Pairwise DNA Oligonucleotide Comparison

By Kayvan Mivehnejad (Kmi7682@gwu.edu)

Graduate Class of 2024 in Biomedical Informatics

School of Medicine and Health Science

The George Washington University

Preceptor Dr. Krystl Haerian

Program Director of Biomedical Informatics

Department of Clinical Research and Leadership

The George Washington University

Abstract

Objective: To Develop an algorithm with C++ that will compare a pair of DNA oligonucleotide sequences for evolutionary correlations with the ability to show similarities, recognize the variations and nominate a complementary RNA for their pathogenesis implications.

Materials and Methods: The overlapping regions of gene fragments (contigs) were aligned using simplified indexing to quantitatively measure scanned sites for the highest identical score and detect a seed. Dynamic programming methods were used to fill the substitution matrix and back-trace for the highly scored global/local alignments.

Results: The locus of the subject sequence with the highest identity score was successfully anchored without any parallel computing techniques or GPU acceleration. Practical testing demonstrated the ability of this API tool to detect seeds within identical sites and suggest a candidate RNA primer in a text-based output and JSON format.

Discussion and Conclusion: This API, like the BLAST tool, utilizes position-independent alignment algorithms, while other tools use position-specific profiles against homologous sequences. For correlation between residues emitted on bases with state paths, probability distribution tools like HMMER3 use the Hidden Markov Model (HMM) to predict the position of 5′/3′splice sites and transition between exons and introns. The developed API shows potential for the design of a complementary RNA primer while aligning a pair of DNA oligonucleotides to aid in various genetic and evolutionary studies.

Kayvan Mivehnejad, kmi7682@gwu.edu

Introduction

Deoxyribonucleic acid (DNA) is a macromolecule polymer in the form of double-helical strands that stores information and is comprised of four monomers of phosphoric acid alternating with deoxyribose sugar on the backbone with appendage bonds of acidic nucleotides like the rungs of a ladder. The nitrogenous bases in nucleotides form hydrogen bonds between one molecule of purine (Adenine or Guanine) and one cyclohexane ring of pyrimidine (Cytosine or Thymine). Therefore, the Adenine can only have a complementary bond with the Thymine, while the Guanine can only bond with the Cytosine. The phosphate group on carbon five prime also have a tendency to link to the oxygen on carbon three prime (Fig. 1.) to form an electromagnetic phosphodiester bond on the backbone.

