MODIFIED HARRIS HAWKS OPTIMIZATION ALGORITHM FOR PROTEIN MULTIPLE SEQUENCE ALIGNMENT

AL-ZAIDI MOHAMMED KHALEEL IBRAHIM

UNIVERSITI SAINS MALAYSIA

2024

MODIFIED HARRIS HAWKS OPTIMIZATION ALGORITHM FOR PROTEIN MULTIPLE SEQUENCE ALIGNMENT

by

AL-ZAIDI MOHAMMED KHALEEL IBRAHIM

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

November 2024

ACKNOWLEDGEMENT

I wish to express my heartfelt gratitude to several individuals who have played pivotal roles in completing my Ph.D. thesis. First and foremost, my sincere thanks go to my dedicated and supportive supervisor, Dr. Umi Kalsom Yusof. Her unwavering guidance, valuable insights, and tireless encouragement throughout this journey have been invaluable. Dr. Umi's mentorship has significantly contributed to the quality and depth of this research; I am profoundly grateful for that. I would also like to extend my deep appreciation to my supervisor, Dr. Nibras Abdullah, for his insightful feedback, continuous support, and encouragement, which have been vital to the success of this work. I also want to extend my profound appreciation to my parents. Their unwavering love, unwavering support, and unending belief in my abilities have been my constant motivation. They have been my pillars of strength and have made countless sacrifices to see me succeed in my academic pursuits. This accomplishment is as much theirs as it is mine. Furthermore, I am grateful to my dear friends, whose encouragement, camaraderie, and support have made this challenging academic endeavor more enjoyable and manageable. Your moral support and shared experiences have been a source of inspiration throughout this PhD journey. I also extend my gratitude to our institution's academic and administrative staff for their support and resources that have facilitated my research. Lastly, my thanks go out to all those who, in one way or another, contributed to the completion of this thesis. Your assistance and encouragement have not gone unnoticed and are greatly appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony
ACO	Ant Colony Optimisation
BaliBase	Benchmark Protein Alignment Database
BLOSUM	Block Substitution Mutation
DNA	Deoxyribonucleic Acid
EHHO	Enhanced Harris Hawks Optimization algorithm
FHHOPCM	First-best-solution with Harris Hawks Optimization algorithm with Profile alignment, Crossover, and Mutation operator
GA	Genetic Algorithm
GA-ACO	Genetic Algorithm and Ant Colony Optimisation
GAPAM	Genetic Algorithm Progressive Alignment Approach
HHO	Harris Hawks Optimization algorithm
HHOP	Harris Hawks Optimization algorithm with Profile alignment
HHOPC	Harris Hawks Optimization algorithm with Profile alignment and Crossover operator
HHOPCM	Harris Hawks Optimization algorithm with Profile alignment, Crossover, and Mutation operator
LS	Local Search
MSA	Multiple Sequence Alignment
NP-hard	Non-deterministic Polynomial-time Hard
PAM	Point Accepted Mutation
PSO	Particle Swarm Optimization algorithm
Q	Alignment Quality
RNA	Ribonucleic Acid
SA	Simulated Annealing
SOP	Sum Of Pairs
TC	Total column Score
TS	Tabu Search

ALGORITMA PENGOPTIMUMAN HARRIS HAWKS YANG TERUBAHSUAI UNTUK PENJELASAN JURUTAN BERBILANG PROTEIN

ABSTRAK

Protein adalah penting untuk kehidupan, memberi kesan kepada banyak aspek kewujudan manusia. Kemajuan terkini dalam teknologi penjujukan generasi akan datang telah menjana sejumlah besar data dalam talian. Walau bagaimanapun, sarjana menghadapi cabaran untuk menavigasi maklumat ini dan menguruskan pengiraan kompleks yang diperlukan untuk membandingkan jujukan protein. Penjajaran jujukan adalah penting dalam konteks ini, dengan implikasi yang ketara untuk meningkatkan diagnosis penyakit awal dan kejuruteraan farmaseutikal. Penjajaran jujukan berbilang (MSA) ialah alat penting dalam bioinformatik untuk menganalisis jumlah data jujukan yang semakin meningkat. Walau bagaimanapun, mencari persamaan merentasi pangkalan data yang besar adalah masalah sukar NP, bermakna ia amat sukar dan memakan masa untuk diselesaikan dengan tepat dalam masa nyata. Ini menyerlahkan keperluan untuk kaedah yang lebih cepat dan lebih tepat. Paradigma metaheuristik baru-baru ini telah menarik perhatian penting di tengah-tengah pelbagai pendekatan untuk menangani masalah yang sangat kompleks seperti MSA. Peserta yang terkenal dalam domain ini ialah algoritma Harris Hawks Optimization (HHO), yang telah membezakan dirinya melalui hasil pengoptimuman yang diterbitkan, meletakkannya sebagai pesaing yang hebat dalam kalangan metaheuristik terkini. Akibatnya, algoritma HHO telah dipilih sebagai ubat yang berpotensi untuk menghadapi dilema ketepatan intrinsik kepada MSA. Penyesuaian algoritma HHO memerlukan pengubahsuaian bernuansa pengendalinya. Secara khusus, tiga lelaran berbeza

dicadangkan: Pertama, gabungan HHO dengan teknik Profil hibrid (HHOP) berusaha untuk mengoptimumkan skor Q dan TC dalam MSA, di mana skor Q dan TC mewakili ukuran kualiti penjajaran output. Kedua, mengkaji kepelbagaian HHO yang diubah suai dengan Profil melibatkan penyepaduan pengendali silang dan mutasi dalam populasi (HHOPCM). Ketiga, penambahbaikan dan penyiasatan lanjut dijalankan untuk meningkatkan keupayaan carian tempatan HHOPCM dan menghasilkan HHO (EHHO) yang dipertingkatkan. Selepas itu, hasil penyesuaian ini tertakluk kepada perbandingan yang ketat dengan banyak kaedah lain yang ditetapkan, memanfaatkan BAliBASE 3.0 sebagai set data penanda aras piawai. Berbanding dengan kaedah yang ditetapkan, kaedah EHHO mencapai hasil yang lebih baik untuk subset RV11 dan RV12 daripada set data Balibase 3.0 dari segi skor (Q dan TC) dengan (75.21, 53.97) untuk RV11 dan (94.95, 87.72) untuk RV12. Penilaian komprehensif ini bertujuan untuk memberikan pandangan dan kemajuan yang berharga dalam domain MSA.

