr. Sanjeeb Prasad Panday for their support and suggestions. I would like to express my sincerest thanks to Dr. Basanta Joshi and Er. Dinesh Baniya for their encouragements.

**ABSTRACT**

The structure of protein has been used to predict the position of thousands of atoms within the protein crystals. Protein crystallization is the process of formation of protein crystals. [Proteins](https://en.wikipedia.org/wiki/Proteins) are biological [macromolecules](https://en.wikipedia.org/wiki/Macromolecules) and function in an [aqueous](https://en.wikipedia.org/wiki/Aqueous) environment, so protein crystallization is predominantly carried out in presence of water. The aim of this study was to produce crystals of protein to study its structure. Subsequent to this, Genetic Algorithm was used to determine the optimal conditions required for the well-ordered crystals of proteins. In this thesis work, three different types of conditions on the basis of alkalinity and acidity were analyzed for, i.e., acidic, neutral and basic. Genetic Algorithm was proved to be equally well with respect to Associative Experimental Design (AED) conducted previously.

Table of Contents

Approval…………………………………………………………………………………..ii

Copyright…………………………………………………………………………...…….iii

Recommendation……………………………………………………………………...….iv

Departmental Acceptance……………………………………………………………...….v

Acknowledgement………………………………………………………………..............vi

Abstract ………………………………………………………………………………….vii

List of Figures ………………………………………………….………………………....x

List of Tables...………………………………………………………………………...…xi

List of Symbols……………………………………………………………………..........xii

Abbreviations……………………………………………………………………………xiii

[1. INTRODUCTION 1](#_Toc497475740)

[1.1. Background Theory 1](#_Toc497475741)

[1.2. Problem statement 2](#_Toc497475742)

[1.3. Objectives 2](#_Toc497475743)

[1.4. Motivation 3](#_Toc497475744)

[1.5. Scope of the Work 3](#_Toc497475745)

[1.6. Organization of the Thesis 4](#_Toc497475746)

[2. LITERATURE REVIEW 5](#_Toc497475747)

[3. METHODOLOGY 8](#_Toc497475748)

[3.1. Data Collection 8](#_Toc497475749)

[3.2. Setting Up Experiment 8](#_Toc497475750)

[3.3. Scoring of Protein Crystals 9](#_Toc497475751)

[3.4. Genetic Algorithm 12](#_Toc497475752)

[3.4.1. Methods 12](#_Toc497475753)

[3.4.1.1. Chromosome Encoding 12](#_Toc497475754)

[3.4.1.2. Selection 14](#_Toc497475755)

[3.4.1.3. Crossover 14](#_Toc497475756)

[3.4.1.4. Mutation 15](#_Toc497475757)

[3.4.1.5. Iteration Parameter 15](#_Toc497475758)

[3.5. Flowchart of the Method 16](#_Toc497475759)

[3.6. Implementations and Tools 22](#_Toc497475760)

[3.6.1. Tools 22](#_Toc497475761)

[3.6.1.1. Analysis and Design Tools 22](#_Toc497475762)

[3.6.1.2. Implementation Tools 22](#_Toc497475763)

[3.6.2. Implementation 22](#_Toc497475764)

[4. RESULTS AND DISCUSSIONS 24](#_Toc497475765)

[4.1. Results 24](#_Toc497475766)

[4.2. Discussions 40](#_Toc497475767)

[5. CONCLUSION AND RECOMMENDATIONS 44](#_Toc497475768)

[5.1. Conclusion 44](#_Toc497475769)

[5.2. Limitations 44](#_Toc497475770)

[5.3. Future Enhancements 45](#_Toc497475771)

[6. REFERENCES 46](#_Toc497475772)

APPENDIX ……………………………………………………………………………...49

Appendix A………………………………………………………………………49

Appendix B………………………………………………………………………54

**LIST OF FIGURES**

**Figures Page No**

Figure 3.1. Sample Plate for Crystallization Screening………………………….9

Figure 3.2. Sample microscopic images of the outcomes…………….................11

Figure 3.3. Flowchart of the Method…………….……………………................16

Figure 3.4. Sample data showing the steps of the methodology….……...............21

Figure 4.1. Crystals formation of AbIPPase ……………………………….........24

Figure 4.2. Crystals formation of KbIPPase…………………………………….25

Figure 4.3. Crystals formation of CjIPPase…………………………….……….25

Figure 4.4. Graph depicting relationship of Fitness Score and No. of Generations for MES Sodium……………………………………………………….……………..41

Figure 4.5. Graph depicting relationship of Fitness Score and No. of Generations for Bicine Chloride……………………………………………………..…………….42

Figure 4.6. Graph depicting relationship of Fitness Score and No. of Generations for Hepes Sodium……………………………………………………….……………42

**LIST OF TABLES**

**Table Page No**

Table 3.1: List of Hampton and Revised scores…………………………….10

Table 3.2: Relationship of Tournament Size and Selection Intensity………20

Table 4.1: Fitness Score for Acidic Buffer Solutions (AbIPPase)………….28

Table 4.2: Fitness Score for Basic Buffer Solutions (KbIPPase)…………...34

Table 4.3: Fitness Score for Neutral Buffer Solutions (CjIPPase)………….38

**LIST OF SYMBOLS**

1. 𝛿r : Reagent
2. ⍴(𝛿r) : Significance Ratio of a reagent
3. ⍴(𝛿b) : Significance Ratio of a buffer
4. ⍴(𝛿p) : Significance Ratio of a precipitant
5. ⍴(𝛿s) : Significance Ratio of a salt
6. μ(δp) : Average score of all cocktails containing precipitant p
7. μ(𝛥−δp): Average score of all cocktails not containing precipitant p
8. AbIPPase: Acidic Buffer Solution
9. CjIPPase: Neutral Buffer Solution
10. KbIPPase: Basic Buffer Solution

**ABBREVIATIONS**

1. AED – Associative Experimental Design
2. CASE – Computer Aided Software Engineering
3. CC – Child Chromosome
4. DBMS – Data Base Management System
5. DFD – Data Flow Diagrams
6. FS – Fitness Score
7. GUI – Graphical User Interface
8. IDE – Integrated Development Environment
9. Ltd. – Limited
10. MS – Microsoft
11. PC – Parent Chromosome
12. pH – Potential of Hydrogen (Phenolphthalein)
13. Pvt. – Private
14. SMS – Sparse Matrix Sampling
15. SQL – Structured Query Language
16. 1D – One - dimensional
17. 2D – Two - dimensional
18. 3D – Three - dimensional

# INTRODUCTION

## 1.1. Background Theory

Protein crystallization is the process of formation of protein crystals. This process depends on a number of experimental conditions such as pH, temperature, ionic concentrations and various other reactions that affect crystal growth. Every protein has a unique primary structure and needs its own set of parameters to be set up in order to yield crystals for that protein [1]. The optimal conditions for crystallization are difficult to predict because the total number of potential solutions to be tested turns out to be in the order of hundreds of thousands, and setting up experiments for all these conditions is highly impractical. Due to this, the conditions for crystallization are determined using screening experiments [2].

Visualization of protein structure via crystallography has a significant role in clinical chemistry [3, 4]. Protein crystallographic methods have been adopted by researchers in drug discovery and optimization for over two decades to aid in compound design [5, 6]. It allows researchers to visualize how drug molecules bind with target proteins. Structural information lends itself to small-molecule design and the generation of new chemical ideas for drug discovery process [7].

Crystallization screening aims to identify the important factors that contribute to the crystallization of proteins. The major hurdle in the screening process is the vast chemical space that is often expanded to hundreds of thousands of conditions. Cost and time constraints make exhaustive trial of all these conditions practically impossible. In order to increase efficiency of the experiments, commercial screens are now available that are likely to produce successful outcomes for some proteins. However, these screens are not guaranteed to work for all proteins especially for the ones that are difficult to crystallize. An alternative would be to analyze the results of existing experiments and derive new conditions based on those results. But this approach is not without faults because of several reasons. The first and most crucial is the vast number of possible conditions and only few successful outcomes. Because of this, the data is highly skewed and identifying factors for crystallization is very difficult [8]. There may not even be a correlation between those factors. The data has no known distribution (normal, linear, exponential, etc.) and so modeling the data using conventional methods is not possible. Also, there is no way to measure the effectiveness of conditions which makes classification difficult.

## 1.2. Problem statement

Protein Crystallization is the process of growing crystals in a solution to identify protein structure. Setting up an experiment with multiple combinations of reagents to make solutions for crystal growth is called screening. Since the number of combinations of reagents can be very high, there are several commercial screens available that help crystallographers to increase experiment efficiency.

In this thesis, it is to be analyzed whether genetic algorithms could be used to generate crystalline conditions by analyzing the results of previous experiments. Given a screen (population) of conditions along with their scores, the goal is to generate a new set of conditions that would generate crystalline conditions. A genetic algorithm makes use of a large and uneven search space to produce optimal results. Genetic algorithm is suitable for protein crystallization screening since it produces a population of conditions (or cocktails) as an output compared to other methods (e.g., classification methods) that evaluate a single condition.

## 1.3. Objectives

The objectives of the thesis work are as follows:-

* To determine the optimal conditions for the crystallization of protein using Genetic Algorithm
* To determine the protein structure using Genetic Algorithm.

## 1.4. Motivation

Methods for protein crystallization screening such as full factorial and incomplete factorial experiments are balanced. These methods do not consider generation of novel conditions. The experiments already done using Associative Experimental Design (AED) method proved that novel conditions can be generated by swapping reagents among cocktails having common elements. The success of AED in generating novel conditions with good crystalline outcomes was the primary motivation for this thesis. However, AED works only with cocktails that have a common reagent, meaning that there could be some potential conditions with a high degree of crystallization that could be missed by AED.

Screening using Genetic Algorithm takes a different approach. It analyzes reagents and their individual scores in the dataset and then generates conditions based on the aggregate score of individual reagents in a condition. It can be said that it is a more generic implementation of AED in that it pairs two conditions that do not necessarily have common reagents, but rather have a high individual score themselves, to generate new conditions. This allows the genetic algorithm to capture conditions that may have been missed by AED.

## Scope of the Work

The study work is focused on determining the structure of protein by the use of Genetic Algorithm which would be useful for hospitals and research laboratories that conduct experiments on proteins and its variants.

## Organization of the Thesis

The first chapter is the background theory: basic principles and objectives of the thesis. It lays out the background for some of the key terminologies associated with protein crystallization that will aid the reader in understanding the discussions that follow. It also provides related work which acted as motivator for developing experiments for protein crystallization. Chapter 2 provides an overview on literature review and describes what have been done in the same context in the past. Chapter 3 explores some essential genetic algorithm terms and how they are utilized. It then goes into explaining the algorithm components in more detail. Chapter 4 provides the theoretical and experimental results of the method along with the analysis of results. Chapter 5 focuses on the conclusion of the thesis work with the limitations and future enhancements. Chapter 6 lists up the references that had have helped to conceptualize the idea for the work. At last, Appendices include the graphical relationships between the fitness score and the number of generations of different buffer solutions together with certain source codes and snapshot of the output.

# 2. LITERATURE REVIEW

Protein Crystallization is the process of growing crystals in a solution to identify protein structure. Setting up an experiment with multiple combinations of reagents to make solutions for crystal growth is called screening. Since the number of combinations of reagents can be very high, there are several commercial screens available that help crystallographers to increase experiment efficiency. However, the available commercial screens are not guaranteed to crystallize difficult proteins. Also, since the number of solutions in these screens is limited, the data collected from these screens also tend to be limited to just those solutions. There are no reliable metrics to measure chemical distances, and so it is hard to establish some sort of correlation between factors that affect crystallization [9]. There are also cost factors that need to be accounted for. It is not practical to set up experiments covering thousands of conditions due to the cost and limited availability of proteins and resources. The success rate for difficult proteins is low.

The existing methods for optimizing protein crystallization screening can be categorized into two classes based on whether they use results from previous experiments [9, 10]. In the first category, chemical factors and levels are defined by user. This approach does not take in data from previously conducted experiments. Full Factorial Design, Incomplete Factorial Design, Block design and Bayesian design are some of the methods that fall under this category [11].

Another method is the Sparse Matrix Sampling (SMS), put forth by Jancarik and Kim in 1991. The SMS method further reduces the amount of proteins needed for initial screening by designing a sparse matrix of trial conditions selected from known crystallization conditions [12]. The major parameter variables chosen were pH and buffer materials, additives and precipitating agents. The values for the pH and buffer were chosen from known conditions with successful crystalline outcomes. The choice of reagents was also based on past experiences of many crystallization trials. Commercial screens based on SMS have been developed.