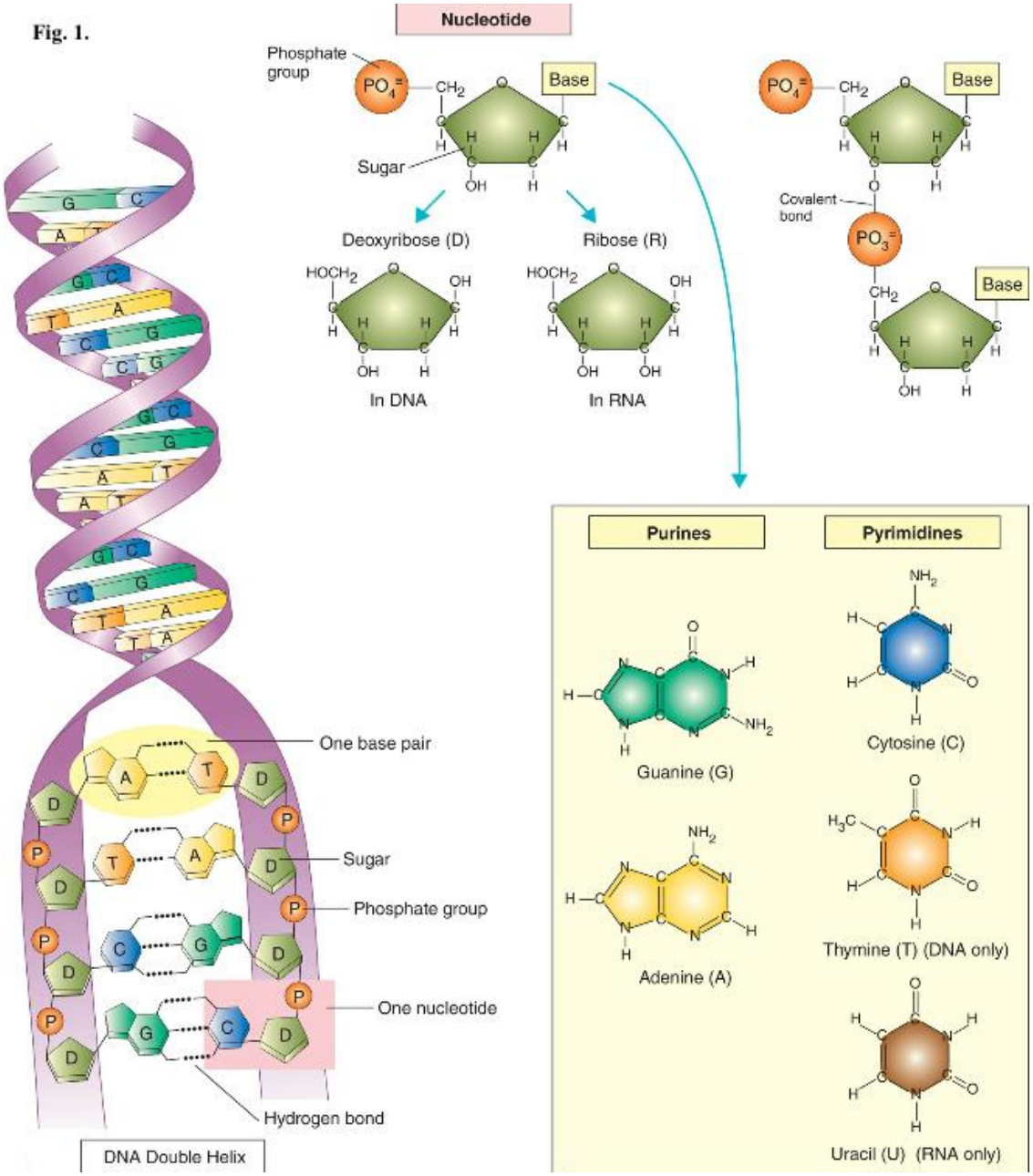
The human genome comprises three billion base pairs with only three million differences (about 0.1%) among individual DNA sequences. The importance of nucleotide sequence comparison is embodied in forensic investigations or when the genetic markers of two species sequences are compared to study their functionality. For instance, Hepatocellular carcinoma(HCC), one of the most common types of liver cancer (Muflikhah & Santoso, 2017, p. 1786), has disclosed similar DNA sequences in hepatocytes with inner core protein of Hepatitis-B antigen (HBcAg).

The Primary point mutation during gametic meiosis can cause genetic diseases when two inherited alleles of homologous genes, one paternal and the other maternal, are mutated aside from chromosomal crossover intermixing. Moreover, Epigenetic changes (e.g., methylation of cytosine or histone modifications) can change the order of these nucleotides for expression of

Kayvan Mivehnejad, kmi7682@gwu.edu

different traits. Post-translational alterations in the functionality of the proteins can also contribute to the pathogenesis of diseases like Cancer or Parkinson's.

Fig. 1.



Note: Double Helix Structure of DNA with Four Nucleotide Bases of Purines and Pyrimidine Forming Hydrogen and Phosphodiester Bonds. Adapted from Reisner, E.G. & Reisner, H.M. (2022). Crowley's An Introduction to Human Disease: Pathology and Pathophysiology Correlations Eleventh Edition, Jones & Bartlett Learning.

Kayvan Mivehnejad, kmi7682@gwu.edu

The study of DNA sequences is also crucial in the pathogenesis of infections and when pathogens flood the bloodstream, and penetrate cells' cytoplasm with endocytosis to hijack cellular structures. Some evolved DNA viruses infiltrate the chain of nucleotides inside the nucleus to multiply their genetic code. Integration of viral DNA with the host genome can disrupt regulatory signal pathways and cause continuous cellular growth and differentiation. Therefore, two biochemists, Jennifer Doudna and Emmanuelle Charpentier, revolutionized the treatment of genetic diseases (Gostimskaya, 2022, p. 777) by adopting the interspaced palindromic repeats observed in the immunity of procaryotic cells to synthesize complementary RNA sequence for the viral DNA (guided RNA) and attaching it to CRISPR-Associated Protein9 (CAS9) scissor enzyme to regulate gene expression or disrupt duplication of a viral DNA.

Methods

A traditional DNA sequencing method called Sanger Sequencing would split the Genome into a myriad of small fragments (about 900 base pairs long) to repeatedly go through three steps of denaturing in high temperate, then annealing with primer sequence in a lowered temperature, and extending nucleotides by DNA polymerase using replication methods such as polymerase chain reaction (PCR). The extension stops at chain-terminating dideoxy fluorescent-labeled nucleotides to visualize different read lengths on gel electrophoresis because the dideoxy nucleotide is missing a hydroxyl group on carbon three prime.