MODIFIED HARRIS HAWKS OPTIMIZATION ALGORITHM FOR PROTEIN MULTIPLE SEQUENCE ALIGNMENT

ABSTRACT

Proteins are essential to life, impacting many aspects of human existence. Recent advances in next-generation sequencing technologies have generated a vast amount of data online. However, scholars face the challenge of navigating this information and managing the complex computations needed for comparing protein sequences. Sequence alignment is crucial in this context, with significant implications for improving early disease diagnosis and pharmaceutical engineering. Multiple sequence alignment (MSA) is a vital tool in bioinformatics for analyzing growing amounts of sequence data. However, finding similarities across large databases is an NP-hard problem, meaning it is extremely difficult and time-consuming to solve exactly in real time. This highlights the need for faster and more accurate methods. The metaheuristic paradigm has recently captured significant attention amid the diverse approaches to address highly complex problems like the MSA. A notable entrant in this domain is the Harris Hawks Optimization (HHO) algorithm, which has distinguished itself through published optimization outcomes, positioning it as a formidable competitor among state-of-the-art metaheuristics. Consequently, the HHO algorithm has been chosen as a potential remedy to confront the accuracy dilemmas intrinsic to MSA. The adaptation of the HHO algorithm entails a nuanced modification of its operators. Specifically, three distinct iterations are proposed: First, the fusion of HHO with a hybrid Profile technique (HHOP) seeks to optimize the Q and TC scores within MSA, where the Q and TC scores represent quality measures of the output alignment.

Second, examining the enhanced diversity of the modified HHO with Profile involves integrating crossover and mutation operators within the population (HHOPCM). Third, further enhancements and investigations are conducted to improve the local search capabilities of the HHOPCM and produce an enhanced HHO (EHHO). Subsequently, the outcomes of these adaptations are subjected to rigorous comparison with many other established methods, leveraging the BAliBASE 3.0 as a standardized benchmark dataset. Compared to the established methods, the EHHO method achieved a better result for the RV11 and RV12 subsets from the Balibase 3.0 dataset in terms of (Q and TC) scores with (75.21, 53.97) for RV11 and (94.95, 87.72) for RV12. This comprehensive evaluation aims to provide valuable insights and advancements in the domain of MSA.

CHAPTER 1

INTRODUCTION

1.1 Background

Bioinformatics refers to analyzing and managing biological information. Computational biology pertains to the application of computational methods, including physical and mathematical simulations, in the study and analysis of biological processes. (Diniz & Canduri, 2017; Tiwary, 2022; Varshney et al., 2022). In other words, Bioinformatics is where researchers create the tools, software, and algorithms that can be used to handle and work with sizeable biological data systems. Likewise, Computational Biology is all about learning and studying biology using the computational tools and software made by Bioinformaticians (Cohen, 2004; Shastry & Sanjay, 2020). Bioinformatics and computational biology (BCB) integrate computer science and molecular biology into a unified discipline, bridging hardware design and molecular biology knowledge.

The field of bioinformatics encompasses three major challenges. (Cohen, 2004; Shastry & Sanjay, 2020). The initial challenge in bioinformatics involves efficiently storing and organizing vast genomic databases. Subsequently, the field requires the development of resources and tools to facilitate the analysis of complex biological data. Finally, bioinformatics endeavors to leverage these tools to derive meaningful interpretations from biological information for a diverse array of crucial applications, including forensic medical sciences and drug discovery. For example, a comprehensive understanding of genetic and protein-related information within biological sequences can significantly

contribute to the development of improved medicines and treatments. This multidisciplinary field has gained considerable traction over the past decade, buoyed by advancements in computing technologies and the successful culmination of the Human Genome Project (HGP) in 2003, owing to its potential to enhance the quality of life.

The challenge facing bioinformatics has transitioned from the initial task of gathering genomic datasets and organizing them in computerized databases to the more nuanced endeavor of developing techniques for processing the accumulated biological data that encompasses DNA and protein sequences. Proteins, among the most abundant organic compounds in living organisms, exhibit diverse functions among all macromolecules. They can fulfill roles such as structural, regulatory, contractile, or protective functions and participate in transport, storage, or membrane-related activities. Some proteins may act as toxins or enzymes. Within a single cell of a living organism, there can be thousands of distinct proteins, each performing a specific function. These proteins exhibit a wide range of structures, reflecting their various functions. Proteins possess multiple structural levels, including primary, secondary, tertiary, and quaternary structures, each contributing to the elucidation of cellular functions and chemical properties. Proteins are essentially polymers of amino acids arranged in a linear sequence (Konieczny et al., 2023; David James Russell, 2014).

Sequence alignment, a cornerstone of bioinformatics, is a pivotal discipline in molecular biology that serves as a fundamental tool for analyzing biological information. It involves the systematic arrangement of nucleotides or amino acids, assuming they share common ancestors, to elucidate their structural and functional relationships. When the number of sequences for alignment equals two, it's termed pairwise alignment. Conversely,

if the number of sequences exceeds two, it's referred to as multiple sequence alignment (MSA) (Chowdhury & Garai, 2017; Y. Zhang et al., 2022).

There are five main approaches to the MSA problem (Amorim et al., 2021; Chatzou et al., 2016; Kapli et al., 2020): Progressive, Iterative, Consistency-based, Block-based, and metaheuristic approach. The primary focus in recent years has been on the metaheuristic approach since it has improved accuracy and speed compared to other MSA approaches by finding near-optimal solutions in a short time.