The second approach would be to design experiments using data from previously conducted experiments [13]. Data mining methods such as regression, classification, association analysis, etc. have been used for experimental design domain [14, 15]. However, these methods did not prove to be effective in the context of screening optimization due to insufficient and highly skewed data and a very large solution space [16, 17].

A third method called Associative Experimental Design (AED), proposed by Dinc et al. [18], uses outcomes of previously conducted experiments to generate new conditions by identifying screening factors that are most likely to produce high scoring crystalline outcomes. It is not an initial screening process as it uses results from the previous screening conditions having a higher probability of producing good crystals. It analyzes conditions that have common reagents and swaps the rest of the reagents among the conditions to produce new candidate conditions to be tested in the lab. AED is different from previously discussed methods in that it analyzes possible interactions between reagents in order to determine new crystalline conditions.

A genetic algorithm-inspired concept for optimizing screening process was put forward by Emmanuel et al, which used the principles of crossover and mutation to produce novel conditions [19]. This method used ‘hits’ (conditions that produced crystals) from previously conducted experiments and applied crossover and mutation to some of the hits to generate new conditions. It does not assign scores to reagents or conditions and the crossover and mutation process and was run only up to the second generation. But even with this extremely simplified version of the genetic algorithm, the researchers were able to generate crystalline conditions from the 50 conditions of the Hampton Research Crystal Screen [19].

# 3. METHODOLOGY

## 3.1. Data Collection

Dataset from the laboratory and research center of a hospital was used to evaluate the conditions for protein crystallization and determine the structure of proteins, using Genetic Algorithm. In this case, data from Birat Aspatal Pvt. Ltd., Biratnagar have been used.

## 3.2. Setting Up Experiment

Protein is generally dissolved in a liquid solution. Formation of protein crystals requires the solution to be super-saturated which means that the concentration of the solute is above the solubility limit. Solutions that are under-saturated or saturated don’t produce crystals. However, a super saturated solution alone does not guarantee crystal growth and the process is affected by several other physical factors such as pH, temperature, activation energy, and other chemical reactions within the solution. For successful crystallization, specific activation energy and an ordered sequence of inter-molecular-interactions are required [20, 21].

For the thesis work, Protein crystals were grown in a saturated solution in the lab. A solution was made up of several crystallization agents such as buffers, precipitants, salts and other additives. The buffers were used to maintain pH of solution and additives to control solubility of protein. Once the solutions were prepared, experiment plates were set up with the solutions and protein.

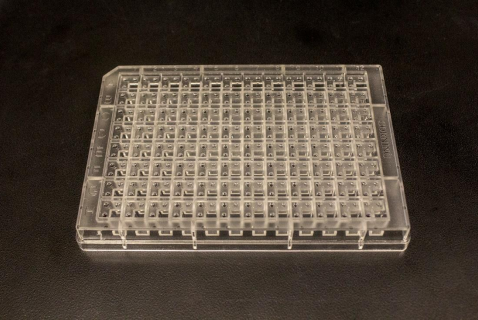


Figure 3.1: Sample Plate for Crystallization Screening

## 3.3. Scoring of Protein Crystals

There are many scoring schemes for protein crystallization outcomes. In this case, Revised Hampton scoring was used which used a range of 0 to 9 to score proteins [22]. The scores for protein crystals were assigned manually as shown in Table 3.1. Likewise, Figure 3.2 shows sample images corresponding to each Hampton score.

Table 3.1: List of Hampton and Revised scores.

|  |  |  |
| --- | --- | --- |
| **Hampton Scoring** | **Revised Scoring** | **Outcomes** |
| - | 0 | Heavy amorphous precipitate |
| 1 | 1 | Clear solution |
| 2 | 2 | Phase change (oiling out) |
| 3 | 3 | Precipitate (light) |
| 4 | 4 | Bright spots or granular precipitate |
| 5 | 5 | Spheroids, dendrites, urchins |
| 6 | 6 | 1D needles |
| 7 | 7 | 2D plates |
| 8 | 8 | 3D Crystals (Small) |
| 9 | 9 | 3D Crystals (Large) |

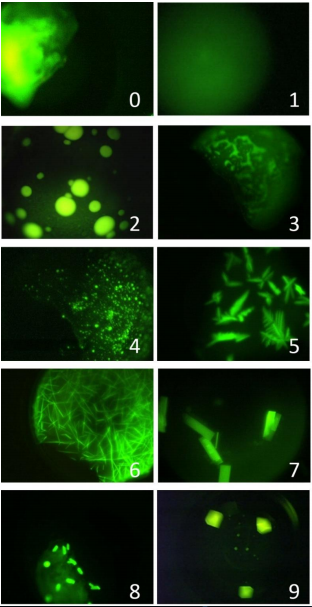


Figure 3.2.: Sample microscopic images of the protein crystallization outcomes. (0) Heavy amorphous precipitate (1) Clear solution (2) Phase change (oiling out) (3) Precipitate (light) (4) Bright spots or granular precipitate (5) Spheroids, dendrites, urchins (6) 1D needles (7) 2D plates (8) 3D crystals, small (9) 3D crystals, large

## 3.4. Genetic Algorithm

## 3.4.1. **Methods**

The key steps in a genetic algorithm are encoding, selection, crossover and mutation. The last three steps are repeated in succession until an optimal result set is achieved.

### 3.4.1.1. Chromosome Encoding

The primary issue to be addressed while implementing a genetic algorithm was the representation of the chromosome. In this case, a chromosome was a single cocktail made up of buffer, precipitant and salt. Hence, the chromosome was of length 3 (buffer, precipitant and salt). The most common way of encoding chromosomes in many genetic algorithms is binary encoding. However, binary encoding was not adopted because it only added to the complexity of the problem. Instead, the conditions were represented by their index values in the dataset.

*D = {(Ci,Si) | (C1,S1),(C2,S2),…,(Cn,Sn)}*  (3.1)

where D is the dataset and the pair *(Ci, Si)* corresponds to the condition *Ci* and its score *Si* for the *ith* entry in the dataset. This method used 3 components of any given condition: pH of the solution, type of precipitant and type of salt. Therefore, a cocktail was represented as,

*Ci = {pHi, Pi, Si}*  (3.2)

where, ‘i’- number of samples in the dataset

pHi - the pH of the ith solution

Pi - the type of precipitant; and

Si - the salt.

A scoring mechanism was to be developed that would make use of available scores from previous experiments. After analyzing the ranking mechanism of AED, “significance ratio” [18] was used to compute scores for individual reagents and then the individual scores were averaged to obtain the overall score for a given condition.

Each reagent (𝛿r) in the dataset had its own significance score ⍴(𝛿r), which was the ratio of the average of the scores of conditions in the dataset having that reagent vs. mean of the average of scores of cocktails without that reagent and average of scores of cocktails with that reagent. As an example, the significance ratio of a precipitant P was calculated as:

where, μ(δp) - the average score of all cocktails containing precipitant P ; and

μ(𝛥−δp)- the average score of all cocktails not containing precipitant P.

The Fitness Score (FS) score for a chromosome was calculated as the average of the significance ratio of all 3 components (i.e. buffer, precipitant and salt) of the corresponding cocktail.

where, *i* - the index of cocktail in the dataset

⍴(𝛿𝑏) - the significance ratio of buffer

⍴(𝛿𝑝) - the significance ratio of precipitant

⍴(𝛿𝑠) - the significance ratio of salt;and

*n* - the total number of data in the dataset whose significance ratio was taken into consideration.

### 3.4.1.2. Selection

The selection process consisted of picking two cocktails from the population with a relatively high fitness score for the crossover step. There are several techniques for selection available such as fitness proportionate selection, reward based selection, stochastic universal sampling, tournament selection, etc. [23]. Screening using Genetic Algorithm maintains diversity in the choice of reagents and does not completely eliminate the weaker reagents. For this reason, the tournament selection technique was chosen. The idea behind tournament selection was to select the fittest candidate (one with the highest fitness score) out of a small subset of the total population. So, if there were 100 cocktails in the main population, a subset of (say 5) randomly chosen cocktails were created, and then the fittest out of that subset was selected. Tournament size parameter set the size of the tournament during the parent chromosome selection process. If the tournament size was set to 1, a single random chromosome was selected from the population. If the tournament size was set 3, then 3 random conditions were selected from the population and the fittest of the three was chosen as a parent chromosome. This selection process was done twice to produce two parent chromosomes.

### 3.4.1.3. Crossover

Crossover is the process of creating a new chromosome using genes from two parent chromosomes. In this case, a single point crossover was chosen to determine gene inheritance. In single point crossover, a number between 1 and the total length of the chromosome was taken at random. That number was the crossover point and was also the index that marked the separation of inheritance for the child chromosome.

*PCi = {pHi, Pi, Si}*  (3.5)

*PCj = {pHj, Pj, Sj}*  (3.6)

*PCi* and *PCj* were two parent chromosomes obtained from the tournament selection process. Let the random crossover point be 2. With this value, the child chromosome received its first two genes from parent *i* and the remaining genes from parent *j* and generated the following child:

*CCij = {pHi, Pi, Sj}* (3.7)

*CCij* was the child chromosome obtained after crossing over two parent chromosomes, *PCi* and *PCj*.

### 3.4.1.4. Mutation

The next step in the algorithm was called mutation, and its purpose was to introduce some variation in the child chromosome. The original screening dataset was maintained and a random gene was picked up from the child chromosome which was replaced by a corresponding gene from a randomly selected candidate in the original dataset. Thus, the child chromosome had a gene from the original dataset that is different from the ones inherited from its parents. Mutation was not applied to the final generation.

### 3.4.1.5. Iteration Parameter

Iteration parameter specified the number of iterations the algorithm should be run. This is related to the number of cycles of crossover and mutation to be applied to the original dataset. For lower number of iterations, high ranking cocktails might not have been identified completely. It was thought to keep the number of within the range of 100-200 to make sure the algorithm did not converge to the local maxima. But, later on, as the highest possible fitness value was 2.0, the number of iterations was kept on the basis of the fitness values.

## 3.5. Flowchart of the Method

Start

Read Input Dataset

Generate Initial Population

Tournament Selection

Crossover

Mutation

Remove prohibited conditions

Apply Ranking

Generate Output

End

Figure 3.3: Flowchart of the method

1. Input Data Set

The data was collected on the basis of three reagents. The three reagents were: Buffer solution, Precipitate and Salt.

1. Buffer Solution

The solution that resists the change in pH value on the addition of acids or bases in it is called Buffer Solution. 17 samples of Buffer Solution were collected. However, the samples were classified into three types:-

1. Acidic – AbIPPase
2. Neutral – CjIPPase
3. Basic – KbIPPase

i. Acidic Buffer Solution (AbIPPase):-

-Buffer Solution with pH value less than 7 is called Acidic Buffer Solution.

* Cacodylate Sodium
* Sodium Phosphate Citrate
* MES Sodium
* Sodium Citrate
* Sodium Acetate
* Ethanoic Acid
* Sodium Ethanoate

ii. Basic Buffer Solution (KbIPPase):-

-Buffer Solution with pH value greater than 7 is called Basic Buffer Solution.

* Bicine Chloride
* Tris Chloride
* Ches Sodium
* Ammonium Phosphate
* Ammonium Chloride

iii. Neutral Buffer Solution (CjIPPase):-

-Buffer Solution with pH value equal to 7 is called Neutral Buffer Solution.

* Hepes Sodium
* Bis Tris Chloride
* Bis Tris Methane
* Bis Tris Propane
* Cholamine Chloride

1. Precipitant

A substance that causes the precipitation of a specified substance is called precipitant. Six samples of precipitants were collected for the thesis work.

* Ammonium Phosphate
* Trisodium Citrate
* Potassium Phosphate
* PolyEthylene Glycol 8000
* PolyEthylene Glycol 4000
* PolyEthylene Monomethyl Ether

1. Salt

A salt is an ionic compound that can be formed by the neutralization reaction of an acid and a base.

Five samples of salt were collected for the purpose.

* Sodium Chloride
* Ammonium Dichlorate
* Magnesium Sulphate
* Sodium Bicarbonate
* Magnesium Chloride

1. Generate Initial Population

Different Precipitants were mixed with different salts and the mixture was put into the buffer solution. There were six different types of precipitants and five different types of salts and seventeen different types of buffer solutions. Altogether, there were 510 mixtures out of which the best two ones were selected as two parent chromosomes.