The fragments (contigs) with overlapping regions are reassembled to build longer scaffolds with recorded quality scores and then get passed to alignment tools like Burrow-Wheeler Aligner

(BWA) or bowtie to use FM-indexing (full-text) with typical Graphical Processing Units (GPU) acceleration or other parallel computing methods to map these contigs against reference sequence for highlighting variations in file formats like Variant Call Format (VCF). The new sets of DNA sequencing technologies are called Next Generation Sequencing (NGS), which essentially runs large numbers of Sanger reactions in parallel for fast and cost-reduced sequencing.

Two types of dynamic programming algorithms can solve the alignment by dividing it into many small subproblems. The global alignment algorithms map the entire length of sequences for similarity to fill a scoring matrix and back-trace scores penalties for returning the possibilities with the highest scores. The local alignment algorithms look for similarity in sequences' specific regions and calculate a pointing system for match/mismatch nucleotides. The Basic Local Alignment Sequencing Tool (BLAST) is the most well-known local sequence alignment tool maintained by the National Institute of Health (NIH).

BLAST is developed in C and C++, so I decided to investigate the domain of pairwise DNA sequence comparison by practical design of a C++ style Application Programming Interface (API) that can quantitatively measure the most extended identical segments of two sequences for seeds detection and represent suboptimal variations engaging features of dynamic programming. We can scan the subject sequence only in highly exact regions for seeds to save memory and design a fixed-length primer candidate using a method similar to FM-index that locates the occurrence of position rather than using hash tables (Liu et al., 2022, p. 4).

The designed algorithm can align a query DNA oligonucleotide with M base pairs with the subject DNA sequence of N nucleotides to anchor the identical sites (Ezz El-Din Rashed et al., 2021, p. 109522) while enhancing the asymptotic runtime of O(MN) for further processing of the highest scored site. A combination of seed-extend technique (Bayat et al., 2019, p. 1) and dynamic programming split the chain of subject (target) sequence to depict the primary alignment with seeds, then engages the smith-waterman substitution matrix to quantitatively assess the alignment with limited possibilities of single point mutations and INDEL variations (Zhao et al., 2013, p. 2).

Smith-Waterman Algorithm

Local alignment is interested in finding the most similar substring pair among two sequences. Smith and Waterman proposed a solution for local alignment that computes the optimal local alignment using dynamic programming. If we have two strings S[1, …, n] and Q[1, ..., m], then the maximum score V(i , j) is for the global alignment of A and B when all substring A of S end at i, and all substring B of Q end at j, where $1 \leq i \leq n$ and $1 \leq j \leq m$. The formula is divided into two cases: (1) for when i=0 or j=0, and (2) for when both i>0 and j >0. In case (1), we will have empty strings found on S, or Q so the equation will become (Sung, 2010, p. 39) :

$$V(i,0) = 0 \text{ for } 0 \leq i \leq n$$
$$V(0,j) = 0 \text{ for } 0 \leq j \leq m$$

For case (2), where both i>0 and j>0, there are two scenarios. First, for when both Strings S and Q are empty or V(i, j) =0. For the second scenario, within best alignment of some substring of S ending at i and some substring of Q ending at j, the last pair of aligned characters should be

Kayvan Mivehnejad, kmi7682@gwu.edu

either match/mismatch, delete, or insert. To get the optimal score, we choose the maximum value among zero and these three cases:

$$V(i,j) = \max \begin{cases} 0 & \textit{align empty strings} \\ V(i-1,j-1) + \delta(S[i], Q[j]) & \textit{match/mismatch} \\ V(i-1,j) + \delta(S[i], \_) & \textit{delete} \\ V(i,j-1) + \delta(\_, Q[j]) & \textit{insert} \end{cases}$$

The optimal local alignment score is $\max_{i,j} V(i , j)$. Smith and Waterman proposed that this score can be computed by filling in the matrix V row by row using the above recursive equations. For example, consider Q = ACAATCG and S = CTCATGC. Assume a match score is +2 and an insert/delete score is -1. Matrix V shows a maximum score of 6 (Fig. 2.). Through back-tracing from V(7,6) and according to scores filled for match/mismatch nucleotides, we can show in which sequence insertion or deletion occurred.

S= C– AT– G
Q= CAATCG

Fig. 2.