A metaheuristic is an algorithm crafted to tackle a broad spectrum of challenging optimization problems without necessitating deep adaptation to each specific problem. The Greek prefix "meta" in its name denotes that these algorithms are "higher-level" heuristics, contrasting with problem-specific heuristics. Metaheuristics are typically employed for problems lacking satisfactory problem-specific algorithms to address them. Most metaheuristics draw inspiration from nature, leveraging physics and biology principles and incorporating stochastic elements involving random variables. Additionally, they entail several parameters that require customization to suit the specific problem at hand. Broadly, metaheuristic methods can be categorized into single-solution and population-based approaches. In contrast, population-based metaheuristics are highly dependent on the quality of the initial population to find the solution in the search space (Abdollahzadeh et al., 2021; Boussaïd et al., 2013).

Within the realm of population-based metaheuristics, the Harris Hawks Optimization Algorithm (HHO) stands as a noteworthy exemplar. Drawing inspiration from the predatory behavior of Harris's hawks in nature, the HHO emulates the collaborative

hunting strategies observed in these avian species. In the optimization context, the HHO leverages the principles of prey localization and coordinated attack to explore solution spaces efficiently. By iteratively updating the position of search agents based on a combination of global and local information exchange, the HHO endeavors to converge toward optimal or near-optimal solutions. Notably, the efficacy of the HHO is contingent upon the quality of the initial population, emphasizing the importance of an adept initialization strategy in guiding the algorithm toward favorable outcomes within the solution landscape.

Profile alignment techniques are commonly utilized in collective endeavors such as remote homology detection and fold recognition (Du et al., 2021; Gribskov et al., 1987; G. Kumar et al., 2022; Söding, 2005; von Öhsen et al., 2002; Von Öhsen et al., 2004; Zheng et al., 2022). A substantial extent of work has been arranged in identifying Profile scoring functions that distinguish well between feebly homologous sequences and nonhomologous sequences (Edgar & Sjölander, 2004a; G. Kumar et al., 2022; von Ohsen & Zimmer, 2001). Although one might assume that a Profile scoring function that accomplishes nice classification should give precise multiple sequence alignments, experimental trials have shown only slight differences in alignment quality resulting from numerous Profile scoring schemes (Edgar & Sjölander, 2004b; Ohlson et al., 2004; Shinwari et al., 2022; Vicedomini et al., 2022). However, the combination of metaheuristics with the Profile alignment technique has shown promising results while solving nonhomologous multiple sequence alignment (Ali, 2016; Amorim et al., 2021; Hussein et al., 2019; Y. Zhang et al., 2022).

One of the most widely used stochastic algorithms to solve the MSA problem is the Genetic Algorithm (GA). It Is an evolutionary algorithm with underlying genetic operators that motivate many other algorithms (X.-S. Yang & He, 2019). One of the most powerful features of the GA is that fitness functions are separated from the GA parts, which gives great flexibility to the GA to solve a wide range of problems (Chowdhury & Garai, 2017; Tutumlu & Saraç, 2023). Many attempts have been recorded to solve the MSA problem by the use of GA and have shown competitive outcomes when compared to state-of-the-art sequence aligners, like SAGA (Chelly Dagdia et al., 2021; Notredame & Higgins, 1996), GAPAM (Dabba et al., 2020; Naznin et al., 2012a), and MO-SAStrE (Chowdhury & Garai, 2020; Ortuno et al., 2013).

While genetic Algorithms take the information from a few parents to produce an original solution, the algorithms built on swarm intelligence bring new individuals, taking into account information not only from their parents but also from the rest of the population. Also, hybrid metaheuristic approaches have proven to be highly competitive when compared to other GAs available in the literature for answering the MSA problem (Amorim et al., 2021; Rubio-Largo et al., 2016).

1.2 Problem Statement

Multiple Sequence Alignment (MSA) is a bioinformatics method for aligning and comparing several biological sequences, such as DNA, RNA, or protein sequences, to identify similarities, differences, and conserved regions. Deciphering the Multiple Sequence Alignment (MSA) challenge holds the potential for scientists to mine and recognize sequences within the human genome (Amorim et al., 2021). In this realm of

research, the assembly of sequences containing two or more residues, maximizing similarities among sequence alignments, assumes substantial importance. The primary aim of the Multiple Sequence Alignment problem is to enhance the accuracy of alignment methods for multiple sequences. This accuracy is critical because it aids in reconstructing phylogenetic trees and determining the function of previously unknown proteins by aligning their sequences with those of acknowledged proteins (Y. Zhang et al., 2022).

Through literature, many MSA methods have relied on tree-based approaches, constructing alignments based on pairwise comparisons and assumed evolutionary relationships between sequences (Amorim et al., 2021; Y. Zhang et al., 2022). While initially intuitive, these methods suffer from two critical limitations. Firstly, they rely on the existence of a perfect evolutionary tree, which is rarely the case. Errors in the tree structure translate directly into errors in the alignment, compromising accuracy. Secondly, the pairwise alignment steps can accumulate errors, further diminishing the overall quality of the final MSA (Zhongmin Wang et al., 2020). In bioinformatics, pursuing methods to elevate the overall accuracy of Multiple Sequence Alignment (MSA) has emerged as a progressively challenging endeavor (Y. Zhang et al., 2022). Conversely, the range of techniques currently available for addressing this problem is somewhat constrained regarding their accuracy (Amorim et al., 2021).

Numerous heuristic and metaheuristic approaches have been developed to achieve optimized solutions for complex problems (Chelly Dagdia et al., 2021; Y. Zhang et al., 2022), with notable success in both population-based and local search-based methods (Abdollahzadeh et al., 2021; Amorim et al., 2021; Velasco et al., 2024). The Harris Hawks Optimization (HHO) algorithm, a recent advancement in metaheuristics, has shown

promise in solving various optimization problems by mimicking the hunting behavior of Harris Hawks (Heidari et al., 2019). However, HHO has not yet been explored to tackle the Multiple Sequence Alignment (MSA) problem. Moreover, like many metaheuristics, HHO suffers from limitations such as random initialization, which can affect the algorithm's ability to find optimal solutions (Amorim et al., 2021; Shehab et al., 2022). Despite these challenges, ongoing research aims to improve HHO's performance by addressing these limitations (Agrawal et al., 2021).

