A Probabilistic Method to Find and Visualize Distinct Regions in Protein Sequences

Morteza Hosseini
IEETA/DETI
University of Aveiro
Aveiro, Portugal
seyedmorteza@ua.pt

Diogo Pratas
IEETA/DETI
University of Aveiro
Aveiro, Portugal
pratas@ua.pt

Armando J. Pinho
IEETA/DETI
University of Aveiro
Aveiro, Portugal
ap@ua.pt

***Abstract*—Studies on identification of species-specific protein regions, i.e., unique or highly dissimilar regions with respect to close species, will lead us to understanding of evolutionary traits, which can be related to novel functionalities or diseases. In this paper, we propose an alignment-free method to find and visualize distinct regions between two collections of proteins. We applied the proposed method, FRUIT, on multiple synthetic and real datasets to analyze its behavior when different rates of substitutional mutation occur. Testing with different k-mer sizes showed that the higher the mutation rate, the higher the relative uniqueness. We also employed FRUIT to find and visualize distinct regions in modern human proteins relatively to the proteins of Altai, Sidron and Vindija Neanderthals. The results show that four of the most distinct proteins, named ataxin-8, 60S ribosomal protein L26, NADH-ubiquinone oxidoreductase chain 3 and cytochrome c oxidase subunit 2 are involved in SCA8, DBA11, LS and MT-C1D, and MT-C4D diseases, respectively. There is also Interferon-induced transmembrane protein 3, among others, which is part of the immune system. Besides, we report the most similar primate exomes to the found modern human one, in terms of identity, query cover and length of sequences. The reported results can give us insight to the evolution of proteomes.**

***Index Terms*—palaeoproteomics, Neanderthals, alignment-free method, relative uniqueness, Bloom filter**

I. Introduction

Palaeoproteomics is an emerging field that focuses on the study of ancient proteomes and intersects evolutionary biology, archaeology and anthropology. It has the potential to provide researchers with the information about new or existing phylogenetic trees, species identification and past migrations [1]–[5].

Proteins can last longer than DNA, since they have more stable bonds for connecting them, are deposited in greater volumes and have more degradation-proof molecular structures. This makes them appropriate for recovering information from much longer periods back in time [6]. Even with the best preservation conditions, the oldest DNA samples date back to 0.4–1.5 million years ago, while the oldest proteins are hundreds of millions years old [7], [8].

One of the closest hominins to modern humans are Neanderthals, who lived within Eurasia from 400,000 until 40,000 years ago [9], [10]. Their sequences has been provided in the literature, as pieces [11]–[13], complete mitochondrial [14], genome draft [15], complete genome [16], [17] and complete exome [18]. In this paper, we use the complete exomes of a ∼50,000 year old Neanderthal from Denisova Cave in the Altai Mountains in Siberia, a ∼49,000 year old one from El Sidron Cave in Spain and a ∼44,000 year old one from Vindija Cave in Croatia [18].

In [19], it has been suggested that modern human interbred with Neanderthals when they arrived to Europe. Considering this plus the similarity reported between these two hominins [20], [21] would mean that unique regions in the proteome of modern humans may be of limited extent. In this paper, we use Altai, Sidron and Vindija Neanderthals proteins to find and visualize distinct regions of the reference proteome of modern human. This has been studied at the genomic level [22]; here, we study it at the proteomic level.

In the following section, we propose an alignment-free probabilistic method to find unique regions in collections of proteins and describe it in detail. Then, we present the results of running the implemented tool, FRUIT, on a collection of Neanderthal and modern human proteins as well as synthetic sequences. We also analyze the effect of mutations in the mentioned datasets. Finally, we draw some conclusions.

II. Method

We tackle the problem of finding the regions in target sequences which do not exist in reference sequences. For this purpose, a model is required that can search the references for the existence of all words of a certain size, k, in the targets. It should also be able to report the positions of unique words. The model that we use is described next.