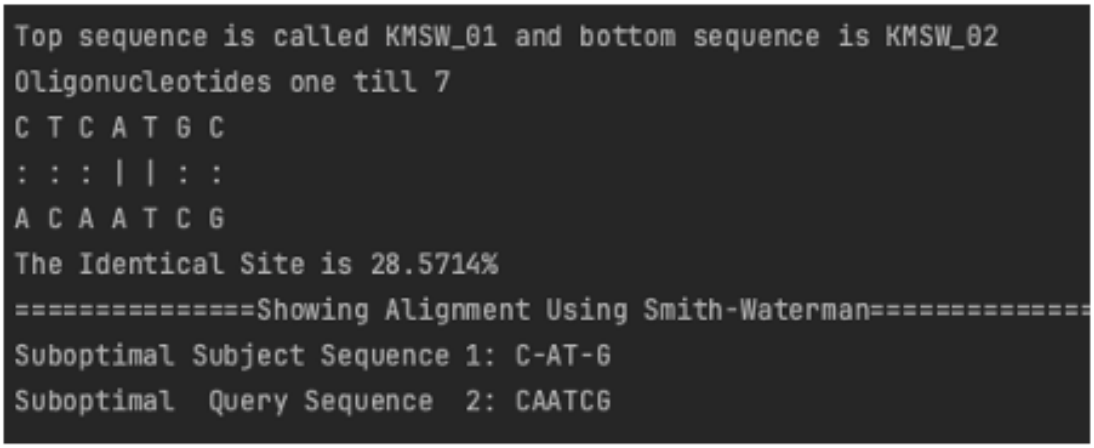
	_	C	T	C	A	T	G	C
_	0	0	0	0	0	0	0	0
A	0	0	0	0	2	1	0	0
C	0	2	1	2	1	1	0	2
A	0	1	1	1	4	3	2	1
A	0	0	0	0	3	3	2	1
T	0	0	2	1	2	5	4	3
C	0	2	1	4	3	4	4	6
G	0	1	1	3	3	3	6	5

Note: The Matrix for Local Alignment between Query=ACAATCG and Subject=CTCATGC.

Kayvan Mivehnejad, kmi7682@gwu.edu

Here in Figure 3., we are also showing the output of the pairwise comparison of two sequences in the API, which initially drew an alignment for an identity score of 28.57%, then used the Smith-Waterman algorithm to fill the scoring matrix and back-trace scores for the depiction of our local alignment. This is a condition in which both sequences have seven nucleotides; moreover, in scenarios with different oligonucleotide lengths, after the seeding step, the same method will be used to calculate the local alignment from the highest identical site.

Fig. 3.



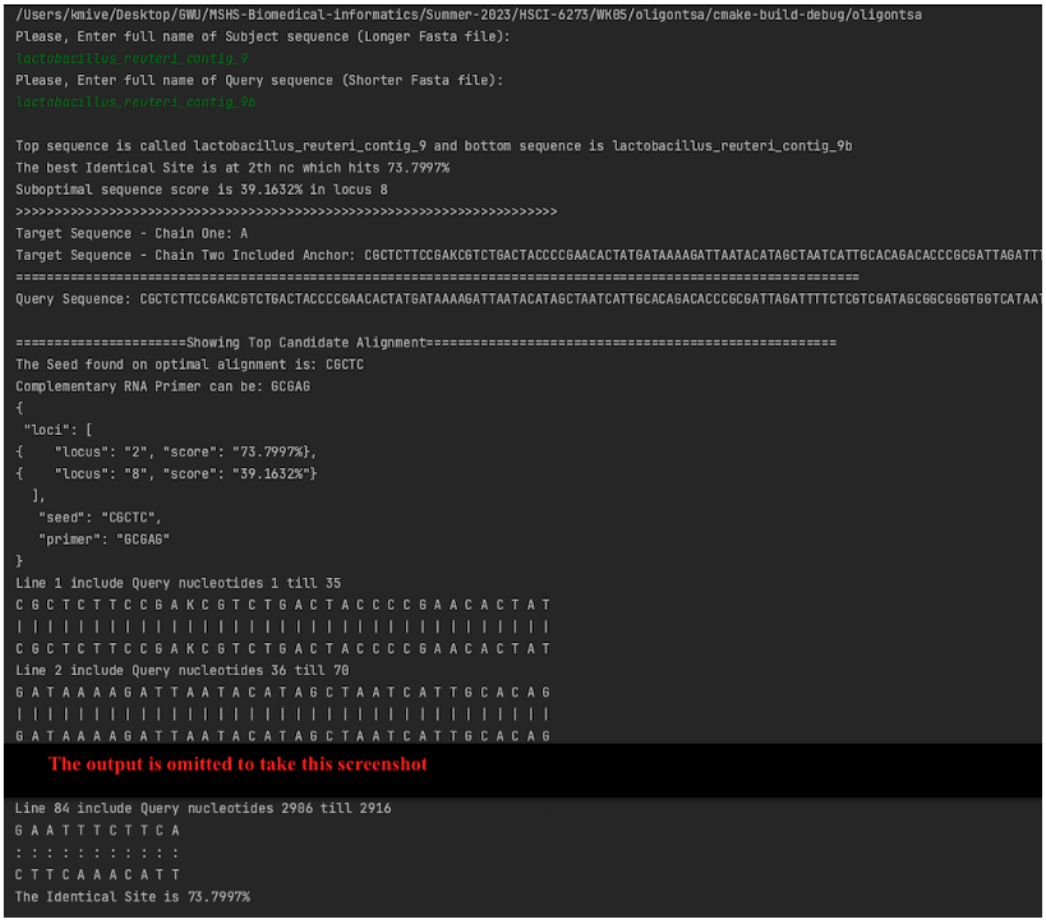
Note: The Output from Developed C++ API Shows The Identical Site and Local Alignment.