Another area for improvement in the Harris Hawks Optimization (HHO) algorithm is its low diversity (Shehab et al., 2022). While the algorithm has effectively tackled optimization problems, its exploration phase can be hindered when dealing with large-scale datasets or complex optimization landscapes. The low diversity may hinder the algorithm's ability to reach an optimal solution efficiently. Addressing this limitation is vital for enhancing the algorithm's applicability in real-world optimization scenarios, like solving the MSA problem (Shehab et al., 2022).

An additional area for enhancing the Harris Hawks Optimization (HHO) algorithm lies in improving its exploitation phase (Shehab et al., 2022). The HHO algorithm may exhibit limitations in effectively exploiting promising solutions when dealing with large-scale datasets or complex optimization problems. This aspect of the algorithm impacts its ability to thoroughly exploit the search space and identify optimal solutions promptly (Velasco et al., 2024). By strengthening the exploitation component of the HHO algorithm, researchers aim to enhance its overall efficiency and effectiveness in solving optimization problems across various domains, like the MSA problem.

1.3 Research Questions

The central research inquiry revolves around enhancing Multiple Sequence Alignment's (MSA) accuracy. This research aims to delve into the following key questions:

1. How effective is the integration of the Harris Hawks Optimization (HHO) algorithm with the Profile technique in overcoming the limitations of HHO's random initialization of the initial solution and improving the accuracy of MSA solutions?

2. How significantly can the inclusion of crossover and mutation operators in the Harris Hawks Optimization (HHO) algorithm with Profile alignment enhance the algorithm's diversity and, consequently, the accuracy of MSA solutions in comparison to the previously discussed proposed algorithm?

3. How can the exploitation part of the algorithm from the previous point be enhanced by incorporating the elitism technique in the selection of potential solutions, thereby striving for improved accuracy in MSA solutions?

1.4 Research Objectives

The proposed methods have a primary objective: to empirically analyze the utilization of the HHO algorithm as a metaheuristic approach for resolving the MSA problem and improving accuracy. To fulfill this principal aim, the study sets forth the following objectives:

1. To enhance the capability of the Harris Hawks algorithm (HHO) by utilizing the Profile technique (HHOP) to improve the accuracy of the MSA solution.

2. To improve the exploration part of the HHOP algorithm by utilizing genetic operators of crossover and mutation (HHOPCM), seeking better MSA accuracy.

3. To optimize the exploitation part of the proposed HHOPCM algorithm by utilizing the elitism technique in selecting potential solutions (EHHO), hence aiming for better MSA solution accuracy.

1.5 Research Contributions

To the best of our information, the proposed methods represent the pioneering effort in adapting the Harris Hawks Algorithm (HHO) to address the Multiple Sequence Alignment (MSA) problem. Notably, the three proposed methods have been developed sequentially, each designed for the MSA problem.

The anticipated contributions of this study can be succinctly reviewed as follows:

1. By incorporating the Profile technique into the Harris Hawks Optimization (HHO) algorithm, this study addresses the limitation of random initialization of solution, significantly improving MSA accuracy. This novel HHOP approach demonstrates the effectiveness of hybrid techniques in bioinformatics.

2. Introducing crossover and mutation operators into the HHOP algorithm (HHOPCM) enhances exploration capabilities and increases solution diversity. This modification leads to substantial improvements in MSA accuracy compared to the original HHOP.

3. Incorporating the elitism technique in the HHOPCM algorithm (EHHO) improves the exploitation phase by selecting the best potential solutions. This ensures higher-quality solutions and further boosts MSA accuracy.

1.6 Scope and Limitations

This research evaluates and analyzes the results produced by aligning the protein sequences of the MSA problem using a newly introduced metaheuristic called Harris Hawks Optimization (HHO). The indication of HHO algorithm issues of processing MSA is the parameter tuning and limited ability of the exploitation part in the proposed algorithm. The use of genetic operators and profile alignment technique, together with the extra improvement in HHO exploitation, help improve the overall HHO algorithm precision.

The main limitation of this investigation is the run time resulting from the complex and exponentially growing nature of the MSA problem, besides the utility of the Profile alignment technique. Another limitation is the non-homogeneous outcomes of applying the proposed algorithm to different subsets from the protein datasets.

1.7 Research Organization

Chapter Two provides an extensive examination of MSA approaches and methodologies. It encompasses a comprehensive literature review that elucidates and evaluates the existing MSA techniques. Furthermore, it furnishes an overview of the Harris Hawks Optimization (HHO) algorithm.

Chapter Three is the research methodology. This chapter accommodates the general proposed framework and the experimental design approach.

Chapter Four embarks on an initial exploration of adapting the Harris Hawks Algorithm (HHO) for Multiple Sequence Alignment (MSA). This involves the introduction of a hybrid Profile technique designed to enhance the efficacy of the MSA-solving process. The Profile technique is incorporated into the HHO algorithm to address the limitation inherent in the initial alignment constructed using the HHO.

Chapter Five introduces a hybrid approach combining the Harris Hawks Algorithm (HHO) with genetic crossover and mutation operators to improve the MSA quality. Building on the successful method from Chapter 4, which uses the Profile technique to enhance the alignment of near-optimal solutions, this chapter integrates HHO with genetic operators to boost exploitation capabilities using crossover and mutation operators. Additionally, this chapter presents experimental tests and results for the proposed MSA methods, showcasing their performance.

Chapter Six presents enhancements to the Harris Hawks algorithm's local search, aiming to improve the Multiple Sequence Alignment quality by incorporating features from the top two solutions. This optimization facilitates effective search space exploration and efficient selection of the superior solution. The chapter concludes with a comparison of the suggested method's accuracy against common approaches.

Chapter Seven presents the concluding remarks of the research findings and suggestions and recommendations for future research endeavors.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides a comprehensive review of essential concepts and methods in genomics and sequence alignment. It begins with an overview of genomic data and databases, laying the groundwork for understanding how genetic information is stored and utilized. The discussion then moves to sequence alignment, a crucial technique for comparing DNA, RNA, or protein sequences. Key scoring schemes for alignment are examined in detail. The chapter further explores Multiple Sequence Alignment methods, from progressive and iterative techniques to block-based and metaheuristic approaches.