A. Model

For the purpose of checking the existence of all k-mers of a target in a reference sequence, it is impractical to use a binary vector which considers all the possible situations. We give an example; assume the cardinality of the alphabet representing a protein sequence is 20, and the k-mer size is 14. This needs a gigantic amount of $20^{14} \simeq 1.5$ million terabytes of memory. To tackle this problem, we use a space-efficient probabilistic data structure, named Bloom filter [23], [24]. In this data structure, false negative matches are impossible but false positives are not, i.e., it enables to test whether the k-mers of a target sequence are definitely not members of a reference, or they possibly are. A Bloom filter, with optimal number of

978-9-0827-9703-9/19/$31.00

hash functions and an appropriate size, can provide the results only slightly different than the deterministic approach.

An empty Bloom filter is a bit vector of size m. This model requires h different hash functions which map, separately, each k-mer to one of the bit vector positions; this will generate a uniform random distribution. The number of hash functions, h, which minimizes the false positive probability, p, is proportional to the number of amino acids in a reference sequence, n, and is obtained by

$$h = \frac{m}{n} \ln 2. \tag{1}$$

It has been proven, in [25], that considering the optimal value of h, the false positive probability is at most

$$\left(1 - e^{-\frac{h(n+0.5)}{m-1}}\right)^h. \tag{2}$$

To hash k-mers, we use universal hashing. Assume our intention is to map k-mers from some universe U to m bins, labeled as $[m] = \{0, 1, \ldots, m-1\}$. We need to randomly select a function from a family of hash functions. A family of functions $F = \{f : U \to [m]\}$ is called a universal family if,

$$\forall x, y \in U,\ x \neq y: \quad \Pr_{f \in F}[f(x) = f(y)] \leq \frac{1}{m}. \tag{3}$$

$1/m$ is the probability of collision when a hash function maps a key to a truly random element. To obtain a universal hash function, we use the state-of-the-art multiply-add-shift scheme [26]:

$$f_{a,b}(x) = ((ax + b) \bmod 2^w) \operatorname{div} 2^{w-M}, \tag{4}$$

in which, w is the number of bits in a machine word, e.g., 64, M is $\log_2 m$, assuming the number of bins, m, is a power of two, a is a random positive integer less than 2^w and b is a random non-negative integer less than 2^{w-M}.

The algorithm of the proposed method is shown in Fig. 1.

B. Implementation

The proposed method has been implemented in C++ language, under the name of FRUIT, and the executables are publicly available at [27], under GPLv3. The implemented tool contains three programs: `fruit-map` (to map the relatively unique regions), `fruit-filter` (to filter the regions and save the positions) and `fuit-visual` (to visualize positions of the relatively unique regions). The three file formats of FASTA, FASTQ and SEQ (including solely amino acid letter codes, e.g., M, A, R, D, etc.) can be fed to this tool.

III. Results

We tested the proposed tool on a machine which had an 4-core 3.40 GHz Intel® Core™ i7-6700 CPU with 32 GB RAM. The results presented in this paper can be replicated by the script `run.sh`, available at [27].

To analyze the behavior of FRUIT for different rates of substitutional mutations, we applied it to real and synthetic datasets, shown in Table I, and measured uniqueness ratios. These datasets are available at [27], however, they can be downloaded from [28] and [29], too.

```
Initialize the size of Bloom filter, m, and the k-mer size, k
for each r in reference proteins do
    Calculate the optimal no. of hash functions, h, using (1)
    Calculate the false positive probability, p, using (2)
    for each k-mer string, s, in r do
        for each h_i in hash function family do
            Hash s to position o_{s,i} in the Bloom filter, using (4)
            Update the Bloom filter on the position o_{s,i}
        end for
    end for
    for each t in target proteins do
        for each k-mer string, s, in t do
            for each h_i in hash function family do
                Hash s to position o_{s,i} in the Bloom filter
                Query the Bloom filter for the position o_{s,i} and
                save the Boolean result to the unique file u_{r_i,t_j}
            end for
        end for
    end for
end for
for each t in target proteins do
    Perform bit-wise or on all Boolean elements of unique
    files u_{r_i,t} and save the result to the unique file u_t
    Filter u_t by a window of size w and save the relatively
    unique positions in the file o_t
end for
Illustrate the positions, saved in o_t files, in an SVG image
```

Fig. 1. The algorithm of the proposed method.