Results

Biological plausibility was exerted to the debugging process by referencing a study by Diaz et al. 2015, regarding to most types of cheese being a community of gram-positive bacteria (Lactic Acid Bacteria) that contain Pyruvoyl-dependent histidine decarboxylase enzyme (encoded by hdca gene). This enzyme can catalyze the synthesis of toxic dietary amine histamine from essential amino acid histidine, so the study compared gram-positive strands of bacteria downloaded from Genbank (as a marker for hdca) against amplicon sequence separations of a

few different types of Spanish cheese on acrylamide gel using PCR-DGGE (Polymerase Chain Reaction Denatured Gradient Gel-Electrophoresis).

Fig. 4.

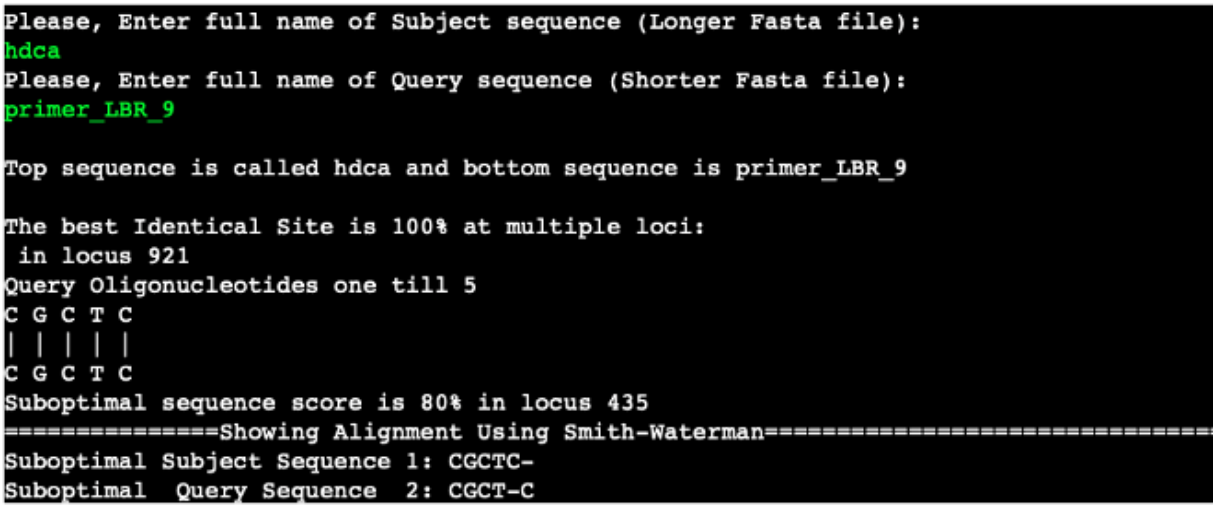


Note: Output from Binary Executable Showing the Pairwise Alignment of Two Lactobacillus-Reuteri Contigs with Indication of a Primer Candidate.