Additionally, the profile technique alignment is covered, illustrating its role in enhancing alignment outcomes. The chapter also includes a review of algorithms specific to multiple sequence alignment and a focused discussion on the Harris Hawks Optimization algorithm, showcasing its relevance in optimization tasks. Concluding with an analysis of research gaps, this chapter aims to provide a thorough understanding of current advancements and methodologies in the field.

2.2 Genomics

Genomics is an interdisciplinary field of biology concentrating on the structure, function, evolution, mapping, and editing of genomes (Ramsden, 2023). Cells within living organisms, encompassing humans, plants, and animals, predominantly consist of protein and nucleic acids, namely deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA, a pivotal molecule, furnishes the requisite instructions within the nucleus of every

cell, orchestrating biological operations comprising four nucleic acids: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA serves as the blueprint for genetic information storage. Genes, comprising segments of DNA, encode the instructions for cells within living organisms to synthesize other vital molecules termed proteins.

2.2.1 Genomic data

Within the human body, genes vary significantly in size, ranging from several hundred DNA residues to over 2 million residues or bases. Proteins, the building blocks of cells, are primarily composed of combinations of the 20 primary amino acids and are synthesized from DNA through two essential processes: transcription and translation. During transcription, the genetic information encoded in DNA is transcribed into a molecule known as messenger ribonucleic acid (mRNA). This mRNA molecule serves as a single-stranded copy of the gene and acts as a template for protein synthesis during the translation process. The translation represents the second step of gene expression, where the mRNA is decoded according to the rules of the genetic code, facilitating the conversion of the four-letter DNA code into the sequence of the 20 primary amino acids, as illustrated in Figure 2.1 (Weitzman & Weitzman, 2020). Table 2.1 summarizes the genetic code for amino acids, providing their three-letter code, the corresponding amino acid name, and its single-letter representation (Zeng et al., 2020).

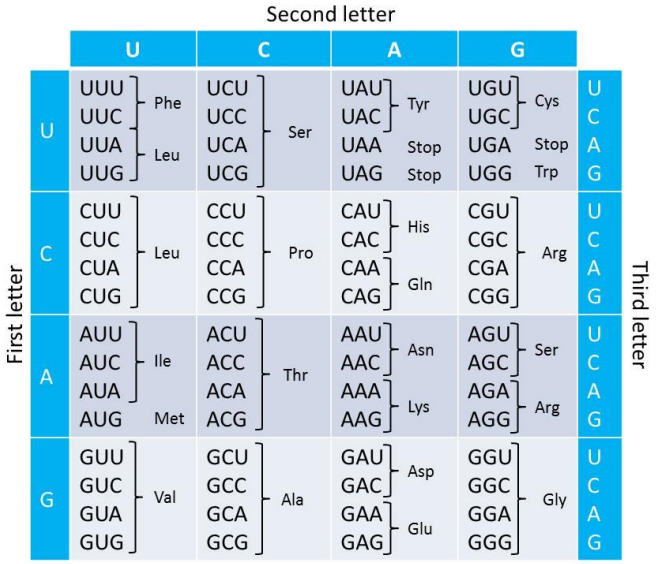

Figure 2.1 The genetic codes which relate the DNA to the amino acids

Table 2.1 The primary amino acids and their corresponding three-letter (genetic code) and single-letter codes

Amino acid	Three letter code	One letter code
alanine	ala	A
arginine	arg	R
asparagine	asn	N
aspartic acid	asp	D
asparagine or aspartic acid	asx	B
cysteine	cys	C
glutamic acid	glu	E
glutamine	gln	Q
glutamine or glutamic acid	glx	Z
glycine	gly	G
histidine	his	H
isoleucine	ile	I
leucine	leu	L
lysine	lys	K
methionine	met	M
phenylalanine	phe	F
proline	pro	P
serine	ser	S
threonine	thr	T
tryptophan	trp	W
tyrosine	tyr	Y

Various combinations of these 20 bases or residues give rise to diverse protein sequences, each possessing distinct biological functionalities. For example, within the

human body, blood comprises red blood cells responsible for oxygen transport. These cells utilize a protein known as hemoglobin to capture and convey oxygen throughout the body. Biological phenomena such as mutation and selection contribute to alterations in the genetic codes of both DNA and proteins. Consequently, these changes alter the characteristics and functions of cells in living organisms.

Bioinformatics employs specialized tools called sequence alignments to discern alterations in genetic codes within biological sequences resulting from various biological processes. The fundamental function of sequence alignment is to ascertain whether biological sequences exhibit biological relatedness or have arisen randomly (Tamposis et al., 2019).

In pairwise sequence alignment, a discovered biological sequence, referred to as the query sequence, undergoes comparison against each subject sequence within a database. Conversely, in multiple sequence alignment, the query sequence is simultaneously compared against multiple sequences. The underlying principle behind this foundational operation is primarily to identify regions of similarity among the sequences under investigation. Such regions of similarity may yield valuable insights into their functional, structural, evolutionary, or other intriguing characteristics, considering that the biological sequences have diverged from a common ancestral origin (Tamposis et al., 2019).

The mutational process encompasses various alterations within a sequence, including residue substitution, the addition of new residues, or the removal of existing residues. Specifically, the replacement of a residue in a sequence with a different one is termed substitution. Moreover, gaps refer to the residue's insertions or deletions in the

sequence. Gaps serve a crucial role in aligning sequences to accommodate underlying biological models. For instance, the DNA sequence (C-A-G-T) may be generated through the insertion of the residue (T) into the sequence (C-A-G), resulting in (C-A-G-T), or through the deletion of the residue (T) from the sequence (C-A-G-T-T).