The uniqueness ratio falls within the range $[0.0, 1.0]$, and is obtained by size of the relatively unique region divided by size of the sequence. It shows the portion of a target file which does not exist in the reference, with respect to the k-mer size.

Fig. 2 demonstrates uniqueness ratios versus mutation rates, for k-mer sizes of 5 to 10, for a synthetic dataset plus three samples of Altai, Sidron and Vindija Neanderthals. To mutate the datasets, we use the `mutate` tool [27]. Note that when $k = 1, 2, 3, 4$, the uniqueness ratio is 0 for all mutation rates, i.e., all words of size up to 4 in the targets are found in the references. Fig. 2 shows that the higher the mutation rate, the higher the uniqueness ratio, and also, the greater the k, the greater is the uniqueness ratio. As an example, with 50% mutation, given a uniform distribution, we expect a target to be highly dissimilar to the reference. This can be seen in all of the datasets, for $k \geq 7$.

For the next experiment, we picked as targets all the 20,412 proteins of modern human [29], and found their unique regions relatively to the Altai, Sidron and Vindija Neanderthals [28], as references. For this purpose, we first used `fruit-map` to map amino acids of the targets to the files showing their existence in the references. In this phase, we considered as data structure a highly accurate Bloom filter with the false positive probability of 0.00001. Then, we used `fruit-filter` to filter the results of `fruit-map` and find the positions of

TABLE I
DATASETS USED IN THIS PAPER, INCLUDING SYNTHETIC AND REAL DATA FROM NEANDERTHALS AND MODERN HUMAN.

Reference sequences					**Target sequences**				
Dataset	***Species***	***# amino acids***	***# reads***	***Cardin.***[a]	***Dataset***	***Species***	***# amino acids***	***# reads***	***Cardin.***[a]
Synthetic	–	5,000,000	-	20	Syn. mutated[b]	–	5,000,000	–	20
Altai	*H. neanderthalensis*	22,829,171	42,394	21	A. mutated[b]	–	22,829,171	42,394	21
Sidron		22,829,205	42,394		S. mutated[b]		22,829,205	42,394	
Vindija		22,829,173	42,394		V. mutated[b]		22,829,173	42,394	
Altai	*H. neanderthalensis*	22,829,171	42,394	21	Modern human[c,d]	*H. sapiens*	11,374,527	20,412	21
Sidron		22,829,205	42,394						
Vindija		22,829,173	42,394						

[a]Cardinality: number of different amino acids in a protein sequence.
[b]This dataset is a copy of the reference file which is mutated from 1% to 50%, using `mutate` tool [27]. It contains 50 different files with the same number of amino acids and number of reads.
[c]Reviewed reference proteome–manually annotated.
[d]The modern human multi-FASTA file is divided into 20,412 FASTA files. Each one of them is considered as a target, and the three samples of Altai, Sidron and Vindija Neanderthals are considered as references altogether.

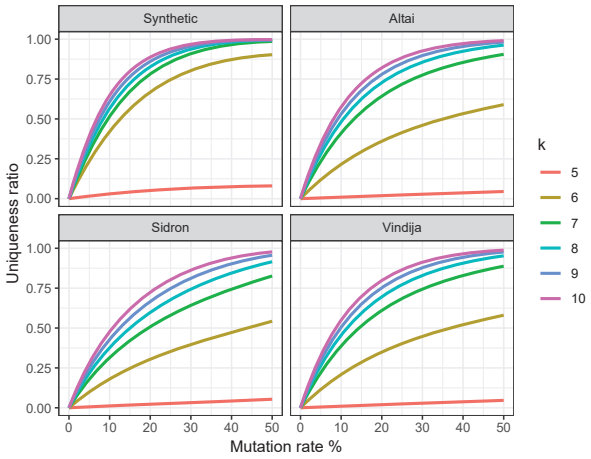

Fig. 2. Uniqueness ratios for different rates of mutation and different k-mer sizes applied on synthetic and real (Neanderthals) datasets.

