Journal of **Concepts in Structural Biology & Bioinformatics**

GENOMICS & PROTEOMICS ARTICLES

Nucleic Acid Code Translation to Proteins and Reverse (NCTPR): A Versatile Online Bioinformatics Tool for Research and Educational Applications

RACHEDI Abdelkrim*

Laboratory of Biotoxicology, Pharmacognosy and biological valorisation of plants, Faculty of Natural and Life Sciences, Department of Biology, University of Saida - Dr Moulay Tahar, 20100 Saida, Algeria.

***Correspondence:** RACHEDI Abdelkrim – Email: abdelkrim.rachedi@univ-saida.dz, bioinformatics@univ-saida.dz
Published: 03 October 2023

Abstract

The elucidation of nucleic acid sequences is a cornerstone of genetic research and education, necessitating tools that can accurately translate DNA and RNA sequences into proteins and vice versa. "Nucleic Acid Code Translation to Proteins and Reverse" (NCTPR) is an innovative online bioinformatics tool designed to meet this need in the post-genomic era where large genomic data are made available and which are subject of gene discovery. NCTPR enables users to translate nucleotide sequences into amino acids and perform reverse translations, accommodating various genetic codes applicable to eukaryotic and prokaryotic organisms. The tool provides options to translate across individual reading frames, all 3'-5' frames, all 5'-3' frames, or all six frames simultaneously, offering a flexibility for in-depth genetic analysis. Rigorous validation against known sequences and established tools like Expasy - Translate has confirmed its accuracy, making NCTPR a reliable resource for basic research and teaching. With a user-friendly web interface, NCTPR facilitates its integration into educational curricula, aiding in the practical understanding of molecular biology for students. This paper describes the development, functionality, and validation of NCTPR, highlighting its significance as a research and educational tool in the evolving landscape of bioinformatics.

Availability: https://bioinformatics.univ-saida.dz/nctpr/

Key words: Bioinformatics, Genomic Research, Nucleic Acid Translation, Reverse Translation, Protein Synthesis, Sequence Analysis, Synthetic Biology, Molecular Biology Education.

Introduction

The discovery of DNA and its pivotal role in genetics has been one of the most significant scientific advancements of the 20th century. The foundational work of Watson and Crick in 1953, which elucidated the double helix structure of DNA, marked the beginning of a new era in understanding genetic information and inheritance (Watson & Crick, 1953). The concept of genes as units of inheritance, initially proposed by Gregor Mendel in the 19th century through his work on pea plants, found a molecular basis through the understanding of DNA structure and function (Mendel, 1866).

Following the elucidation of DNA's structure, the central dogma of molecular biology, formulated by Francis Crick, provided a framework for understanding the flow of genetic information from DNA to RNA to protein (Crick, 1958). This dogma underscored two critical processes: transcription and translation. Transcription involves the conversion of DNA into messenger RNA (mRNA), a process that is regulated and highly complex (Berg et al., 2002). Translation, the subsequent step, involves the decoding of mRNA to synthesize proteins, which are essential for various cellular functions (Alberts et al., 2002).

The completion of the Human Genome Project at the dawn of the 21st century, often referred to as the post-genomic era, marked a significant milestone in genetics and bioinformatics (Collins et al., 2003). This era has been characterized by an explosion of genetic data, necessitating the development of robust, efficient, and user-friendly bioinformatics tools. Such tools are essential for translating the vast amounts of DNA and RNA sequence data into meaningful protein sequences, thereby enabling a deeper understanding of gene function and regulation (Mount, 2004). The translation of nucleic acid sequences to protein sequences is not just a routine task in molecular biology and genetics research but also a fundamental step in various applications, including disease research, drug development, and synthetic biology (Lander, 2011).

In this context, the need for accessible and accurate bioinformatics tools for sequence translation is more pronounced than ever. While several tools exist for this purpose, challenges in terms of user interface, accuracy, and flexibility in handling different reading frames persist. Addressing these challenges is crucial for advancing our understanding of genetic information and its implications in the post-genomic era.

