Using Set Covering Problem Varaints to Detect Genes Responsible for Diseases

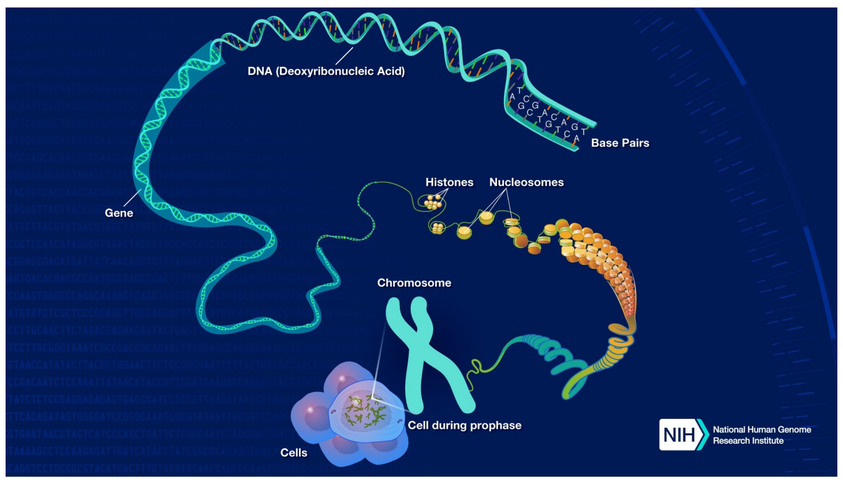
Author: Abiola Babatunde
Supervisors: AmirHosein Sadeghimanesh and Matthew England

Introduction

Our research: Using combinatorial optimisation methods to detect genes responsible for diseases.

Benefits:

- Saving costs of gene sequencing analysis.
- Finding the genes responsible for diseases can lead to improved preventive and curative measures.



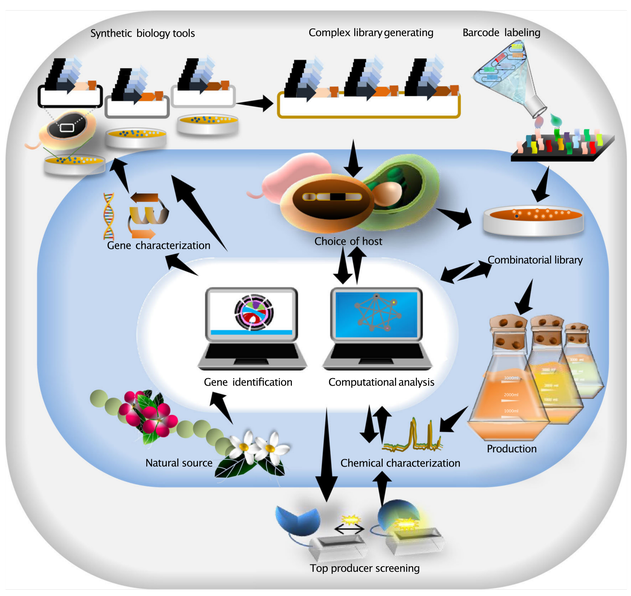
Gene Sequencing

What is gene sequencing?

The process of determininig the sequence of nucleotides in DNA is called gene sequencing. Gene sequencing is the first step in identifying the parts of the DNA responsible for the activities causing the disease.