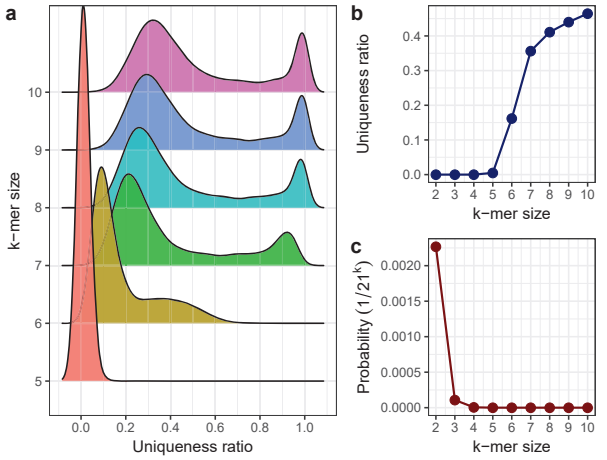

Fig. 3. a) Distribution of uniqueness ratios, b) total uniqueness ratios and c) probability of a target word being seen in the reference, for different k-mers.

relatively unique regions. Finally, we used `fruit-visual` to visualize the relative positions found in the previous step.

Fig. 3a shows distributions of uniqueness ratios for modern human sequences relatively to Neanderthals sequences, for different k-mer sizes. As can be seen, the shapes of distributions changes for $k = 5, 6, 7$, but thereafter, it changes only slightly with k. Fig. 3b demonstrates the total uniqueness ratios for different k values. For $k = 7$, the sign of the second derivative changes from positive to negative, meaning that it is the lowest upper-bound that can be chosen for our purpose. Fig. 3c shows the probability of a target word being seen in the reference, considering different k-mer sizes. As shown, its value is $\sim$0 for $k \geq 4$. The probabilities of the k-mers are considered to be $1/21^k$, in which 21 is the maximum cardinality of alphabet representing the target and references.

The top ten unique modern human proteins relatively to the Neanderthals proteins are described in Table II, in detail, and illustrated in Fig. 4, using `fruit-visual`. Each color in the figure represents a continuous relatively unique region. As an example, 2 out of 145 amino acids in 60S ribosomal protein L26 are not relatively unique, leading to the separation of the protein into two regions, each one represented by a distinct color. For a more complete list of detected proteins, see [27].

Table III describes the most similar exomes of non-human primates to the ones listed in Table II. This table shows that the found modern human exomes exist, fully or partially, in the listed primates, but not in the Neanderthals. This can have multiple reasons, such as ambiguity of computational models in prediction of proteins, inclusion of contaminant exogenous sources [30] and ancient DNA damage [31].

IV. CONCLUSIONS

We proposed a probabilistic method, FRUIT, to find dissimilarity between two collections of proteins. Testing on synthetic and real (Neanderthals) datasets, the impact of different rates of substitutional mutation on uniqueness ratios were analyzed. The FRUIT tool was also employed to map, filter and visualize unique proteins in modern humans relatively to Altai, Sidron and Vindija Neanderthals. The top ten distinct proteins are reported in this paper. Some of the found proteins are associated

TABLE II
THE MOST UNIQUE PROTEINS OF MODERN HUMAN ABSENT IN THE NEANDERTHALS.