In the wake of these scientific milestones, the development of practical and educational tools for understanding genetic processes has become a critical need. The tool has been crafted with this dual purpose in mind. Not only does it serve as a robust bioinformatics tool for professionals in the field of molecular biology,

genetics, and bioinformatics, but it has also been meticulously designed to aid in the educational context, particularly for master's students specializing in Biology, Biochemistry, and Biotechnology.

The educational aspect of this tool is particularly significant. In an academic setting, students often encounter challenges in grasping the complexities of molecular genetics, especially in understanding the intricacies of transcription and translation. This tool offers a hands-on experience, allowing students to input real sequences and observe the translation process. It effectively bridges the gap between theoretical knowledge and practical application, enabling students to visualize and comprehend the fundamental processes of molecular biology. By integrating this tool into coursework and assignments, educators can provide a more interactive and engaging learning experience, enhancing students' understanding of the subject matter.

Moreover, in a discipline that is rapidly evolving, such as bioinformatics, keeping pace with the latest advancements and methodologies is crucial. This tool, with its user-friendly interface and versatile functionality, is a step towards making complex bioinformatics processes accessible to a broader audience, including students and researchers alike. Its ability to translate DNA/RNA sequences into protein sequences across various frames, as well as perform reverse translations, makes it an invaluable resource in both research and education.

The tool available at "https://bioinformatics.univ-saida.dz/nctpr/" is not just a technical asset for scientific research but also a significant educational resource. It plays a dual role in advancing the understanding of genetic translation in the post-genomic era and in fostering educational growth in the field of molecular biology and bioinformatics.

Methods

The bioinformatics tool NCTPR is meticulously designed to encompass a wide range of genetic codes, catering to both eukaryotic and prokaryotic organisms. This section explains the methodology employed, particularly focusing on the tool's capability to handle various genetic codes in both translation and reverse translation processes.

Web Interface and Accessibility

Both the translation and reverse translation functionalities are accessible through a user-friendly web interface. The interface is built using HTML, ensuring compatibility and accessibility across various web browsers and devices. Users can input their sequences directly into the provided fields on the webpage. This design

choice enhances the tool's accessibility, making it suitable for both academic and research settings. The interface is intuitive, allowing users to easily choose between different translation frames or opt for reverse translation, depending on their requirements, see **Figure 1**.