1. Tournament Selection

Tournament selection was the preferred choice for parent chromosome selection in the algorithm. The advantage of this approach was that it could identify top ranking reagents among the list and prioritize pairings with the best ranked reagents.

Tournament Selection lessens the possibility of random noise. One advantage of the tournament selection is its ability to handle either minimization or maximization problems without any structural changes. It is more efficient to code as it works on parallel architectures as well.

The parameter for tournament selection was the tournament size. Tournament size parameter set the size of the tournament during the parent chromosome selection process. If the tournament size was set to 1, a single random chromosome was selected from the population. If the tournament size was set to 3, then 3 random chromosomes were selected from the population and the fittest of the three was chosen as a parent chromosome. The second parent chromosome was selected by the same process.

Table 3.2: Relationship between Tournament Size and Selection Intensity

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tournament Size | 1 | 2 | 3 | 5 | 10 | 30 |
| Selection Intensity | 0 | 0.56 | 0.85 | 1.15 | 1.53 | 2.04 |

On the contrary, the tournament size should be maintained at a low value to ensure variation in the final output. It was observed that a tournament size between 1 and 3 usually had the best results.

1. CrossOver

The components of the parent chromosomes were swapped and the result was analyzed. Single point crossover was used. One of the three constituents was swapped among each other.

1. Mutation

The original dataset was maintained and a random gene picked up from the child chromosome was replaced by a corresponding gene from a randomly selected candidate in the original dataset.

1. Remove Prohibited Conditions

Certain combinations of reagents are known, either from literature or empirically, to produce bad results (precipitates or phase separation). Such combinations are defined in a list from the outcomes of AED experiment conducted previously. The final population generated by the algorithm was scanned for any prohibited conditions. All conditions that were deemed forbidden were removed from the final population.

1. Apply Ranking

The ranking was applied to the protein crystals formed on the basis of Hampton Scores and Revised Scores illustrated in Table 3.1 and shown in Figure 3.2.

1. Generate Output

The output was generated on the basis of the rankings applied and the equations of significance ratio of Precipitant P and Fitness score for a chromosome given by equation (3.3) and equation (3.4) respectively.

PC1

SODIUM CHLORIDE

HEPES SODIUM

AMMONIUM PHOSPHATE

PC2

MAGNESIUM CHLORIDE

TRIS CHLORIDE

POLYETHYLENE GLYCOL 4000

C R O S S O V E R

MAGNESIUM CHLORIDE

POLYETHYLENE GLYCOL 4000

HEPES SODIUM

CC

MUTATE

MAGNESIUM SULPHATE

HEPES SODIUM

POLYETHYLENE GLYCOL 4000

CC

Figure 3.4: Sample data showing the entire steps of the methodology

## 3.6. Implementations and Tools

### 3.6.1. Tools

Tools refer to the systems and software that may be used to develop the proposed system or to conduct any researches. These tools may be used to model some system, draw charts or perform mathematical analysis on the data. The tools that had been used for the development of the system can be given as:

### 3.6.1.1. Analysis and Design Tools

The analysis and design tools that were used can be given as:

* Data Flow Diagrams (DFD)
* Structured English
* Pseudo Code

### 3.6.1.2. Implementation Tools

The implementation tools that were used by the proposed system can be given as:

* CASE Tools
* IDE - Microsoft Visual Studio (C#)
* DBMS – Microsoft SQL Server

### 3.6.2. Implementation

At the end of the experiment, different proteins were used for analysis. For a single protein, the algorithm was run using different values for the algorithm parameters (i.e., population size, tournament size, number of iterations and mutation rate) and the output obtained from different sets would be combined to form a single output file.

AED analysis was already done on 3 proteins and the results were tested in the lab. The output for the same proteins was generated by using Genetic Algorithm and the results so obtained were compared to the results of AED for the same proteins.

This was implemented as a C# application using .NET 4.0 framework in Visual Studio 2015. The application has a simple GUI which allows users to upload a Microsoft Excel file with the results of a screening experiment. The file must be in a specific format for it to be read by the application (proper headers and data types). The output of the application was an MS Excel file that lists the candidate cocktails sorted by their ranks.

# 4. RESULTS AND DISCUSSIONS

## 4.1. Results

Acidic Buffer Solution produced the best result among the three types of buffer solutions classified on the basis of acidity and alkalinity. MES Sodium was found to be the best among all which produced 3D crystals when combined with Trisodium Citrate and Sodium Chloride.

Basic Buffer Solution was relatively the worst compared to the Acidic and the Neutral Buffer Solutions. The solution of Bicine Chloride produced the best result among Basic Buffer Solutions.

The quality of crystals Neutral Buffer Solution produced was better than what Basic Buffer Solution produced but poorer than what Acidic Buffer Solution produced.. Hepes Sodium among all neutral buffer solutions produced the best output when combined with Trisodium Citrate and Sodium Chloride.

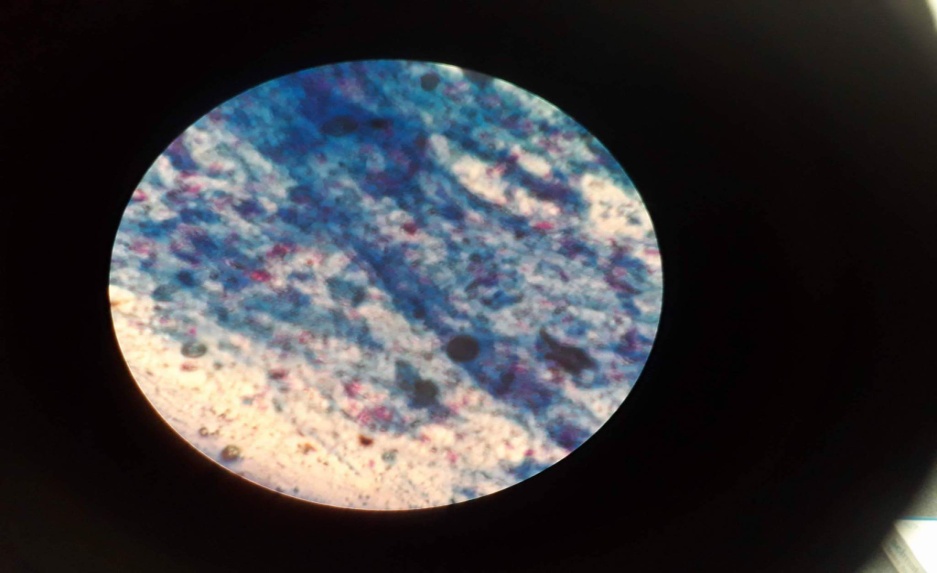


Figure 4.1: Crystals formation of a AbIPPase (Buffer: MES Sodium, Precipitant: Trisodium Citrate, Salt: Sodium Chloride; Used Stain: Methylene)

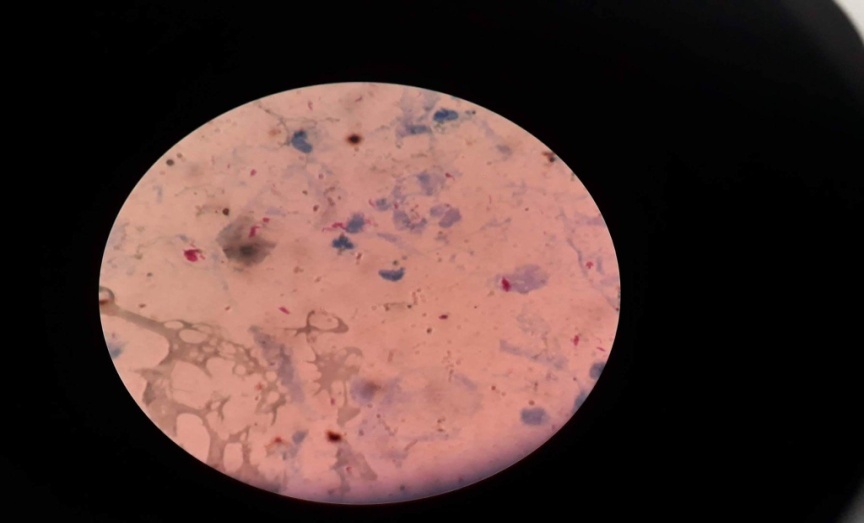


Figure 4.2: Crystals formation of a KbIPPase (Buffer: Bicine Chloride, Precipitant: Trisodium Citrate; Salt: Sodium Chloride; Used Stain: Carbol Fuschin)



Figure 4.3: Crystals formation of a CjIPPase (Buffer: Hepes Sodium, Precipitant: Trisodium Citrate, Salt: Sodium Chloride; Used Stain: Carbol Fuschin)

Figure 4.1 shows the protein crystals formation when the combination of Precipitant Trisodium Citrate and Salt Sodium Chloride was put into the Acidic Buffer Solution (AbIPPase) of MES Sodium. It took 23 generations for the output crystals to appear. The protein crystals thus formed appeared cleared milky white. The stain, Methylene, was put to give the milky white crystals a bright contrast so that it would be easier to read the appearance of the crystals. Methylene helped to make the solution bluish and the crystals were visible more clearly. The fitness score of the solution was calculated out to be 2.00.

Figure 4.2 depicts the formation of protein crystals when the Precipitant Trisodium Citrate and Salt Sodium Chloride were put into the Basic Buffer Solution (KbIPPase) of Bicine Chloride. It took 21 generations for the output crystals to appear. Milky white protein crystals appeared. So, Carbol Fuschin stain was used to make the solution orange to have a better view of the crystals. The fitness score of the solution was calculated out to be 1.8666.

Figure 4.3 illustrates the result of formation of Protein Crystals when Precipitant Trisodium Citrate and Salt Sodium Chloride was placed in the Neutral Buffer solution of Hepes Sodium. Orange crystals of protein were detected after 24 generations. The intensity of the colour was not as high as required for the proper illustrations of the crystals. Carbol Fuschin stain which produces orange colour was used to increase the intensity of the colour and make the crystals appear clearer. The fitness score of the solution was calculated out to be 1.9233.

In this thesis, since only the 2D and 3D crystals were considered fit for consideration, the highest possible fitness score was 2.00 for a combination of Buffer Solution (either Acidic, Basic or Neutral), precipitant and salt. On the basis of Hampton and Revised scores illustrated in Table 3.1 and depicted in Figure 3.2, integers from 0 to 9 are chosen to label the outcomes. For best case, the value 9 is selected and 0 is chosen for the worst case. Hence, the maximum possible significance ratio is,

= = 2.0

Ultimately, the maximum possible Fitness Score is 2.0.

But, the combination of Buffer Solution, Precipitate and Salt whose fitness score was below 1.35 was discarded as it was too low to be taken into consideration as per the objective of the thesis.

The combinations obtained so far are listed in the table as follows:-

1. Acidic Buffer Solutions (AbIPPase)

MES Sodium when mixed with TriSodium Citrate and Sodium Chloride produced the maximum possible fitness score of 2.0 and was the best combinations of all.

Sodium Phosphate Citrate was observed to be the second in the rank (after MES Sodium) to give the best output. When the mixture of PolyEthylene Glycol 8000 and Sodium Chloride was plunged into the solution of Sodium Phosphate Citrate, the fitness score of the output was obtained to be 1.77. Trisodium Citrate was the best precipitate to produce protein crystals. But, it didn’t produce desirable crystals when combined with Sodium Phosphate Citrate.

Similarly, the best obtained fitness score for Cacodylate Sodium was 1.73 when put together with Trisodium Citrate and Sodium Chloride.

Acidic Buffers like Sodium Citrate and Sodium Acetate were relatively less efficient in terms of the output and hence the fitness score obtained, 1.618 and 1.601 being their best respectively.

Other Acidic Buffer Solutions as Ethanoic Acid and Sodium Ethanoate didn’t prove better results for the first four generations so they were not used in the later generations of Genetic Algorithm.