The pursuit of sequence homology is a cornerstone tool in molecular biology, enabling the attainment of various objectives such as aiding in drug engineering, deducing protein functions from amino acid sequences, conducting genomic sequencing, and constructing evolutionary trees. Figure 2.2 (Behl et al., 2021) exemplifies the application of sequence alignment in facilitating drug discovery. In this scenario, the query sequence consists of a human amino acid sequence with a damaged protein segment resulting from DNA transcription errors (Behl et al., 2021).

Recent advancements in genomic sequencing technologies have revolutionized our understanding of genetic information and its applications. Innovations such as next-generation sequencing (NGS) have significantly increased sequencing speed and accuracy while reducing costs. Technologies like single-molecule sequencing and long-read sequencing are now capable of generating more detailed and comprehensive genomic data, allowing for better characterization of complex genomic regions and structural variants (Abdi et al., 2024). Additionally, improvements in bioinformatics tools and computational methods have enhanced the ability to analyze and interpret vast amounts of sequencing data, facilitating breakthroughs in personalized medicine, disease genomics, and evolutionary studies. These advancements are driving rapid progress in both research and clinical applications, transforming genomics into a more powerful and accessible field (K. R. Kumar et al., 2024).

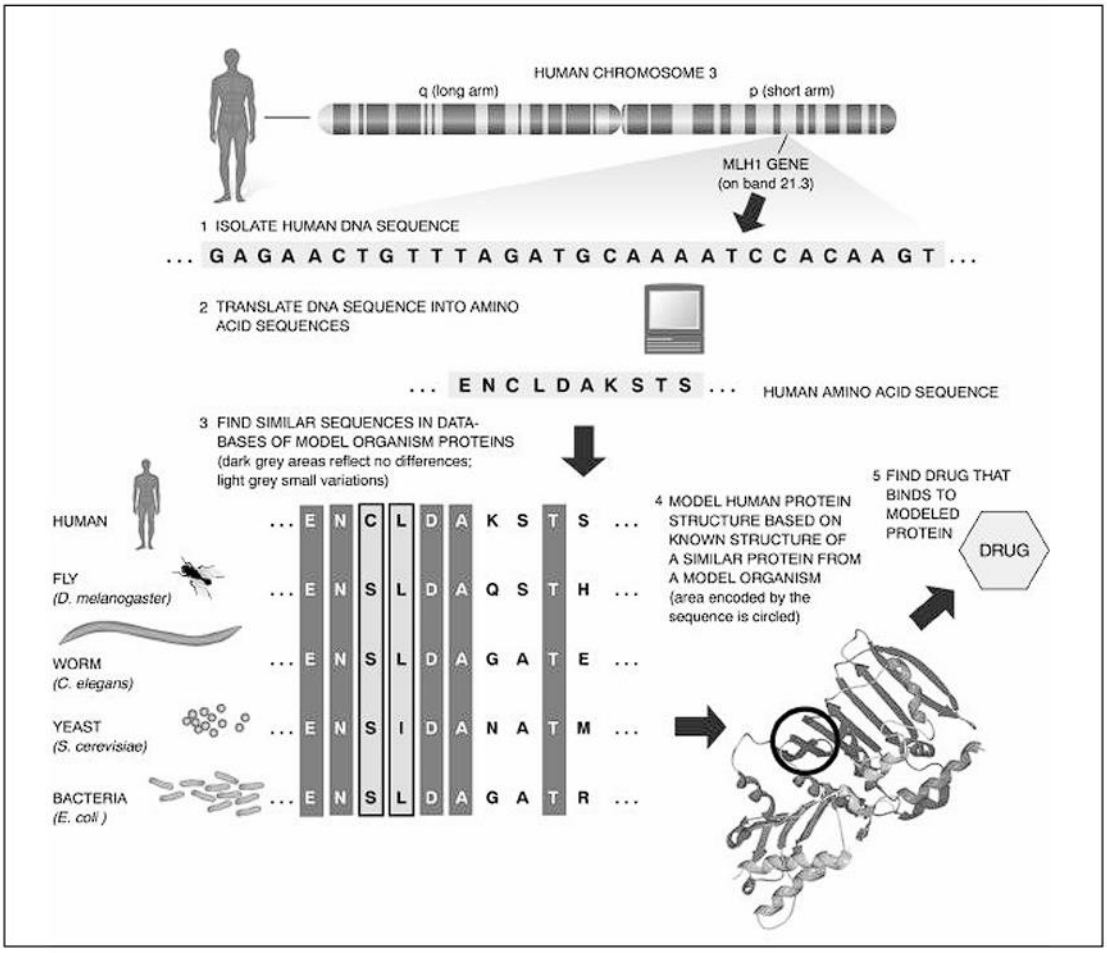

Figure 2.2 Example of the application of biological sequence alignment

Transcription, the process of transcribing genetic information from DNA to produce human protein, in this instance, is integral to the biological pathway. Within the context of pairwise sequence alignment, the query sequence is compared against other sequences (HUMAN, FLY, WORM, YEAST, and BACTERIA) stored in the database. Through the homology search, a segment of the HUMAN sequence within the database is identified as identical to the query sequence, as depicted in Figure 2.2 (Behl et al., 2021). Subsequently, leveraging the biological attributes of the best-matched sequence, it becomes possible to model the damaged structure of the HUMAN protein. This modeling process facilitates the

design of molecules tailored to act as drugs capable of binding to the damaged HUMAN protein structure.

2.2.2 Genomic database

The example of sequence alignment illustrated in Figure 2.2 only featured a limited number of subject sequences. However, it's worth noting that biological databases often contain an extensive array of sequences, numbering in the millions. Thanks to the completion of the Human Genome Project in 2003, vast amounts of biological information, including nucleotide and protein sequences, have been successfully stored, organized, and indexed within computerized databases. Each sequence in these databases is assigned a unique identifier known as an accession number. Notable examples of such computerized and publicly available biological databases include GenBank and the Universal Protein Resource (UniProt). GenBank, developed by the United States National Center for Biotechnology Information (NCBI) at the National Institute of Health (NIH), serves as a comprehensive genetic sequence database, functioning as an archive of primary sequence data. As of December 2019, GenBank contained an impressive 211,281,415 sequence entries. (Sayers et al., 2019). Additionally, GenBank collaborates with two daily data exchange partners as part of the International Nucleotide Sequence Database Collaboration (INSDC): the DNA Data Bank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL). GenBank serves as a repository for both protein and DNA data, while UniProt is primarily dedicated to providing protein data exclusively (Sorokina & Steinbeck, 2020).