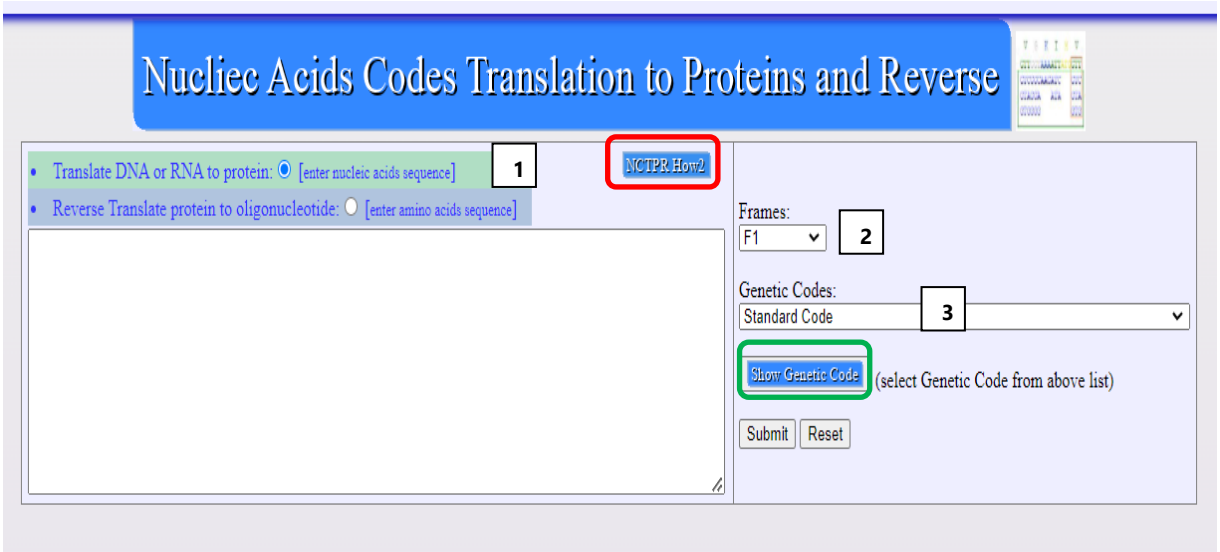

Figure 1. Screenshot shows The web interface of the bioinformatics tool, illustrating the input fields for nucleic acid and protein sequences, the options for translation frames, and the genetic code selection menu. Option 1 to sets whether to do Translation or Reverse Translation, option 2 to select what open reading frames to show and option 3. Encircled in red is an option link that open an instruction guide for the usage of the NCTPR tool. Encircled in green is an option link to display tables of genetic codes as per selected genetic code (from option 3)

Translation Algorithm with Multiple Genetic Codes

Developed using PHP, the translation algorithm is not limited to the standard genetic code, **Table 1**, but includes an extensive array of genetic codes pertinent to different organisms. For another genetic code example, see **Table 2** that represents the Vertebrate Mitochondrial genetic code. Notice the highlighted difference between the two genetic codes.

Table 1. Standard (Homo sapiens) genetic code

Genetic code: Standard			
Codon AA	Codon AA	Codon AA	Codon AA
TTT F	CTT L	ATT I	GTT V
TTC F	CTC L	ATC I	GTC V
TTA L	CTA L	ATA I	GTA V
TTG L	CTG L	ATG M	GTG V
TCT S	CCT P	ACT T	GCT A
TCC S	CCC P	ACC T	GCC A
TCA S	CCA P	ACA T	GCA A
TCG S	CCG P	ACG T	GCG A
TAT Y	CAT H	AAT N	GAT D
TAC Y	CAC H	AAC N	GAC D
TAA *	CAA Q	AAA K	GAA E
TAG *	CAG Q	AAG K	GAG E
TGT C	CGT R	AGT S	GGT G
TGC C	CGC R	AGC S	GGC G
TGA *	CGA R	AGA R	GGA G
TGG W	CGG R	AGG R	GGG G
AA: Amino Acids | *: Stop codon			

Table 2. Vertebrate Mitochondrial genetic code

Genetic code: Vertebrate Mitachondrial			
Codon AA	Codon AA	Codon AA	Codon AA
TTT F	CTT L	ATT I	GTT V
TTC F	CTC L	ATC I	GTC V
TTA L	CTA L	ATA M	GTA V
TTG L	CTG L	ATG M	GTG V
TCT S	CCT P	ACT T	GCT A
TCC S	CCC P	ACC T	GCC A
TCA S	CCA P	ACA T	GCA A
TCG S	CCG P	ACG T	GCG A
TAT Y	CAT H	AAT N	GAT D
TAC Y	CAC H	AAC N	GAC D
TAA *	CAA Q	AAA K	GAA E
TAG *	CAG Q	AAG K	GAG E
TGT C	CGT R	AGT S	GGT G
TGC C	CGC R	AGC S	GGC G
TGA W	CGA R	AGA *	GGA G
TGG W	CGG R	AGG *	GGG G
AA: Amino Acids | *: Stop codon			

This comprehensive approach ensures the tool's applicability across various biological domains. The inclusion of these codes allows for accurate translation of DNA/RNA sequences into amino acids, considering organism-specific variations in codon usage (Osawa et al., 1992). The list of genetic codes incorporated in the algorithm are shown in **Table 3**.

Table 3. The list of species for genetic codes included in the translation algorithms of the NCTPR tool.