Accession number[a]	Modern human protein	Gene	Len.[b]	Unique. ratio	Note
Q3SY05	putative uncharacterized protein encoded by LINC00303	*LINC00303*	128	0.9922	product of a dubious CDS prediction[c]
Q5QFB9	protein PAPPAS	*PAPPA-AS1*	102	0.9902	product of a dubious CDS prediction[c]
Q156A1	ataxin-8	*ATXN8*	80	0.9875	involved in spinocerebellar ataxia 8 (SCA8) disease[d]. It is unknown whether this protein exists in non-SCA8 individuals
Q5VT33	putative uncharacterized protein encoded by LINC01545	*LINC01545*	79	0.9873	protein predicted
P61254	60S ribosomal protein L26	*RPL26*	145	0.9862	component of the large ribosomal subunit. It is involved in Diamond-Blackfan anemia 11 (DBA11) disease[d]
A8MTZ7	uncharacterized protein C12orf71	*C12orf71*	265	0.9851	protein predicted
Q01628	interferon-induced transmembrane protein 3	*IFITM3*	133	0.9850	IFN-induced antiviral protein which disrupts intracellular cholesterol homeostasis
H3BRN8	uncharacterized protein C15orf65	*C15orf65*	121	0.9835	experimental evidence at transcript level
P03897	NADH-ubiquinone oxidoreductase chain 3	*MT-ND3*	115	0.9826	core subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I) that is believed to belong to the minimal assembly required for catalysis. It is involved in Leigh syndrome (LS) and mitochondrial complex I deficiency (MT-C1D) diseases[d]
P00403	cytochrome c oxidase subunit 2	*MT-CO2*	227	0.9824	cytochrome c oxidase is the component of the respiratory chain that catalyzes the reduction of oxygen to water. Subunits 1-3 form the functional core of the enzyme complex. Subunit 2 transfers the electrons from cytochrome c via its binuclear copper A center to the bimetallic center of the catalytic subunit 1. It is involved in mitochondrial complex IV deficiency (MT-C4D) disease[d]

[a]A unique identifier of an entry in the UniProtKB database [29].
[b]Number of amino acids in the sequences.
[c]Level of evidence is "uncertain"; therefore, it is either a) derived from the erroneous translation of a pseudogene or non-coding RNA, that should be removed from protein database, in case the evidence of pseudogenization is overwhelming for instance, or b) it should be upgraded to the certain level, which has happened to e.g., E.coli pseudogene ymiA that has now been found to produce a protein product.
[d]This disease is caused by mutations affecting the gene represented in this entry.

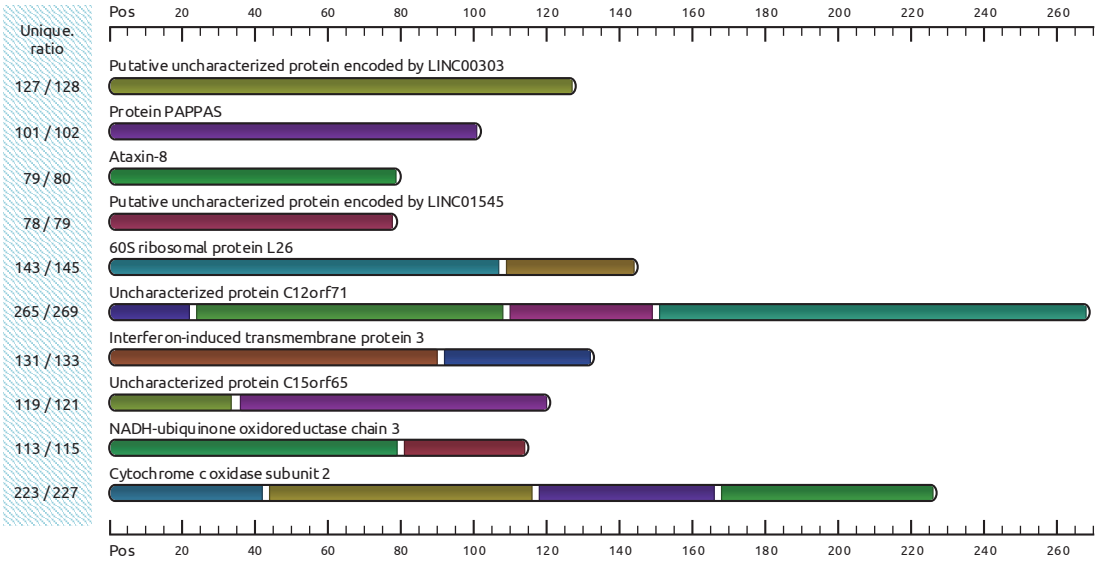

Fig. 4. Modern human proteins with the most distinct regions against Altai, Sidron and Vindija Neanderthals. The format m/n shows that considering k-mer size of 7, m out of n amino acids are relatively unique.

with diseases, including SCA8, DBA11, LS, MT-C1D and MT-C4D, and some others are putative uncharacterized proteins, which are products of dubious CDS prediction. Several other proteins are detected and sorted based on their uniqueness ratios and are available at [27]. Furthermore, we have listed and described the most similar primate exomes to the found modern human ones. Future studies can be done based on the reported proteins to find their expression and meaning in the evolution path. They might reveal unique functionalities in the Neanderthals or modern humans.