During debugging two reads of the Lactobacillus-Reuteri Contig-9 with 2923 base pairs were analyzed from within two fasta files by our compiled binary (posa) to pilot the pairwise comparison, then after detection of seed within the most identical site (Fig. 4.) the complementary RNA primer was suggested by the API as GCGAG. Subsequently, I anchored the CGCTC seed against the hdca gene downloaded from Genbank (Fig. 5.) to validate similarities between the Gene sequence and the primer candidate and used the smith-waterman algorithm to represent one of the possible variations against the identical site.

Kayvan Mivehnejad, kmi7682@gwu.edu

Fig. 5.



Note: The Output of API Shows Pairwise Comparison between hdca Gene and ACTCA Seed.

The typical genomic-wide comparison workflow demands an aligner tool to create the index of a reference file for rapid search; therefore, it could map each read for quantification estimation. The allele(s) from the indexed reference are commonly represented in the variant call format (VCF), the standard format since 2011, which is compatible with upstream and downstream workflows. VCF reflects on single nucleotide polymorphism (SNP), INDEL, or multi-nucleotide variants and uses layered character separators to represent the tree structure of data (Garisson et al., 2022). However, this API uses more versatile standard, the JSON, to generate the conclusive output of identity sites with their respective scores and the primer candidate.

Discussion

The DNA sequence alignment manifested itself in the phylogenic study of many species' DNA transcription to deduce their correlation in evolutionary trees. Methods such as BLAST use

position-independent local alignment algorithms to find similarities against a query sequence; however, position-specific profile databases are available online that can use statistical probabilistic models such as Hidden Markov Model (HMM) for identifying homologous sequences and draw alignments.

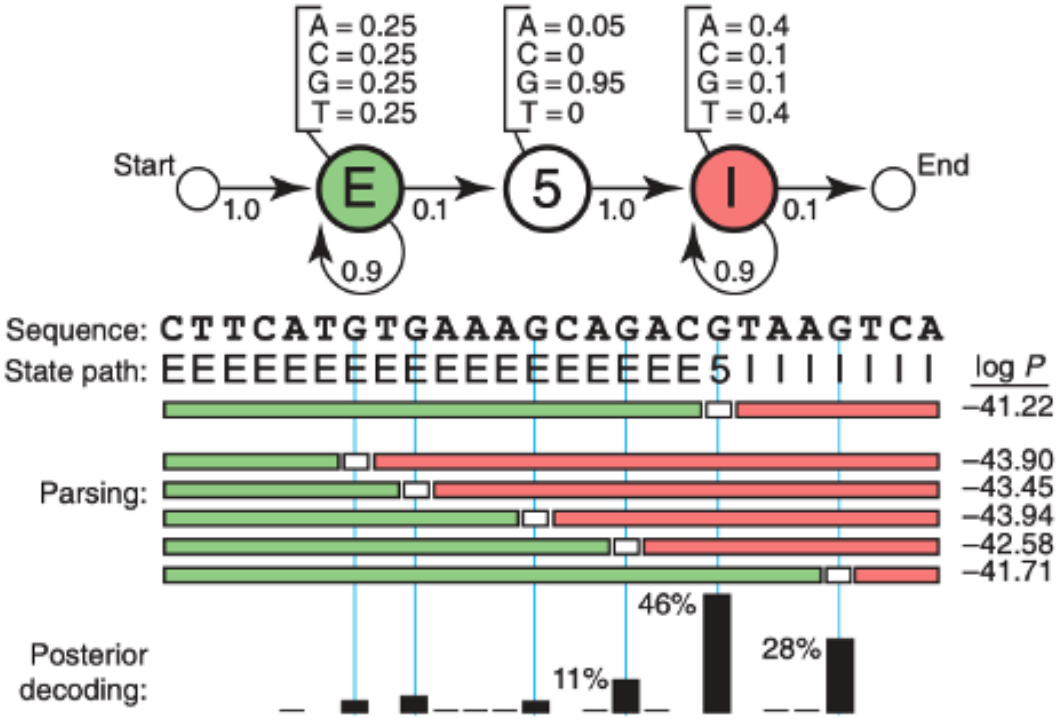
Key-residue of a nucleotide at specific positions can be evolutionarily near-neutral (silent mutations), which means certain insertions or deletions are tolerated better to avoid the changes in the product of translation (protein); some positions may tolerate certain substitutions while conserving physiochemical properties like hydrophobicity, charge or size (missense-conservative) like when codon TTC > TCC, then we will observe the translation of Arginine (Arg/R) with similar chemical structure and electrical charges instead of Lysine (Lys/K).