The UniProt was originally established through the integration of three distinct protein database providers: the Swiss Institute of Bioinformatics and the European

Bioinformatics Institute (EBI), the Translated EMBL Nucleotide Sequence Data Library (TrEMBL) databases, and the Protein Information Resource - Protein Sequence Database (PIR-PSD) from Georgetown University (Sorokina & Steinbeck, 2020). Following their integration, UniProt and TrEMBL continue to exist as separate entities within the UniProt Knowledgebase (UniProtKB). UniProtKB consists of two main components: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot is a manually annotated protein database, whereas UniProtKB/TrEMBL comprises computationally analyzed sequence records sourced from the International Nucleotide Sequence Database Collaboration (INSDC). Figure 2.3 illustrates this distinction between UniProtKB/Swiss-Prot and UniProtKB/TrEMBL (de Haan et al., 2022). Figure 2.3 ("UniProt: the universal protein knowledgebase in 2023," 2023) illustrates the distribution of biological sequences by length (number of residues) within the UniProtKB/Swiss-Prot database as a representative example.

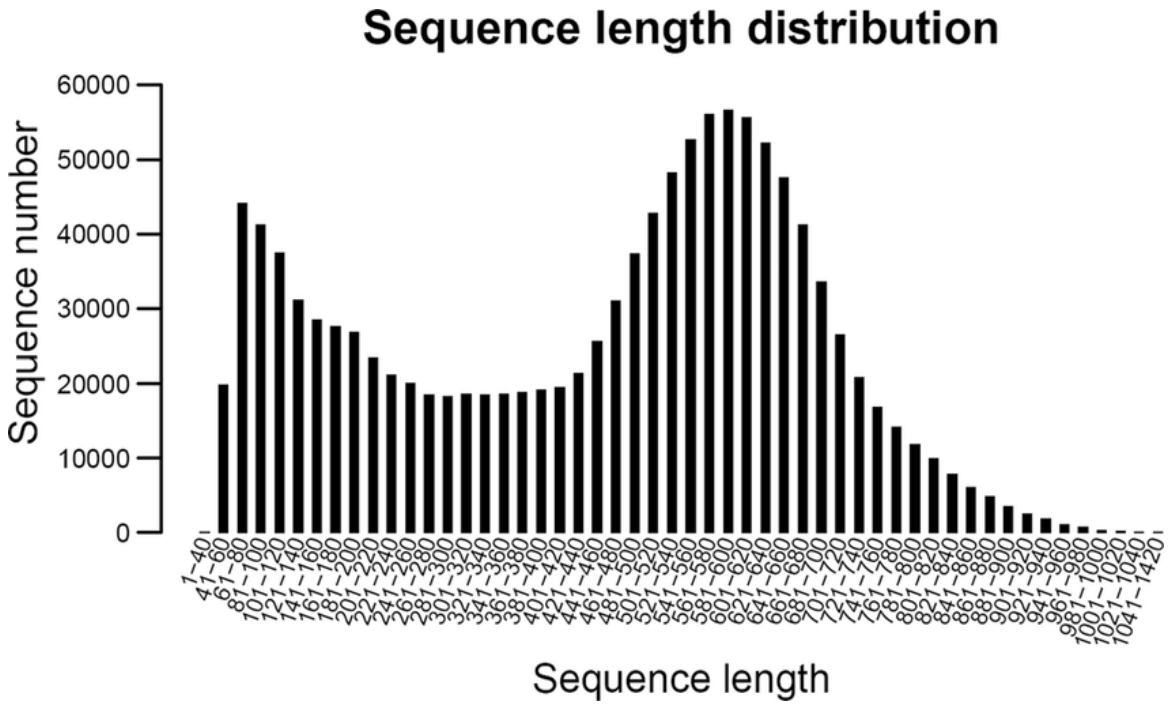

Figure 2.3 The UniProtKB/Swiss-Prot knowledgebase sequences

As of February 2022, the UniProtKB database (Release 2023-02) encompasses a total of 559,228 sequences, corresponding to 200,905,869 amino acids. The average sequence length within the database is 359 amino acids. Notably, the shortest sequence (GWA_SEPOF, accession number P83570) consists of only two amino acids, while the most extended sequence (TITIN_MOUSE, accession number A2ASS6) spans an impressive 35,213 amino acids (de Haan et al., 2022). From a hardware perspective, aligning such lengthy sequences using conventional computers would demand a considerable amount of time. Furthermore, the quantity of biological sequences stored in databases has been increasing exponentially over the years. Figure 2.4 serves as an illustrative example of this trend (Sorokina & Steinbeck, 2020). Figure 2.4 ("UniProt: the universal protein knowledgebase in 2023," 2023) depicts the exponential growth of the UniProtKB/TrEMBL database since the inception of the Human Genome Project (HGP). This trend underscores the necessity for high-performance computing platforms, including processors with multi-core architecture, and high-performance supercomputers, to expedite sequence homology searches and obtain results within a reasonable timeframe.