Table 4.1: Fitness score for Acidic buffer solutions (AbIPPase) with different precipitants and salts.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.No. | Buffers | Precipitants | Salts | No. of Generations | Fitness Score | Rank | AED Rank |
| 1 | MES Sodium | Trisodium Citrate | Sodium Chloride | 23 | 2.0 | 1 | 1 |
| 2 | MES Sodium | Trisodium Citrate | Ammonium Dichlorate | 22 | 1.9046 | 2 | 2 |
| 3 | MES Sodium | Trisodium Citrate | Magnesium Chloride | 21 | 1.88 | 3 | 3 |
| 4 | MES Sodium | Trisodium Citrate | Magnesium Sulphate | 15 | 1.8737 | 4 | 5 |
| 5 | MES Sodium | Ammonium Phosphate | Sodium Chloride | 14 | 1.8666 | 5 | 6 |
| 6 | Sodium Phosphate Citrate | PolyEthylene Glycol 8000 | Sodium Chloride | 19 | 1.7913 | 6 | 12 |
| 7 | MES Sodium | Trisodium Citrate | Sodium Bicarbonate | 18 | 1.7886 | 7 | 4 |
| 8 | MES Sodium | Ammonium Phosphate | Ammonium Dichlorate | 17 | 1.7713 | 8 | 7 |
| 9 | Sodium Phosphate Citrate | PolyEthylene Glycol 4000 | Sodium Chloride | 21 | 1.77 | 9 | 17 |
| 10 | MES Sodium | Ammonium Phosphate | Magnesium Chloride | 16 | 1.7533 | 10 | 10 |
| 11 | MES Sodium | Ammonium Phosphate | Magnesium Sulphate | 20 | 1.7404 | 11 | 8 |
| 12 | Cacodylate Sodium | Trisodium Citrate | Sodium Chloride | 22 | 1.73 | 12 | 21 |
| 13 | Sodium Phosphate Citrate | PolyEthylene Glycol 8000 | Ammonium Dichlorate | 20 | 1.696 | 13 | 13 |
| 14 | Sodium Phosphate Citrate | PolyEthylene Glycol 4000 | Ammonium Dichlorate | 19 | 1.678 | 14 | 18 |
| 15 | Sodium Phosphate Citrate | PolyEthylene Glycol 8000 | Magnesium Chloride | 18 | 1.678 | 15 | 11 |
| 16 | Sodium Phosphate Citrate | PolyEthylene Glycol 8000 | Magnesium Sulphate | 17 | 1.6651 | 16 | 14 |
| 17 | Sodium Phosphate Citrate | PolyEthylene Glycol 4000 | Magnesium Chloride | 16 | 1.66 | 17 | 16 |
| 18 | MES Sodium | Ammonium Phosphate | Sodium Bicarbonate | 19 | 1.6553 | 18 | 9 |
| 19 | Sodium Phosphate Citrate | PolyEthylene Glycol 4000 | Magnesium Sulphate | 15 | 1.6471 | 19 | 19 |
| 20 | Cacodylate Sodium | Trisodium Citrate | Ammonium Dichlorate | 21 | 1.638 | 20 | 22 |
| 21 | Cacodylate Sodium | Trisodium Citrate | Magnesium Chloride | 20 | 1.62 | 21 | 25 |
| 22 | Cacodylate Sodium | PolyEthylene Glycol 4000 | Sodium Chloride | 19 | 1.62 | 22 | 26 |
| 23 | Sodium Citrate | PolyEthylene Glycol 8000 | Sodium Chloride | 21 | 1.618 | 23 | 31 |
| 24 | Cacodylate Sodium | Trisodium Citrate | Magnesium Sulphate | 18 | 1.6071 | 24 | 23 |
| 25 | Sodium Acetate | PolyEthylene Glycol 8000 | Sodium Chloride | 19 | 1.601 | 25 | 41 |
| 26 | Sodium Citrate | PolyEthylene Glycol 4000 | Sodium Chloride | 20 | 1.6 | 26 | 36 |
| 27 | Sodium Acetate | PolyEthylene Glycol 4000 | Sodium Chloride | 18 | 1.5833 | 27 | 46 |
| 28 | Sodium Phosphate Citrate | PolyEthylene Glycol 8000 | Sodium Bicarbonate | 14 | 1.58 | 28 | 15 |
| 29 | Sodium Phosphate Citrate | PolyEthylene Glycol 4000 | Sodium Bicarbonate | 13 | 1.562 | 29 | 20 |
| 30 | Cacodylate Sodium | PolyEthylene Glycol 4000 | Ammonium Dichlorate | 17 | 1.5246 | 30 | 27 |
| 31 | Sodium Citrate | PolyEthylene Glycol 8000 | Ammonium Dichlorate | 19 | 1.5226 | 31 | 32 |
| 32 | Cacodylate Sodium | PolyEthylene Glycol 4000 | Magnesium Chloride | 16 | 1.5066 | 32 | 30 |
| 33 | Sodium Acetate | PolyEthylene Glycol 8000 | Ammonium Dichlorate | 17 | 1.5060 | 33 | 42 |
| 34 | Sodium Citrate | PolyEthylene Glycol 8000 | Magnesium Chloride | 18 | 1.5046 | 34 | 35 |
| 35 | Sodium Citrate | PolyEthylene Glycol 4000 | Ammonium Dichlorate | 17 | 1.5046 | 35 | 37 |
| 36 | Cacodylate Sodium | PolyEthylene Glycol 4000 | Magnesium Sulphate | 15 | 1.4937 | 36 | 28 |
| 37 | Sodium Citrate | PolyEthylene Glycol 8000 | Magnesium Sulphate | 16 | 1.4917 | 37 | 33 |
| 38 | Sodium Acetate | PolyEthylene Glycol 8000 | Magnesium Chloride | 16 | 1.488 | 38 | 45 |
| 39 | Sodium Acetate | PolyEthylene Glycol 4000 | Ammonium Dichlorate | 15 | 1.488 | 39 | 47 |
| 40 | Sodium Citrate | PolyEthylene Glycol 4000 | Magnesium Chloride | 15 | 1.4866 | 40 | 40 |
| 41 | Sodium Acetate | PolyEthylene Glycol 8000 | Magnesium Sulphate | 14 | 1.4751 | 41 | 43 |
| 42 | Sodium Acetate | PolyEthylene Glycol 4000 | Magnesium Chloride | 13 | 1.47 | 42 | 50 |
| 43 | Sodium Acetate | PolyEthylene Glycol 4000 | Magnesium Sulphate | 12 | 1.4571 | 43 | 48 |
| 44 | Cacodylate Sodium | PolyEthylene Glycol 4000 | Sodium Bicarbonate | 14 | 1.4086 | 44 | 29 |
| 45 | Sodium Citrate | PolyEthylene Glycol 8000 | Sodium Bicarbonate | 14 | 1.4066 | 45 | 34 |
| 46 | Sodium Acetate | PolyEthylene Glycol 8000 | Sodium Bicarbonate | 11 | 1.39 | 46 | 44 |
| 47 | Sodium Citrate | PolyEthylene Glycol 4000 | Sodium Bicarbonate | 13 | 1.3886 | 47 | 39 |
| 48 | Sodium Acetate | PolyEthylene Glycol 4000 | Sodium Bicarbonate | 10 | 1.372 | 48 | 49 |

1. Basic Buffer Solutions (KbIPPase)

Bicine Chloride proved to be the best in producing desirable protein crystals among the Basic Buffer Solutions. When the combination of Trisodium Citrate and Sodium Chloride was blended with Bicine Chloride, the best fitness score 1.9166 was obtained.

Tris Chloride was ranked second after Bicine Chloride on the basis of results produced. The fitness score was 1.75 when Trisodium Citrate and Sodium Chloride were mixed with Tris Chloride.

The maximum score that Ches Sodium could yield was 1.608 when PolyEthylene Glycol 8000 and Sodium Chloride were plunged into its solution. Precipitant Trisodium Citrate didn’t prove to be successful for Ches Sodium as only a clear solution (Hampton and Revised Score: 1) was obtained.

Other Basic Buffer Solutions as Ammonium Phosphate and Ammonium Chloride didn’t produce good results hence they were left out for further generations.

Table 4.2: Fitness score for Basic buffer solutions (KbIPPase) with different precipitants and salts.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.No | Buffers | Precipitants | Salts | Fitness Score | No. of Generations | Rank | AED Rank |
| 1 | Bicine Chloride | Trisodium Citrate | Sodium Chloride | 1.9166 | 21 | 1 | 1 |
| 2 | Bicine Chloride | Trisodium Citrate | Ammonium Dichlorate | 1.8213 | 20 | 2 | 2 |
| 3 | Bicine Chloride | PolyEthylene Glycol 8000 | Sodium Chloride | 1.8213 | 19 | 3 | 3 |
| 4 | Bicine Chloride | PolyEthylene Glycol 4000 | Sodium Chloride | 1.8033 | 18 | 4 | 11 |
| 5 | Bicine Chloride | Trisodium Citrate | Magnesium Chloride | 1.8033 | 17 | 5 | 5 |
| 6 | Bicine Chloride | Trisodium Citrate | Magnesium Sulphate | 1.7904 | 16 | 6 | 6 |
| 7 | Tris Chloride | Trisodium Citrate | Sodium Chloride | 1.75 | 20 | 7 | 9 |
| 8 | Bicine Chloride | PolyEthylene Glycol 8000 | Ammonium Dichlorate | 1.726 | 15 | 8 | 8 |
| 9 | Bicine Chloride | PolyEthylene Glycol 8000 | Magnesium Chloride | 1.708 | 14 | 9 | 7 |
| 10 | Bicine Chloride | PolyEthylene Glycol 4000 | Ammonium Dichlorate | 1.708 | 13 | 10 | 10 |
| 11 | Bicine Chloride | Trisodium Citrate | Sodium Bicarbonate | 1.7055 | 12 | 11 | 4 |
| 12 | Bicine Chloride | PolyEthylene Glycol 8000 | Magnesium Sulphate | 1.6951 | 11 | 12 | 12 |
| 13 | Bicine Chloride | PolyEthylene Glycol 4000 | Magnesium Chloride | 1.69 | 10 | 13 | 18 |
| 14 | Bicine Chloride | PolyEthylene Glycol 4000 | Magnesium Sulphate | 1.6771 | 9 | 14 | 14 |
| 15 | Tris Chloride | Trisodium Citrate | Ammonium Dichlorate | 1.6546 | 19 | 15 | 16 |
| 16 | Tris Chloride | PolyEthylene Glycol 8000 | Sodium Chloride | 1.6546 | 18 | 16 | 15 |
| 17 | Tris Chloride | Trisodium Citrate | Magnesium Chloride | 1.6366 | 17 | 17 | 13 |
| 18 | Tris Chloride | PolyEthylene Monomethyl Ether | Sodium Chloride | 1.6366 | 16 | 18 | 17 |
| 19 | Tris Chloride | Trisodium Citrate | Magnesium Sulphate | 1.6237 | 15 | 19 | 25 |
| 20 | Bicine Chloride | PolyEthylene Glycol 8000 | Sodium Bicarbonate | 1.61 | 8 | 20 | 19 |
| 21 | Ches Sodium | PolyEthylene Glycol 8000 | Sodium Chloride | 1.608 | 19 | 21 | 20 |
| 22 | Bicine Chloride | PolyEthylene Glycol 4000 | Sodium Bicarbonate | 1.592 | 7 | 22 | 21 |
| 23 | Tris Chloride | PolyEthylene Glycol 8000 | Ammonium Dichlorate | 1.5593 | 14 | 23 | 24 |
| 24 | Tris Chloride | PolyEthylene Glycol 8000 | Magnesium Chloride | 1.5413 | 13 | 24 | 25 |
| 25 | Tris Chloride | PolyEthylene Monomethyl Ether | Ammonium Dichlorate | 1.5413 | 12 | 25 | 26 |
| 26 | Tris Chloride | Trisodium Citrate | Sodium Bicarbonate | 1.5386 | 11 | 26 | 22 |
| 27 | Tris Chloride | PolyEthylene Glycol 8000 | Magnesium Sulphate | 1.5284 | 10 | 27 | 23 |
| 28 | Tris Chloride | PolyEthylene Monomethyl Ether | Magnesium Chloride | 1.5233 | 9 | 28 | 27 |
| 29 | Ches Sodium | PolyEthylene Glycol 8000 | Ammonium Dichlorate | 1.5126 | 18 | 29 | 28 |
| 30 | Tris Chloride | PolyEthylene Monomethyl Ether | Magnesium Sulphate | 1.5104 | 8 | 30 | 29 |
| 31 | Tris Chloride | PolyEthylene Monomethyl Ether | Sodium Bicarbonate | 1.5066 | 7 | 31 | 30 |
| 32 | Ches Sodium | PolyEthylene Glycol 8000 | Magnesium Chloride | 1.4946 | 17 | 32 | 32 |
| 33 | Ches Sodium | PolyEthylene Glycol 8000 | Magnesium Sulphate | 1.4817 | 16 | 33 | 31 |
| 34 | Tris Chloride | PolyEthylene Glycol 8000 | Sodium Bicarbonate | 1.4433 | 6 | 34 | 34 |
| 35 | Ches Sodium | PolyEthylene Glycol 8000 | Sodium Bicarbonate | 1.3966 | 15 | 35 | 33 |

1. Neutral Buffer Solutions (CjIPPase)

Of all the Neutral Buffer Solutions, Hepes Sodium produced the best fitness score 1.9233 when combined with Trisodium Citrate and Sodium Chloride.