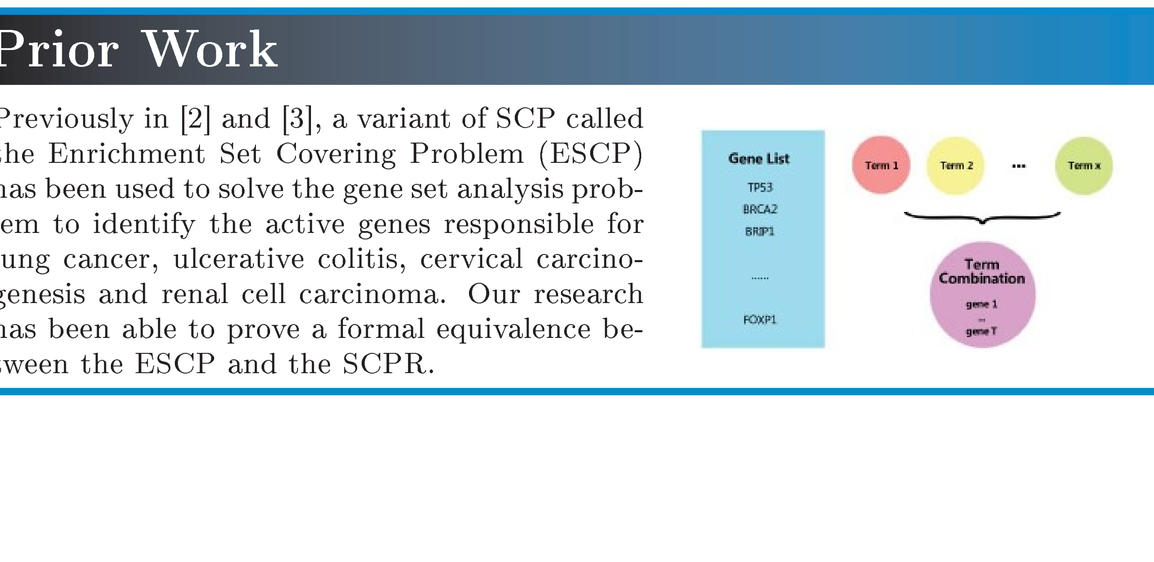
Why is gene sequencing related to combinatorial optimisation?

- The gene sequencing process is repeated multiple times and the data obtained is used to rule out the irrelevant parts.
- The process of analysing gene sequencing results can be written as a combinatorial optimisation problem.



Process of Gene Set Analysis

The figure shows the process of gene set analysis modeled as a combinatorial optimisation problem.



Prior Work

Previously in [2] and [3], a variant of SCP called the Enrichment Set Covering Problem (ESCP) has been used to solve the gene set analysis problem to identify the active genes responsible for lung cancer, ulcerative colitis, cervical carcinogenesis and renal cell carcinoma. Our research has been able to prove a formal equivalence between the ESCP and the SCPR.

Definition of SCPR

We can write the gene sequencing analysis as a combinatorial optimisation problem. Suppose we have a universal set, $U = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, $n \in \mathbb{N}$; a set of reasons, $R = \{\rho_1, \rho_2, \ldots, \rho_r\}$, $r \in \mathbb{N}$; and a set of pairs of subsets E,