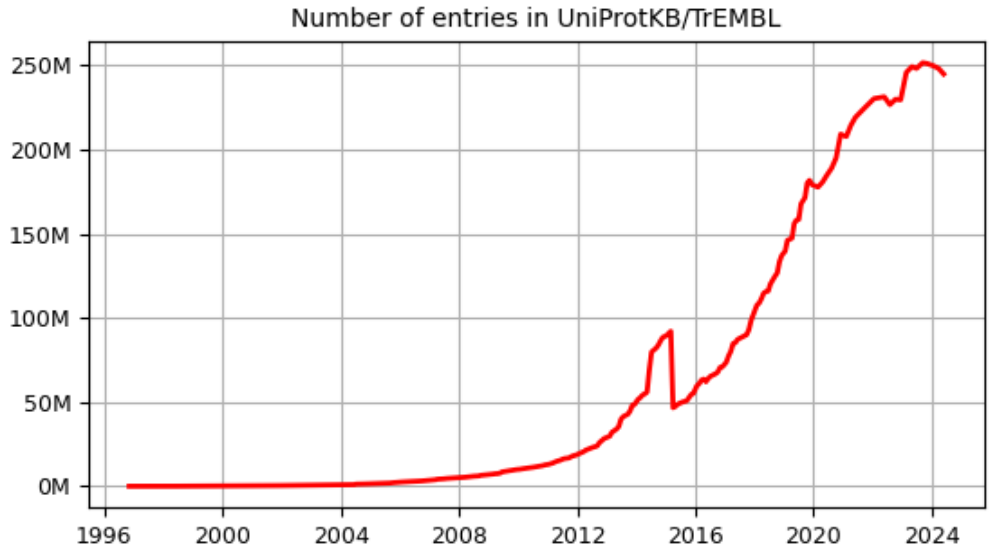

Figure 2.4 Exponential growth of the number of biological sequences in the UniProtKB/TrEMBL protein database over the years

To evaluate the efficiency of Multiple Sequence Alignment (MSA) programs, researchers often use benchmark datasets containing reference alignments. Among these, the BAliBASE dataset stands out as the most widely used for protein sequence alignment comparisons. The BAliBASE 3.0, proposed by Thompson and his team in 2005 (J. D. Thompson et al., 2005), is highly regarded in the research community for its comprehensive and well-curated reference alignments.

BAliBASE 3.0 includes 6255 sequences across 218 test cases and is divided into six reference subsets, each designed to test different aspects of alignment algorithms:

- RV11: Comprises 76 groups of equidistant sequences with less than 20% identities and fewer than 35 insertions, challenging algorithms with low sequence similarity.
- RV12: Contains 88 groups of sequences from families not included in RV11, with identities between 20% and 40%, testing the alignment of moderately similar sequences.

- RV20: This group consists of 82 groups with more than 40% identities, including diverse sequences, and assesses the ability to handle high similarity with significant variations.
- RV30: Includes 60 groups of sequences from various subfamilies with more than 40% identities within subfamilies but less than 25% between them, focusing on subfamily-specific alignment.
- RV40: Comprises 49 groups with more than 20% identities and substantial terminal insertions, evaluating the handling of large terminal gaps.
- RV50: Contains 31 groups with more than 20% identities and extensive internal insertions, testing the management of large internal gaps.

The importance of the BAliBASE dataset lies in its rigorous structure, which provides a standardized and challenging platform for evaluating and comparing the performance of MSA programs. Researchers widely use BAliBASE due to its detailed categorization and comprehensive test cases, making it an invaluable tool for advancing the field of protein sequence alignment (Amorim, 2021; Paruchuri, 2022; Zhang, 2022).

2.3 Sequence Alignment

In bioinformatics, sequence alignment is a way of positioning the sequences of DNA, RNA, or protein to recognize regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences (Chao et al., 2022). The goal of sequence alignment is to locate an optimum match between the sequences being examined. Sequence alignment is a case of text alteration because the sequences are characterized as text strings over a given alphabet. For example, a DNA

sequence draws from an alphabet of four characters (A, C, G, and T) demonstrating the four nucleotides, while a protein sequence draws from an alphabet of 20 symbols, each symbolizing an amino acid (Kille et al., 2022).

The objective of alignment is to organize multiple strings in a manner that they are vertically aligned optimally, emphasizing both similarities and differences (Gururaj & Siddesh, 2020). Empty spaces, known as gaps, are strategically inserted into the strings to ensure all sequences are extended to the same length. This alignment allows symbols in each string to match vertically with corresponding symbols in other strings as frequently as possible. The optimal alignment is achieved when the highest number of symbols from one sequence align with those of another sequence, considering all potential gap placements (Askr et al., 2023).

Pairwise alignment, as the name suggests, involves aligning two sequences, while Multiple Sequence Alignment (MSA) deals with aligning three or more sequences. MSA is particularly valuable for highlighting similarities among sequences, aiding in the detection of distantly related proteins. It is also employed in predicting protein secondary structures, identifying patterns of mutational change, and assessing the evolutionary relationships among family members, often leading to the construction of evolutionary trees (Sievers & Higgins, 2021).

Alignment can be classified as either global or local. Global alignment aims to find the best alignment for the entire sequence, while local alignment focuses on identifying and aligning similarities in smaller, specific regions of the sequences (Bucak & Uslan, 2011; Sievers & Higgins, 2021). The Smith-Waterman algorithm (Smith & Waterman, 1981) is a

central example of a best local alignment algorithm, while the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) is the most popular global alignment algorithm.

Regardless of the approach chosen, a similarity measure is essential to assess how well the sequences align in a given alignment. The similarity scores obtained from various candidate alignments are compared to identify the optimal alignment. This similarity measure must consider changes in the sequences resulting from insertion, deletion, or mutation during evolutionary processes. One common strategy involves strategically inserting additional gaps within the sequences. Assuming that inserting a gap results in more symbol matches and, thus, a higher score, the similarity measure needs to incorporate a penalty for gap insertions to discourage an excessive number of gaps. Typically, two penalty values are used: one for introducing a new gap into a sequence and another for extending the length of an existing gap (Das et al., 2008; Sievers & Higgins, 2021).

In addition to discouraging excessive gap insertions, a good similarity metric should also promote logical and realistic symbol pairings. To achieve this, most metrics rely on substitution matrices, typically from families like PAM or BLOSUM, such as PAM250 or BLOSUM62. These matrices assign scores to each possible amino acid substitution, with higher values assigned to symbol mutations that are more likely to occur naturally. As a result, an alignment that vertically aligns matching symbols will receive a higher score. However, mismatches that are common or biologically explainable will score better than those that are extremely rare or illogical (Chao et al., 2022; Rodriguez et al., 2007).