Likewise, BisTris Chloride was considered to be the second in the rank which gave the fitness score of 1.60 when combined with PolyEthylene Glycol 4000 and Sodium Chloride.

No other Neutral Buffer Solutions, namely, Bis Tris Methane, Bis Tris Propane, Cholamine Chloride could produce effective results hence they had to be ruled out to compete in further generations.

Table 4.3: Fitness score for Neutral buffer solutions (CjIPPase) with different precipitants and salts.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.No. | Buffers | Precipitants | Salts | Fitness Score | No. of Generations | Rank | AEDRank |
| 1 | Hepes Sodium | Trisodium Citrate | Sodium Chloride | 1.9233 | 24 | 1 | 1 |
| 2 | Hepes Sodium | Trisodium Citrate | Ammonium Dichlorate | 1.90 | 10 | 2 | 2 |
| 3 | Hepes Sodium | Trisodium Citrate | Magnesium Chloride | 1.81 | 22 | 3 | 3 |
| 4 | Hepes Sodium | PolyEthylene Glycol 4000 | Sodium Chloride | 1.81 | 23 | 4 | 4 |
| 5 | Hepes Sodium | Trisodium Citrate | Magnesium Sulphate | 1.7971 | 15 | 5 | 8 |
| 6 | Hepes Sodium | Ammonium Phosphate | Sodium Chloride | 1.79 | 19 | 6 | 11 |
| 7 | Hepes Sodium | PolyEthylene Glycol 4000 | Ammonium Dichlorate | 1.7146 | 18 | 7 | 18 |
| 8 | Hepes Sodium | Trisodium Citrate | Sodium Bicarbonate | 1.712 | 14 | 8 | 5 |
| 9 | Hepes Sodium | PolyEthylene Glycol 4000 | Magnesium Chloride | 1.6966 | 12 | 9 | 6 |
| 10 | Hepes Sodium | Ammonium Phosphate | Ammonium Dichlorate | 1.6946 | 15 | 10 | 7 |
| 11 | Hepes Sodium | PolyEthylene Glycol 4000 | Magnesium Sulphate | 1.6837 | 17 | 11 | 13 |
| 12 | Hepes Sodium | Ammonium Phosphate | Magnesium Chloride | 1.6766 | 19 | 12 | 12 |
| 13 | Hepes Sodium | Potassium Phosphate | Sodium Chloride | 1.6711 | 16 | 13 | 10 |
| 14 | Hepes Sodium | Ammonium Phosphate | Magnesium Sulphate | 1.6637 | 11 | 14 | 9 |
| 15 | Bis Tris Chloride | PolyEthylene Glycol 4000 | Sodium Chloride | 1.60 | 13 | 15 | 17 |
| 16 | Hepes Sodium | PolyEthylene Glycol 4000 | Sodium Bicarbonate | 1.5986 | 121 | 16 | 16 |
| 17 | Hepes Sodium | Ammonium Phosphate | Sodium Bicarbonate | 1.5786 | 9 | 17 | 15 |
| 18 | Hepes Sodium | Potassium Phosphate | Ammonium Dichlorate | 1.5757 | 8 | 18 | 14 |
| 19 | Hepes Sodium | Potassium Phosphate | Magnesium Chloride | 1.5577 | 7 | 19 | 19 |
| 20 | Hepes Sodium | Potassium Phosphate | Magnesium Sulphate | 1.5448 | 6 | 20 | 20 |
| 21 | Bis Tris Chloride | PolyEthylene Glycol 4000 | Ammonium Dichlorate | 1.5113 | 18 | 21 | 21 |
| 22 | Bis Tris Chloride | PolyEthylene Glycol 4000 | Magnesium Chloride | 1.4933 | 17 | 22 | 23 |
| 23 | Bis Tris Chloride | PolyEthylene Glycol 4000 | Magnesium Sulphate | 1.4804 | 16 | 23 | 25 |
| 24 | Hepes Sodium | Potassium Phosphate | Sodium Bicarbonate | 1.4597 | 5 | 24 | 24 |
| 25 | Bis Tris Chloride | PolyEthylene Glycol 4000 | Sodium Bicarbonate | 1.3953 | 15 | 25 | 22 |

## 4.2. Discussions

The crystals obtained using AED were 2D while Genetic Algorithm produced 3D crystals for the same combination of reagents for 12 cases out of 108 different samples analyzed. Due to this reason, the ranks of such outcomes were higher in position using Genetic Algorithm compared to that when AED was used.

On the other hand, some crystal outputs were placed lower in ranks as compared to the same when AED was used. The only reason behind it was the higher ranks been taken by the 3D crystals, as mentioned above. Had the significance ratio been same and ultimately the fitness function, the fitness score of those crystals analyzed using AED would have been lower than the fitness score of the crystals analyzed using Genetic Algorithm with the same input constituents.

Figure 4.4: Graph depicting relationship of Fitness Score and No. of Generations for MES Sodium

Figure 4.5: Graph depicting relationship of Fitness Score and No. of Generations for Bicine Chloride

Figure 4.6 : Graph depicting relationship of Fitness Score and No. of Generations for Hepes Sodium

In Figure 4.4, highest score was obtained in 23 iterations for MES Sodium while highest score was obtained in 21 iterations for Bicine Chloride and in 24 iterations for Hepes Sodium as shown in Figure 4.5 and Figure 4.6 respectively.

There are some fluctuations in the score and higher number of iterations does not always mean higher score. There is a limit on the highest score that can be obtained from a dataset and that score fluctuates based on the number of iterations. The exact number of iterations required cannot be known initially, but it can be estimated by running the algorithm for several iterations.

# 5. CONCLUSION AND RECOMMENDATIONS

## 5.1. Conclusion

Genetic Algorithm was introduced to produce crystalline crystals by analyzing results of existing experiments. For proteins AbIPPase, CjIPPase and KbIPPase, Genetic Algorithm had an average 25.92% (28 of 108) overlap with respect to the AED crystalline conditions. The output conditions of Genetic Algorithm for protein was tested and 32 crystalline conditions (not including 28 overlaps with the AED), were obtained, which was 20 for the AED. This shows that GenScreen is an effective protein crystallization screening method.

Acidic Buffer Solutions were determined to be the best to produce wel-ordered crystals of proteins followed by Neutral Buffer Solutions. Basic Buffer Solutions were deduced to be the worst among the three on the basis of the quality of results produced.

The output of the algorithm was determined by 4 parameters (iteration count, population size, tournament size and mutation rate). At the time of the analysis, the values were assigned intuitively by observing the output. For high ranking conditions, higher values of tournament size and iterations were favorable. Variations in the family were generally desired. Due to the randomized nature of the algorithm, it did not produce the same set of results at each run

## 5.2. Limitations

The main limitation of using Genetic Algorithm was its tendency to converge at a local optimum and there was no guarantee of finding the global optimum. Likewise, a few of other limitations are as follows:-

* Time taken for convergence
* Genetic Algorithms specifically assume discrete domains and perform poorly on continuous domains.

## 5.3. Future Enhancements

* The problem of getting stuck into a local optimum can be overcome by hybridization by selecting a local optimum by Genetic Algorithm and using Simulated Annealing based on the initial solutions found from Genetic Algorithm.
* Genetic Algorithm with incorporation of Grid Screening can study each and every case on a continuous domain.
* The salts can be further broken down into Cations and Anions and analyzed more precisely.

# 6. REFERENCES

[1] Giege, R. (2013). A historical perspective on protein crystallization from 1840

to the present day. FEBS *Journal*, 280(24):6456-6497.

[2] McPherson, A. and Gavira, J. A. (2014). Introduction to protein crystallization.

*Acta Crystallographica Section F: Structural Biology Communications*, 70(1):2-20.

[3] Blundell TL, Patel S. High-throughput X-ray crystallography for drug discovery*. Curr Opin Pharmacol* 2004, 4:490-496.

[4] Kuhn P, Wilson K, Patch MG, Stevens RC. The genesis of high-throughput structure-based drug discovery using protein crystallography. *Curr Opin Chem Biol* 2002, 6:704-710.

[5] Hardy LW, Malikayil A. The impact of structure-guided drug design on clinical agents. *Curr Drug Discov* 2003, (Dec): 15-20.

[6] Sharff A, Jhoti H. High-throughput crystallography to enhance drug discovery. *Curr Opin Chem Biol* 2003, 7:340-345.

[7] Williams, Shawn P., Lee F. Kuyper, and Kenneth H. Pearce. Recent applications of protein crystallography and structure-guided drug design. *Current opinion in chemical biology* 9.4 (2005): 371-380.

[8] Sigdel, Madhav, Marc L. Pusey, and Ramazan S. Aygun. CrystPro: Spatiotemporal Analysis of Protein Crystallization Images. *Crystal growth & design* 15.11 (2015): 5254-5262.

[9] Montgomery, Douglas C. *Design and analysis of experiments*. John Wiley & Sons, 2008.

[10] Hurrion, R. D., and S. Birgil. A comparison of factorial and random experimental design methods for the development of regression and neural network simulation metamodels. *Journal of the operational research society*50.10 (1999): 1018-1033.

[11] Creswell, John W. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2013.

[12] Carter Jr, Charles W., and Charles W. Carter. Protein crystallization using incomplete factorial experiments. *J. Biol. Chem* 254.23 (1979): 12219-12223.

[13] Jancarik, J. A. K. S., and S-H. Kim. "Sparse matrix sampling: a screening method for crystallization of proteins." *Journal of applied crystallography* 24.4 (1991): 409-411.

[14] DeLucas, Lawrence J., et al. "Efficient protein crystallization." *Journal of structural biology* 142.1 (2003): 188-206.

[15] Johnson, David H., W. William Wilson, and Lawrence J. DeLucas. "Protein solubilization: A novel approach." *Journal of Chromatography B* 971 (2014): 99-106.

[16] Saveski, Martin, and Amin Mantrach. "Item cold-start recommendations: learning local collective embeddings." *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 2014.

[17] Sharma, Lalita, and Anju Gera. "A survey of recommendation system: Research challenges." *International Journal of Engineering Trends and Technology (IJETT)* 4.5 (2013): 1989-1992.

[18] Dinç, İmren, Marc L. Pusey, and Ramazan S. Aygün. "Optimizing Associative Experimental Design for Protein Crystallization Screening." *IEEE transactions on nanobioscience* 15.2 (2016): 101-112.

[19] Saridakis, Emmanuel. "Novel Genetic Algorithm-inspired concept for macromolecular crystal optimization." *Crystal Growth & Design* 11.7 (2011): 2993-2998.

[20] Rupp, Bernhard. "Origin and use of crystallization phase diagrams." *Acta Crystallographica Section F: Structural Biology Communications* 71.3 (2015): 247-260.

[21] Asherie, Neer. "Protein crystallization and phase diagrams." *Methods* 34.3 (2004): 266-272.

[22] Hampton Research Scoring. https://hamptonresearch.com/ documents/product/hr005490 2-130 user guide.pdf. Accessed: 2015-11-01.

[23] Alabsi, Firas, and Reyadh Naoum. Comparison of selection methods and crossover operations using steady state genetic based intrusion detection system. *Journal of Emerging Trends in Computing and Information Sciences*3.7 (2012): 1053-1058.

**APPENDIX**

1. **APPENDIX – A**

Figure A-1 to A-4 represent the data for Acidic Buffer Solutions while Figure A-5 and A-6 depict the same for Basic Buffer Solutions. Likewise, , Figure A-7 represents the data for Neutral Buffer Solution(Bis Tris Chloride).

In Figure A-1, highest score was 1.77 obtained in 21 iterations for Sodium Phosphate Citrate while highest score obtained for Cacodylate Sodium was 1.73 in 22 iterations shown in Figure A-2. Likewise, the maximum score for Sodium Citrate was 1.618 in 21 iterations and for Sodium Acetate, it was 1.601 in 19 iterations as shown in Figure A-3 and Figure A-4 respectively.