$$E = \{(A_1, R_1), \ldots, (A_m, R_m)\}$$

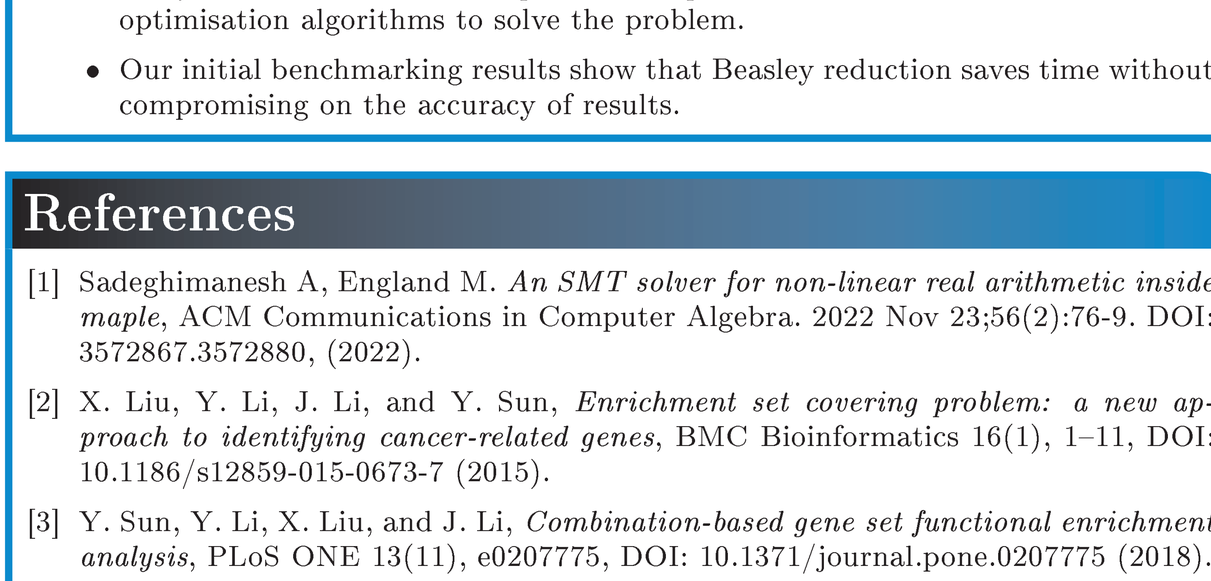
where $A_i \in \mathscr{P}(U)$ are called covering sets, and $R_j \in \mathscr{P}(R)$ are accompanying reason sets. So we have $E \in \mathscr{P}(U) \times \mathscr{P}(R)$.
Then we define the *Set Covering Problem with Reasons (SCPR)* as the problem of finding the minimal subset of reasons such that the covering sets A_i whose accompanying reason sets R_i are covered, themselves cover the entire universal set U. I.e. $S \subset R$ is the minimal subset of R such that

$$\bigcup_{i \text{ such that } R_i \subset S} A_i = U.$$

Relating this to gene sequencing analysis, U is the set of genes that are important and must be included, R is the set of the rest of the genes from the gene sequencing data, and E is built from biological processes. The goal is to find a set of biological processes that cover all of U but minimum number of genes from R using different optimisation algorithms.

Algorithms

Given the relatively recent introduction of SCPR, one of the aims of our research is to ascertain the most effective algorithms for these types of problems. Our research involves a comprehensive comparison of various exact and heuristic algorithms, including linear programming, SAT solvers, genetic algorithm, greedy algorithm, particle swarm optimisation (PSO), simulated annealing (SA), and the multi-level scoring element state configuration checking (MLSES-CC) algorithm. We also considered a method of reducing data complexity called the Beasley Reduction process. The Beasley Reduction process reduces data complexity by removing subsets that are dominated by other subsets. The results of this comparison will provide valuable insights into the most efficient and accurate algorithms for SCPR, thereby contributing to the development of more effective gene set analysis methodologies.



Conclusions

- Gene sequencing is important for understanding biological processes that happen in the cell.
- Understanding how gene sequencing analysis can be modelled as a combinatorial optimisation problem is important for the development of more effective gene set analysis methodologies.
- The introduction of the SCPR provides a new way to model gene sequencing analysis as a combinatorial optimisation problem and allows us to use different optimisation algorithms to solve the problem.
- Our initial benchmarking results show that Beasley reduction saves time without compromising on the accuracy of results.

References

[1] Sadeghimanesh A, England M. *An SMT solver for non-linear real arithmetic inside maple*, ACM Communications in Computer Algebra. 2022 Nov 23;56(2):76-9. DOI: 3572867.3572880, (2022).

[2] X. Liu, Y. Li, J. Li, and Y. Sun, *Enrichment set covering problem: a new approach to identifying cancer-related genes*, BMC Bioinformatics 16(1), 1–11, DOI: 10.1186/s12859-015-0673-7 (2015).

[3] Y. Sun, Y. Li, X. Liu, and J. Li, *Combination-based gene set functional enrichment analysis*, PLoS ONE 13(11), e0207775, DOI: 10.1371/journal.pone.0207775 (2018).

[4] J. E. Beasley, *An algorithm for the set covering problem*, European Journal of Operational Research 31(2), 180–189, DOI: 10.1016/0377-2217(87)90100-6 (1987).