TABLE III
THE MOST SIMILAR NON-HUMAN PRIMATE EXOMES TO THE MODERN HUMAN, THAT ARE LISTED IN TABLE II.

Modern human protein	Similar primate protein	Accession number[a]	Species	Len.	Identity[b]	Query cover[c]
Putative uncharacterized protein encoded by LINC00303	predicted putative uncharacterized protein encoded by LINC00303	XP_004088138.1	*N. leucogenys*	128	92.19%	100%
Protein PAPPAS	predicted protein PAPPAS	XP_015292165.1	*M. fascicularis*	105	94.06%	99%
Ataxin-8	—[d]					
Putative uncharacterized protein encoded by LINC01545	putative uncharacterized protein encoded by LINC01545	XP_008960198.1	*P. paniscus*	79	100.00%	100%
60S ribosomal protein L26	60S ribosomal protein L26 isoform X1	XP_008059327.2	*C. syrichta*	149	100.00%	100%
Uncharacterized protein C12orf71	uncharacterized protein C12orf71 homolog	XP_520810.1	*P. troglodytes*	269	99.26%	100%
Interferon-induced transmembrane protein 3	predicted interferon-induced transmembrane protein 3 isoform X2	XP_004050385.1	*G. gorilla gorilla*	133	100.00%	100%
Uncharacterized protein C15orf65	uncharacterized protein C15orf65 homolog isoform X2	XP_003314728.1	*P. troglodytes*	121	99.17%	100%
NADH-ubiquinone oxidoreductase chain 3	NADH dehydrogenase subunit 3	ABU47841.1	*P. troglodytes*	115	95.65%	100%
Cytochrome c oxidase subunit 2	cytochrome oxidase subunit II (mitochondrion)	ACJ63818.1	*G. gorilla gorilla*	227	99.56%	100%

[a]A unique identifier of an entry in the NCBI database [32].
[b]Describes the percentage of identical characters in proteins.
[c]Describes how much of the primate protein is covered by the modern human protein.
[d]Using QuickBLASTP [33], no similar protein was found. Using DELTA-BLAST [34], which yields better homology detection, we found ataxin-8, partial protein from *Varroa destructor* species with the length of 89, 100.00% identity and 98% query cover, but it does not belong to a primate.

ACKNOWLEDGMENT

This work was supported by Programa Operacional Factores de Competitividade–COMPETE (FEDER); and by national funds through the Foundation for Science and Technology (FCT), in the context of the projects UID/CEC/00127/2019, PTCD/EEI-SII/6608/2014 and the grant PD/BD/113969/2015.