Similarly, Figure A-5 shows the data for Tris Chloride where the highest score was 1.75 in 20 iterations. The maximum fitness score for Ches Sodium was 1.608 in 19 generations shown in Figure A-6.

Likewise, Figure A-7 depicts the data of Bis Tris Chloride. The maximum value of fitness score for Bis Tris Chloride was found to be 1.60 in 19 generations.

There are some fluctuations in the score and higher number of iterations does not always mean higher score. There is a limit on the highest score that can be obtained from a dataset and that score fluctuates based on the number of iterations. The exact number of iterations required cannot be known initially, but it can be estimated by running the algorithm for several iterations.

The graphical relations between Number of Generations and Fitness Score of different Buffer Solutions are as follows:-

1. Sodium Phosphate Citrate

Figure A-1: Graph depicting relationship of Fitness Score and No. of Generations for Sodium Phosphate Citrate

1. Cacodylate Sodium

Figure A-2: Graph depicting relationship of Fitness Score and No. of Generations for Cacodylate Sodium

1. Sodium Citrate

Figure A-3: Graph depicting relationship of Fitness Score and No. of Generations for Sodium Citrate

1. Sodium Acetate

Figure A-4: Graph depicting relationship of Fitness Score and No. of Generations for Sodium Acetate

1. Tris Chloride

Figure A-5: Graph depicting relationship of Fitness Score and No. of Generations for Tris Chloride

1. Ches Sodium

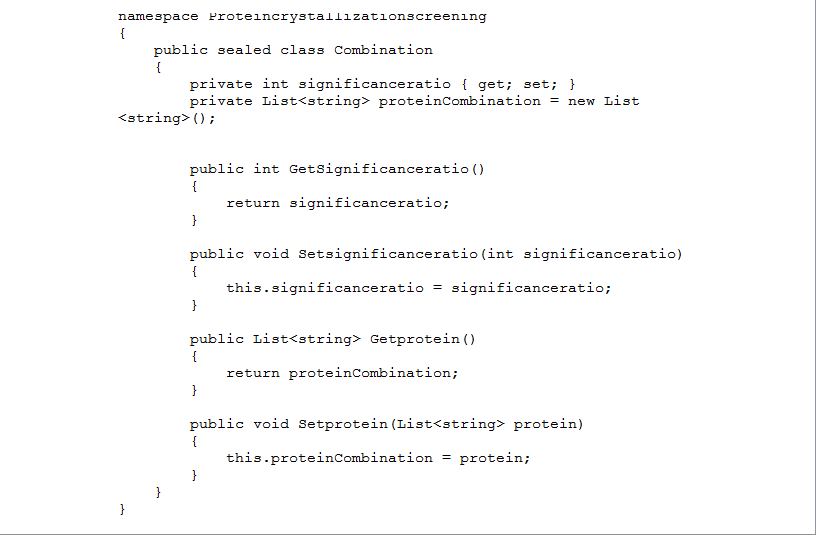
Figure A-6: Graph depicting relationship of Fitness Score and No. of Generations for Ches Sodium

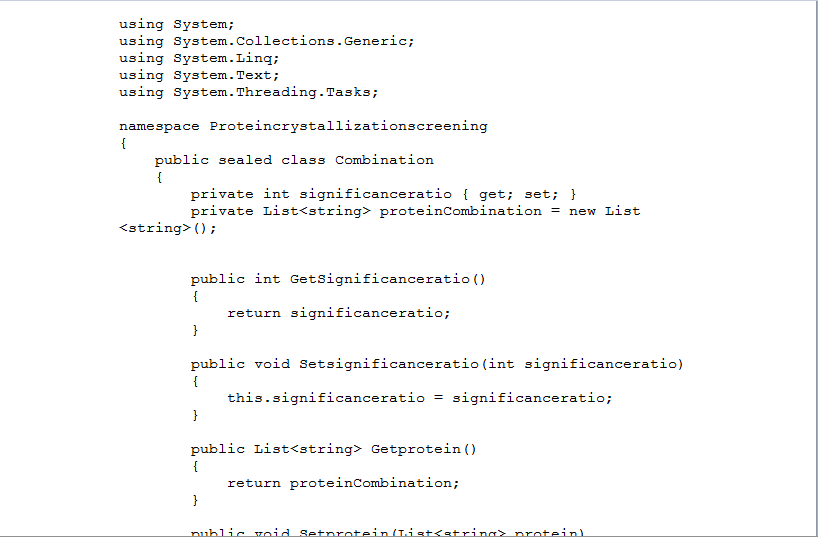
1. Bis Tris Chloride

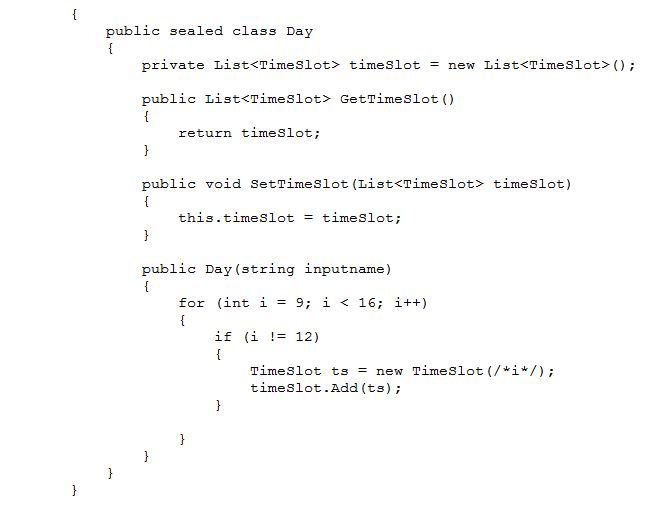
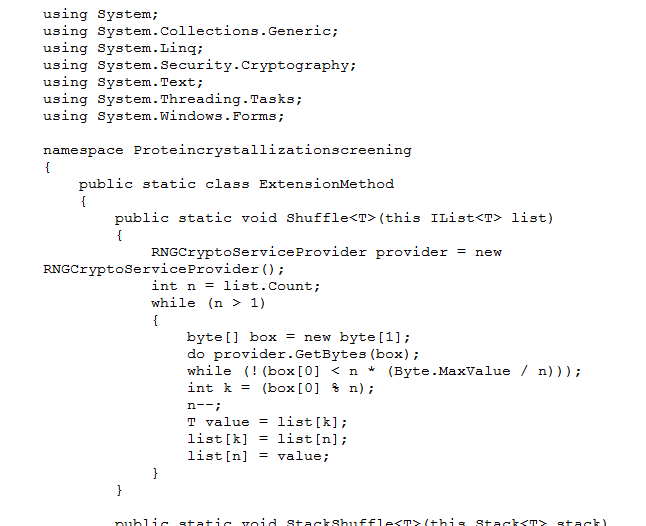
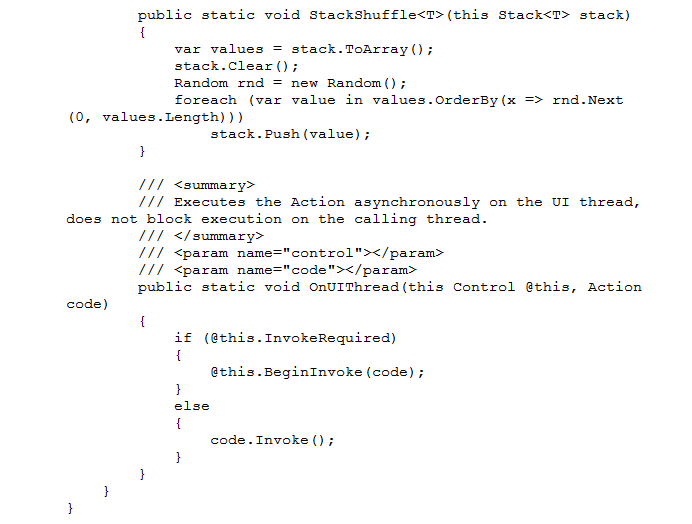
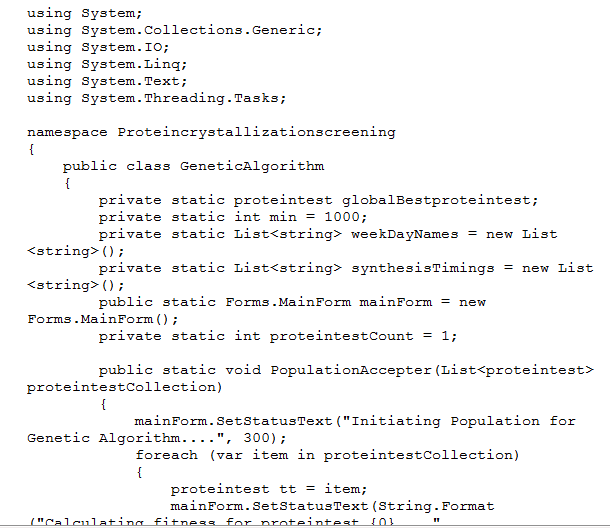
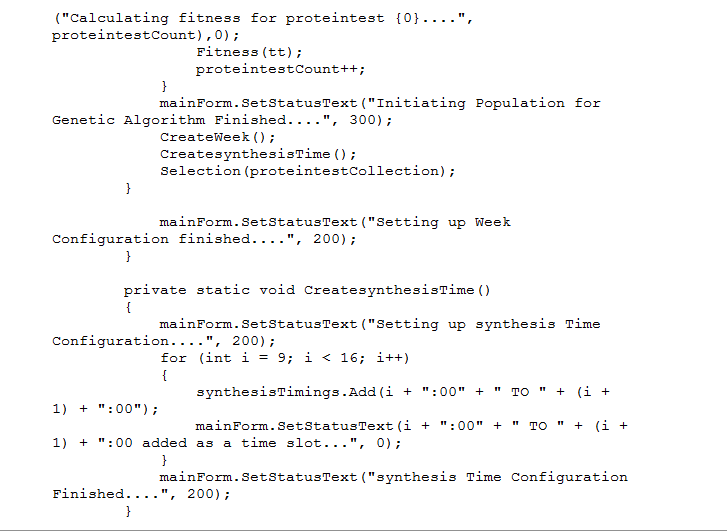
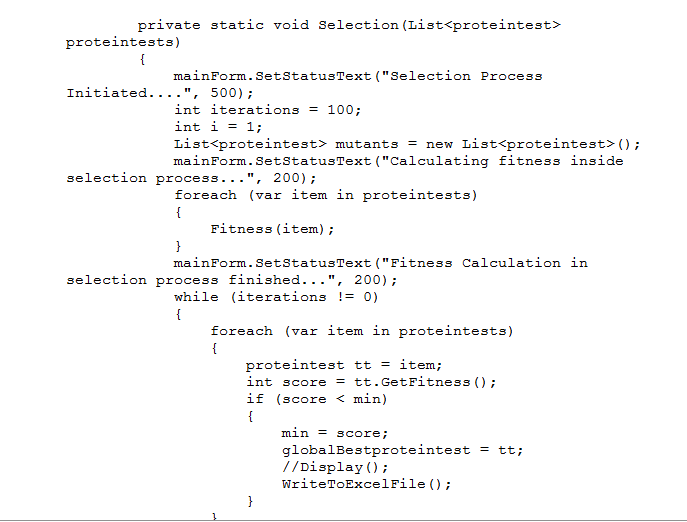
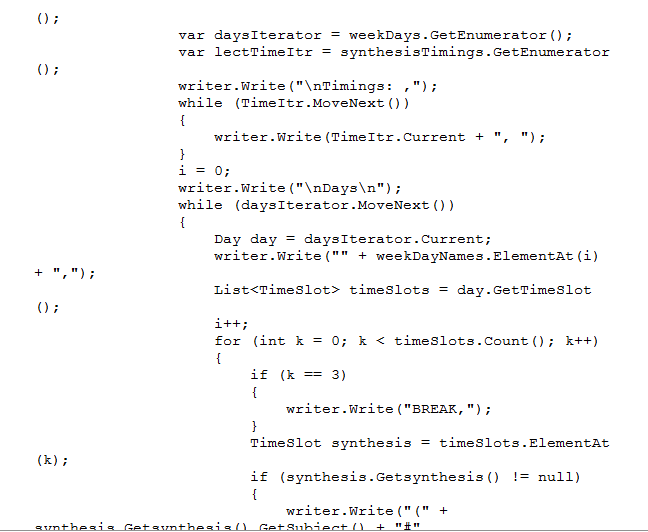
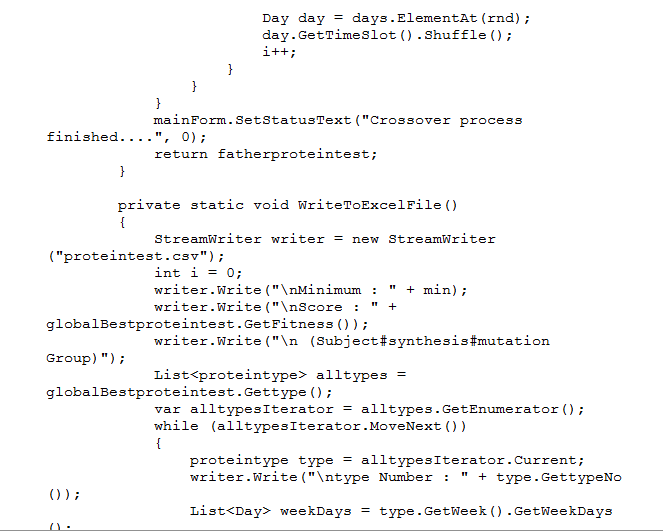
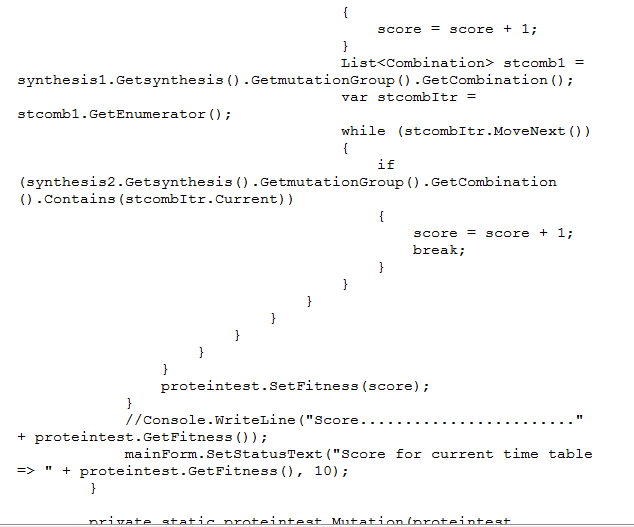
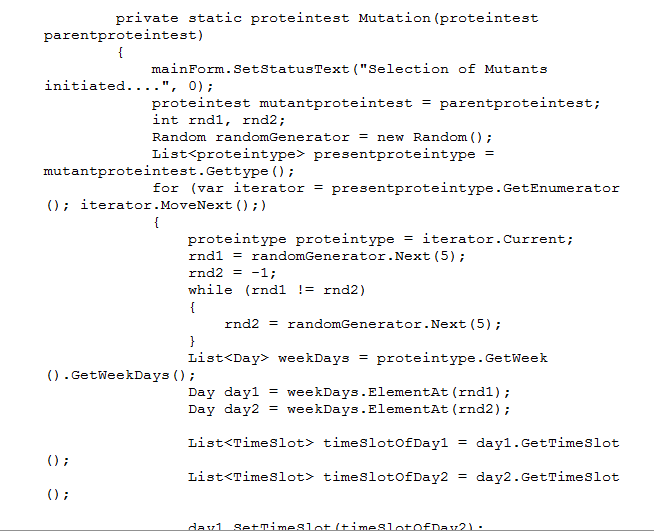
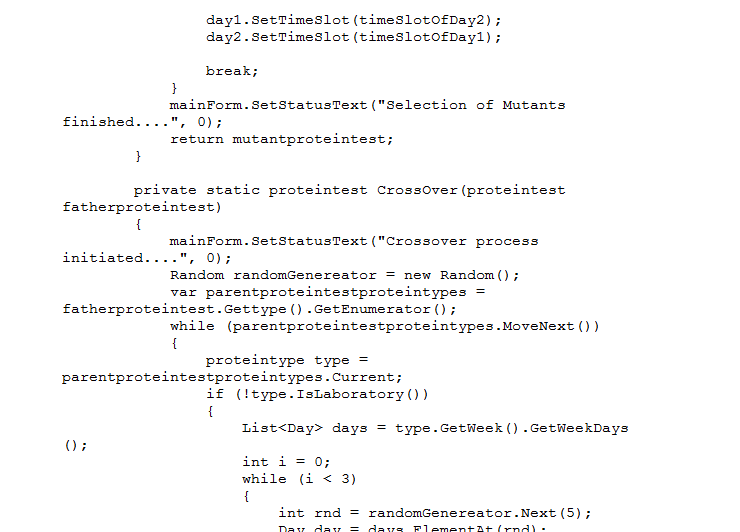
Figure A-3: Graph depicting relationship of Fitness Score and No. of Generations for Bis Tris Chloride