The pairwise DNA oligonucleotide comparison API developed here engaged position-independent methodology similar to BLOSUM and PAM tools because it is built upon the local alignment substitution matrix, so it could provide functionalities similar to BLAST. However, when are working on a specific sequence family with carefully constructed representatives, tools such as HMMER can help to build a position-specific profile from alignment and systematically search DNA sequence databases for more homologs.

One of the questions that we want to answer with aligning DNA sequences is at what site spliceosome protein snip (splice out) the intervening sequences (introns) from gene to fuse the assembly instructions of proteins (exons) and produce a protected and mature mRNA. Here we are showing a DNA sequence (Fig. 6.) with transition from Exon to Intron at a Five prime splice-

site (5´SS). To guess intelligently which nucleotide bases are Exon, intron, and find the location of 5´SS then we need different statistical properties. Exons have uniform base composition of 25% for each base, Introns are A/T rich (40% for each A or T , 10% for each G or C) and consensus nucleotide for 5´SS is almost always a G (95%) leaving 5% for A .

Fig. 6.



Note: A Toy HMM for 5 Prime Splice Site Recognition. Adapted from Eddy, S.R. (2004). What Is a Hidden Markov Model. (22)10. pp. 1315-1316.

We can draw an HMM by invoking three states, each having its own emission probabilities for bases (shown above the states) and transition probabilities (arrows) for linear order in which we expect states to occur. Every base is a residue we emit from the state's emission probability distribution for fourteen possible 5´SS state paths with six highly probable G positions that calculation of logarithmic probability using Viterbi algorithm will yield the highest probability of -41.22 on fifth G to become the position of transition from Exon to Intron (Eddy, 2004).

Kayvan Mivehnejad, kmi7682@gwu.edu

The model can be completed by adding a three prime splice site (3´SS) at the end of Introns with its emission probabilities and linear probability of transition to Exon. The principle of Viterbi Algorithm can find the probability of state l observing i at site x, as the probability of observing i in state l times the maximum value of probabilities found for the previous position x-1 observing j in state k times probability of state k transitioning to state l.

$$p_l(i, x) = e_l(i) \max(p_k(j, x-1).p_{kl})$$

It is convenient to use the log of the probabilities; therefore, we can compute sums instead of products, which is more efficient and accurate:

$$\text{Log } p_l(i, x) = \text{Log } e_l(i) + \max(\text{Log } p_k(j, x-1) + \text{Log } p_{kl})$$

For the sequence shown at Fig. 6. The probabilities of start could be bisected between emission for Exon or Intron sites as:

$$p_E(C, 1) = \text{Ln}(0.5) + \text{Ln}(0.25) = -0.7 + -1.38 = -2.08$$
$$p_I(C, 1) = \text{Ln}(0.5) + \text{Ln}(0.1) = -0.7 + -2.3 = -3$$

Similarly, we can find the probabilities for T at second site and so forth:

$$p_E(T, 2) = e_E(T) \max(P_E(C,1) + P_{EE},\ P_I(C,1) + P_{IE}, P_5(C,1) + P_{5E}, P_3(C,1) + P_{3E})$$

Kayvan Mivehnejad, kmi7682@gwu.edu

$$p_5\,(T, 2) = e_5(T)\ \max(P_E(C,1)+ P_{E5},\ P_I(C,1)+ P_{I5},\ P_5(C,1)+P_{55},\ P_3(C,1)+P_{35})$$

$$p_I\,(T, 2) = e_I(T)\ \max(P_E(C,1)+ P_{EI},\ P_I(C,1)+ P_{II},\ P_5(C,1)+P_{5I},\ P_3(C,1)+P_{3I})$$

$$p_3\,(T, 2) = e_3(T)\ \max(P_E(C,1)+ P_{E3},\ P_I(C,1)+ P_{I3},\ P_5(C,1)+P_{53},\ P_3(C,1)+P_{33})$$

The HMM only deals with the correlation between residues and states; however, there are alignment methods with graph (Ferragina Manzini) based index, like Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT2) that utilizes graph indices to search and incorporate the expanded model of human reference genome with over 14.5 million genomic variants (Kim et al., 2019). HISAT2 can extract identical segments from multiple long genomic sequences to nominate a representative called repeat sequence so it could run reads against that repeat-sequence for accurate, memory efficient and pruned on-disk alignment reckons compared to when we align against the entire genome using linear Burrow-wheeler aligner (BWA) or bowtie.