REFERENCES

[1] J. Hendy *et al.*, "A guide to ancient protein studies," *Nature ecology & evolution*, p. 1, 2018.
[2] E. Cappellini, M. J. Collins, and M. T. P. Gilbert, "Unlocking ancient protein palimpsests," *Science*, vol. 343, no. 6177, pp. 1320–1322, 2014.
[3] M. G. Giuffrida, R. Mazzoli, and E. Pessione, "Back to the past: deciphering cultural heritage secrets by protein identification," *Applied microbiology and biotechnology*, vol. 102, no. 13, pp. 5445–5455, 2018.
[4] M. Sikora *et al.*, "The population history of northeastern Siberia since the Pleistocene," *Nature*, p. 1, 2019.
[5] R. Sawafuji *et al.*, "Proteomic profiling of archaeological human bone," *Royal Society open science*, vol. 4, no. 6, p. 161004, 2017.
[6] M. Buckley, "Paleoproteomics: An introduction to the analysis of ancient proteins by soft ionisation mass spectrometry," pp. 31–52, 2018.
[7] Y.-C. Lee *et al.*, "Evidence of preserved collagen in an Early Jurassic sauropodomorph dinosaur revealed by synchrotron FTIR microspectroscopy," *Nature communications*, vol. 8, p. 14220, 2017.
[8] E. Willerslev *et al.*, "Long-term persistence of bacterial DNA," *Current Biology*, vol. 14, no. 1, pp. R9–R10, 2004.
[9] T. Higham *et al.*, "The timing and spatiotemporal patterning of Neanderthal disappearance," *Nature*, vol. 512, no. 7514, p. 306, 2014.
[10] F. Welker *et al.*, "Palaeoproteomic evidence identifies archaic hominins associated with the Châtelperronian at the Grotte du Renne," *PNAS*, vol. 113, no. 40, pp. 11 162–11 167, 2016.
[11] M. Krings *et al.*, "Neandertal DNA sequences and the origin of modern humans," *cell*, vol. 90, no. 1, pp. 19–30, 1997.
[12] R. E. Green *et al.*, "Analysis of one million base pairs of Neanderthal DNA," *Nature*, vol. 444, no. 7117, p. 330, 2006.
[13] J. P. Noonan *et al.*, "Sequencing and analysis of Neanderthal genomic DNA," *science*, vol. 314, no. 5802, pp. 1113–1118, 2006.
[14] R. E. Green *et al.*, "A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing," *Cell*, vol. 134, no. 3, pp. 416–426, 2008.
[15] ——, "A draft sequence of the Neandertal genome," *science*, vol. 328, no. 5979, pp. 710–722, 2010.
[16] D. Reich *et al.*, "Genetic history of an archaic hominin group from Denisova Cave in Siberia," *Nature*, vol. 468, no. 7327, p. 1053, 2010.
[17] K. Prüfer *et al.*, "The complete genome sequence of a Neanderthal from the Altai Mountains," *Nature*, vol. 505, no. 7481, p. 43, 2014.
[18] S. Castellano *et al.*, "Patterns of coding variation in the complete exomes of three Neandertals," *PNAS*, vol. 111, no. 18, pp. 6666–6671, 2014.
[19] Q. Fu *et al.*, "An early modern human from Romania with a recent Neanderthal ancestor," *Nature*, vol. 524, no. 7564, p. 216, 2015.
[20] F. J. Ayala and C. J. C. Conde, *Processes in Human Evolution: The journey from early hominins to Neanderthals and modern humans*. Oxford University Press, 2017.
[21] J. Hawks, "Significance of Neandertal and Denisovan genomes in human evolution," *Annual Review of Anthropology*, vol. 42, pp. 433–449, 2013.
[22] D. Pratas *et al.*, "Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2017, pp. 235–242.
[23] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.
[24] L. Luo, D. Guo, R. T. Ma, O. Rottenstreich, and X. Luo, "Optimizing Bloom filter: challenges, solutions, and comparisons," *IEEE Communications Surveys & Tutorials*, 2018.
[25] A. Goel and P. Gupta, "Small subset queries and bloom filters using ternary associative memories, with applications," *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 1, pp. 143–154, 2010.
[26] M. Dietzfelbinger, "Universal hashing and k-wise independent random variables via integer arithmetic without primes," in *STACS*. Springer, 1996, pp. 567–580.
[27] FRUIT. [Online]. Available: https://github.com/smortezah/fruit
[28] Max Planck Institute for Evolutionary Anthropology. [Online]. Available: http://cdna.eva.mpg.de/neandertal/exomes/proteins
[29] Universal Protein Resource (UniProt). [Online]. Available: https://www.uniprot.org/uniprot
[30] A. Sajantila, "Editors' pick: Contamination has always been the issue!" *Investigative Genetics*, vol. 5, no. 1, p. 17, 2014.
[31] J. Dabney, M. Meyer, and S. Pääbo, "Ancient DNA damage," *Cold Spring Harbor perspectives in biology*, vol. 5, no. 7, p. a012567, 2013.
[32] National Center for biotechnology Information (NCBI). [Online]. Available: https://www.ncbi.nlm.nih.gov
[33] E. W. Sayers *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic acids research*, vol. 47, no. Database issue, p. D23, 2019.
[34] G. M. Boratyn *et al.*, "Domain enhanced lookup time accelerated BLAST," *Biology direct*, vol. 7, no. 1, p. 12, 2012.