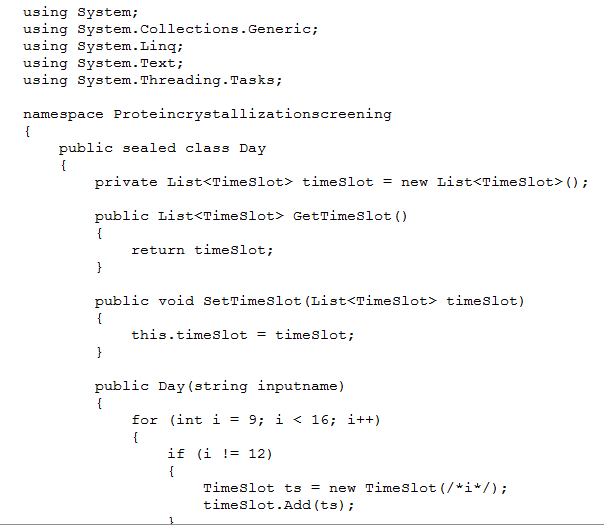
1. **APPENDIX – B**

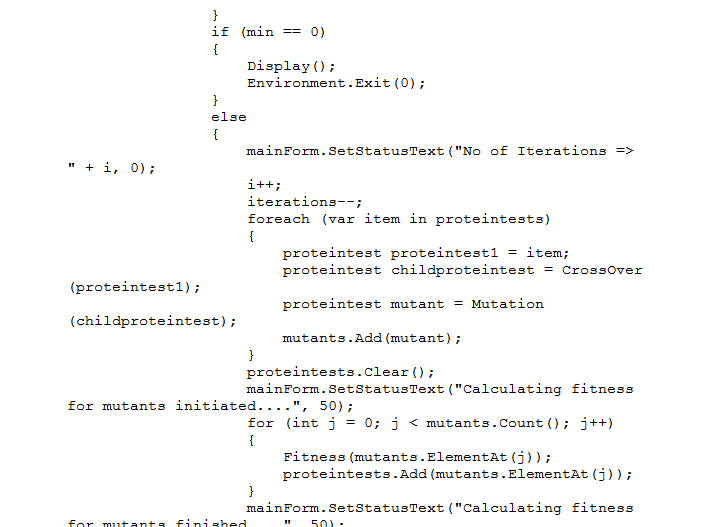
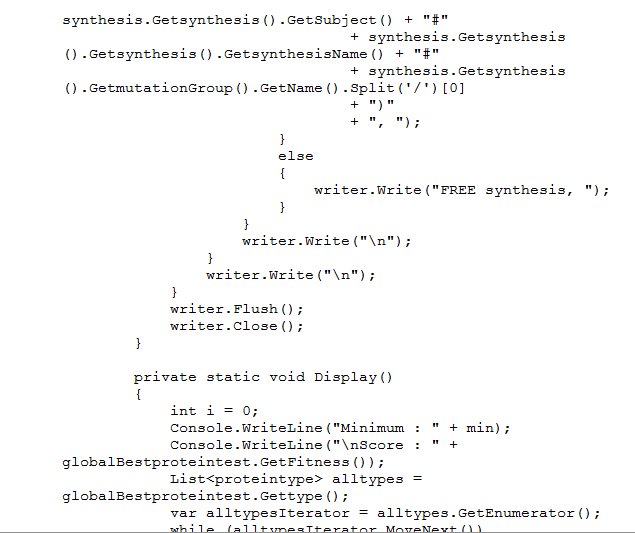
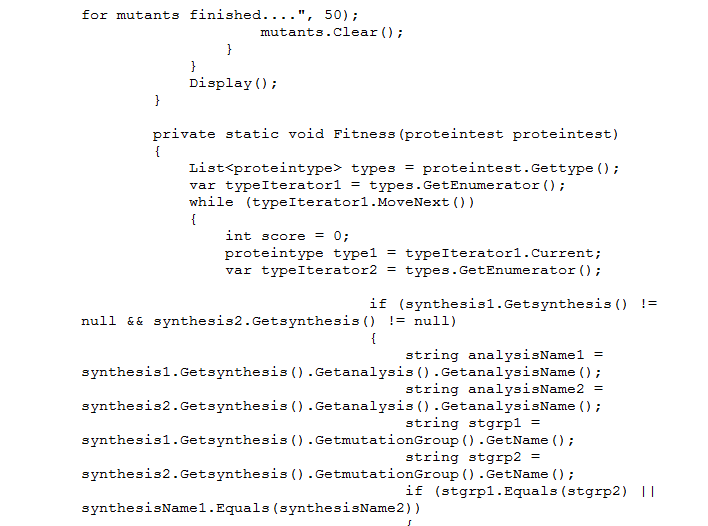
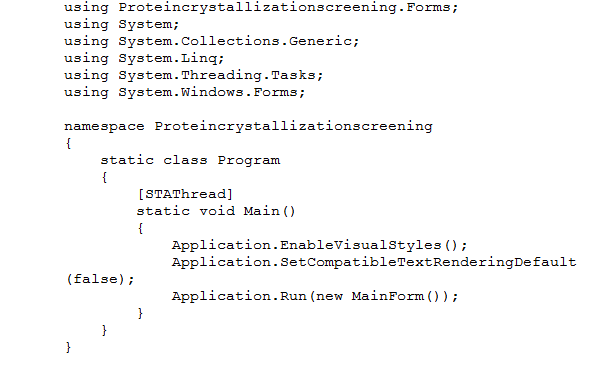
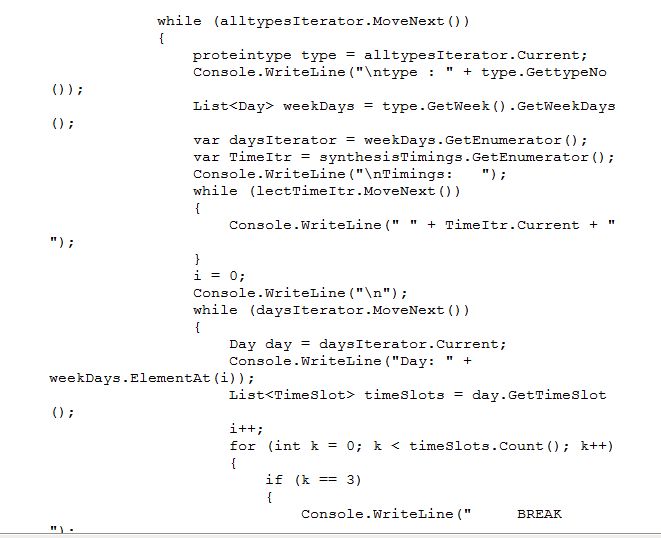
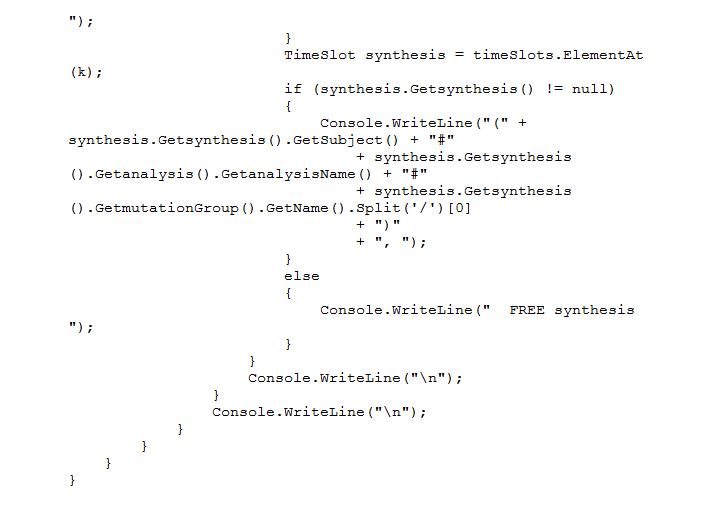
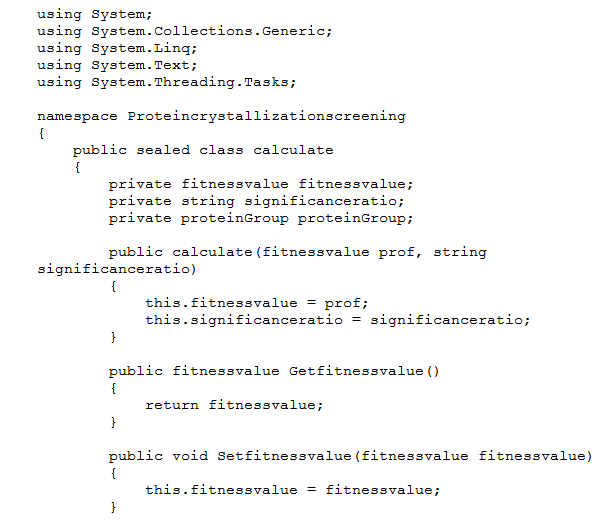
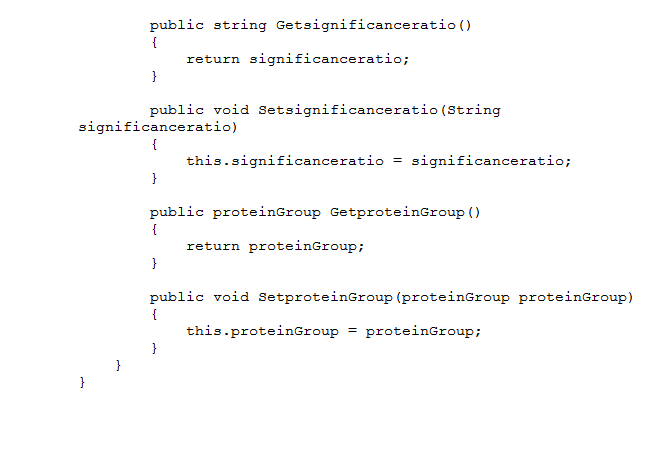
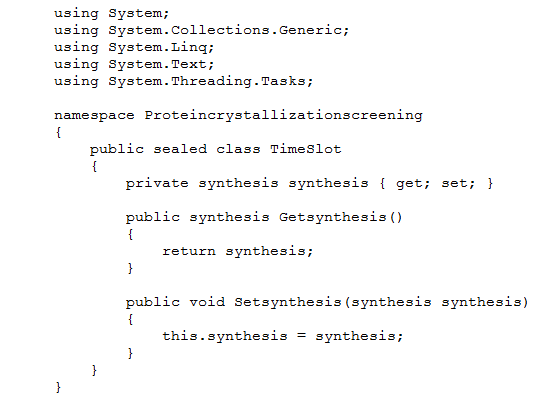
The experimental work was implemented as a C# application using .NET 4.0 framework in Visual Studio 2015. The application has a simple GUI which allows users to upload a Microsoft Excel file with the results of a screening experiment. The snapshots of certain source codes together with the snapshot of the output of Acidic Buffer Solution only, which is an excel file are illustrated below:-