Conclusion

In-depth analysis and comparison of DNA sequences, together with clinical and environmental information, can uncover crucial insights, such as cancer risk prediction like when there are similarities between Hepatocellular carcinoma of hepatocytes and the Hepatitis-B antigen. This comparison can pave the way for groundbreaking advancements in patient outcome improvements through better-targeted therapies. The pairwise Oligonucleotide alignment is quintessential to providing pivotal data for precisely mapping identical segments and the discernment of suboptimal sequences with variations.

Kayvan Mivehnejad, kmi7682@gwu.edu

This C++ API (Mivehnejad, 2023, Supplementary Material 1) accentuates the potential ramifications of pairwise sequence comparison on how a nuanced understanding of genetic markers and mutations can contribute to revolutionary treatments of genetic diseases. The sequence alignment presented in this research can identify seeds at identical sites and may propel further innovations in complementary RNA primer design, providing a foundation for more sophisticated and accurate tools and methodologies in nucleotide sequence comparison. Graph Indexing on a high-performance computing cluster can improve computing time and assist with memory consumption of highly polymorphic genomic sequences to extend functionality of this API for searching numerous alleles of a gene.

References

Bayat, A., Gaeta, B., Ignjatovic, A., & Parameswaran, S. (2019). Pairwise alignment of nucleotide sequences using maximal exact matches. (20)261, https://doi.org/10.1186/s12859-019-2827-0

Diaz, M., Ladero, V., Redruello, B., Sanchez-Llana E., Del Rio, B., Fernandez, M., Martin, M.C., & Alvarez, M.A. (2015). A PCR-DGGE method for the identification of histamine-producing bacteria in cheese. (63), pp. 216-223. https://dx.doi.org/10.1016/j.foodcont.2015.11.035

Diaz, M., Ladero, V., Redruello, B., Sanchez-Llana E., Del Rio, B., Fernandez, M., Martin, M.C., & Alvarez, M.A. (2016). Nucleotide sequence alignment of hdca from Gram-positive bacteria. (6), pp. 674-679. https://doi.org/10.1016/j.dib.2016.01.020

Eddy, S.R. (2004). What is a hidden Markov model. (22)10. pp. 1315-1316. https://www.nature.com/articles/nbt1004-1315

Ezz El-Din Rashed, A., Amer H.A., El-Seddek, M., & El-Din Moustafa H., (2021). Sequence Alignment Using Machine Learning-Based Needleman-Wunsch Algorithm. (9), pp. 109522-109535. https://doi.org/10.1109/ACCESS.2021.3100408

Garrison, E., Kronenberg, Z.N., Dawson, E.T., Pedersen, B.S., & Prins, P. (2022). A Spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, hts-nim and slivar. 18(5). https://doi.org/10.1371/journal.pcbi.1009123

Gostimskaya, I., (2022). CRISPR-Cas9: A History of Its Discovery and Ethical Considerations of Its Use in Genome Editing. (87), p. 777-788. https://doi.org/10.1134/S0006297922080090

Kim, D., Paggi, J.M., Park, C., Bennter, C., & Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. (37), pp. 907 – 915. https://doi.org/10.1038/s41587-019-0201-4

Liu, H., Zou, Q., & Xu, Y., (2022). A novel fast multiple nucleotide sequence alignment method based on FM-index. (23)1, pp. 1-9. https://doi.org/10.1093/bib/bbab519

Muflikhah, L., & Santoso, E. (2017). Pairwise Sequence Alignment between HBV and HCC Using Modified Needleman-Wunsch Algorithm. (15)4, pp. 1785-1793. https://doi.org/10.12928/TELKOMNIKA.v15i4.5813

Reisner, E.G. & Reisner, H.M. (2022). Crowley's An Introduction to Human Disease: Pathology and Pathophysiology Correlations Eleventh Edition, Jones & Bartlett Learning

Sung, W-K, (2010). Algorithms in Bioinformatics: A Practical Introduction. Chapman & Hall/CRC Mathematical and Computational Biology Series. CRC Press, Taylor & Francis Group

Zhao M., Lee W-P, Garrison E.P., & Marth G.T., (2013). SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. (8)12, https://doi.org/10.1371/journal.pone.0082138

Supplementary Materials

1. Mivehnejad, K. (2023). Pairwise DNA Oligonucleotide comparison. GitHub. https://github.com/mivehk/Oligonucleotide_Aligner