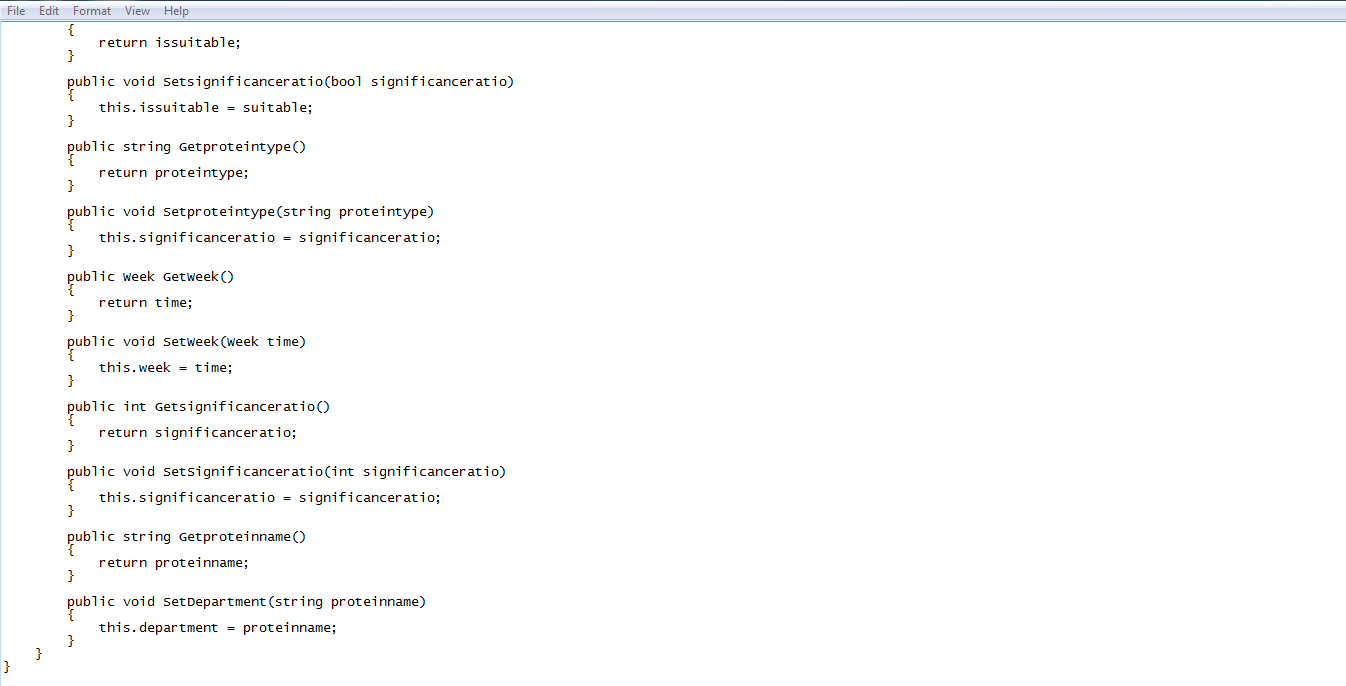
****

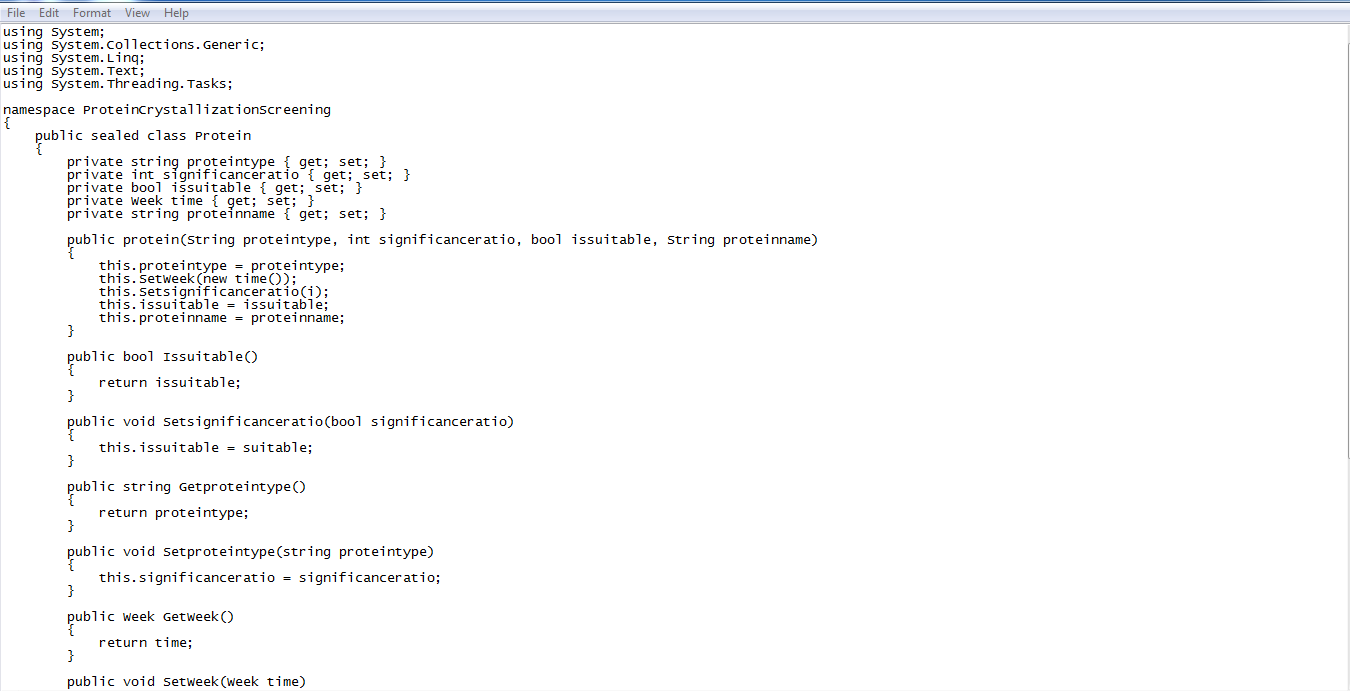
****

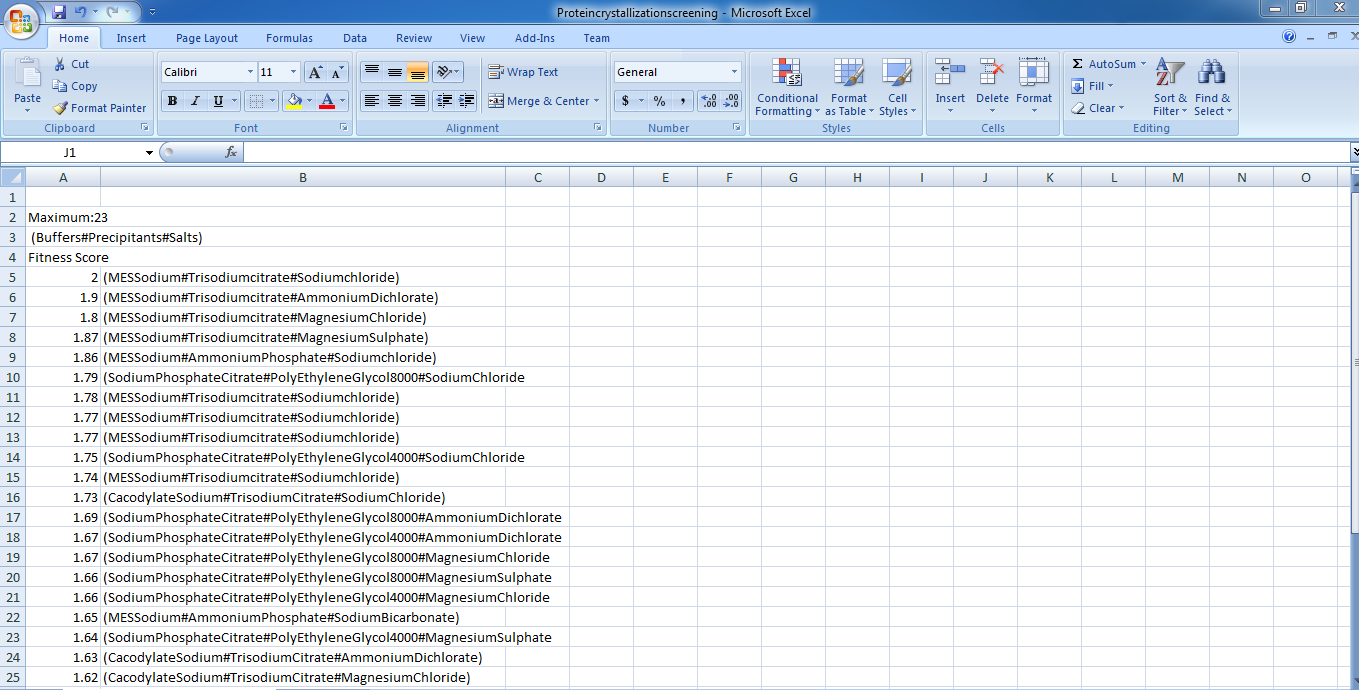
****  ** **

****

**** 





****