- Standard Code
- Vertebrate Mitochondrial
- Invertebrate Mitochondrial
- Yeast Mitochondrial
- Yeast Nuclear (alternative)
- Mitochondrial Mold / Protozoan / Coelenterate + Mycoplasma / Spiroplasma
- Echinoderm and Flatworm Mitochondrial
- Flatworm Mitochondrial (alternative)
- Ascidian Mitochondrial
- Chlorophycean Mitochondrial
- Trematode Mitochondrial
- Scenedesmus obliquus Mitochondrial
- Thraustochytrium Mitochondrial
- Ciliate, Dasycladacean and Hexamita Nuclear
- Euplotid Nuclear
- Bacterial, Archaeal, and Plant Plastid
- Blepharisma Nuclear

Reverse Translation Algorithm with Genetic Code Variations

The reverse translation aspect of the tool is particularly innovative, **Figure 2**. The PHP algorithm, in this case, takes a protein sequence as input and generates all possible DNA sequences that could encode for it. This process is more complex due to the redundancy in the genetic code, where multiple codons can code for the same amino acid (Crick, 1966). The algorithm considers the known codons for each amino acid and generates a comprehensive set of DNA sequence possibilities. This functionality is critical for applications such as synthetic biology and genetic engineering, where designing DNA sequences based on desired protein products is a common task. This capability is significant in synthesizing DNA sequences for a diverse range of organisms, accommodating the genetic code variations found in different species. The algorithm computes all possible DNA sequences from a given protein sequence, taking into account the redundancy of the genetic code and the specific codon preferences of the organism in question (Andersson & Kurland, 1990).

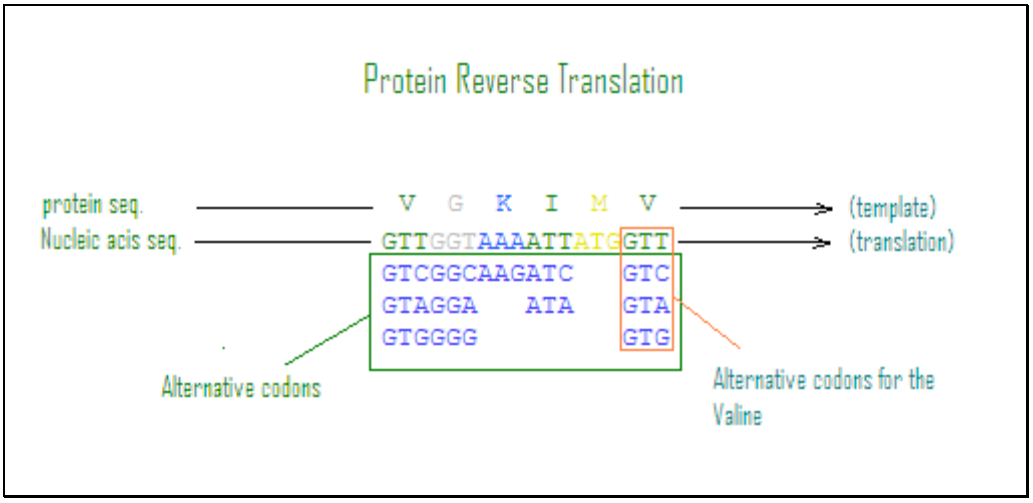

Figure 2. Protein reverse translation by the NCTPR. Top line exhibits a short peptide sequence (template) which is translated into variants of DNA sequences based on the alternative codons associated with each amino acid in the template.

How-to Guide

A clickable link, labeled "NCTPR How2", refer to **Figure 3**, directs users to a comprehensive how-to document that outlines the steps for using the tool. This guide is designed to help users quickly become proficient in the tool's operation, from sequence input to the analysis of translated results. The guide is essential for educational settings, where students may be unfamiliar with bioinformatics tools. The inclusion of this user support feature demonstrates the tool's commitment to accessibility and education, further establishing NCTPR as an invaluable resource for both research and learning in molecular biology.

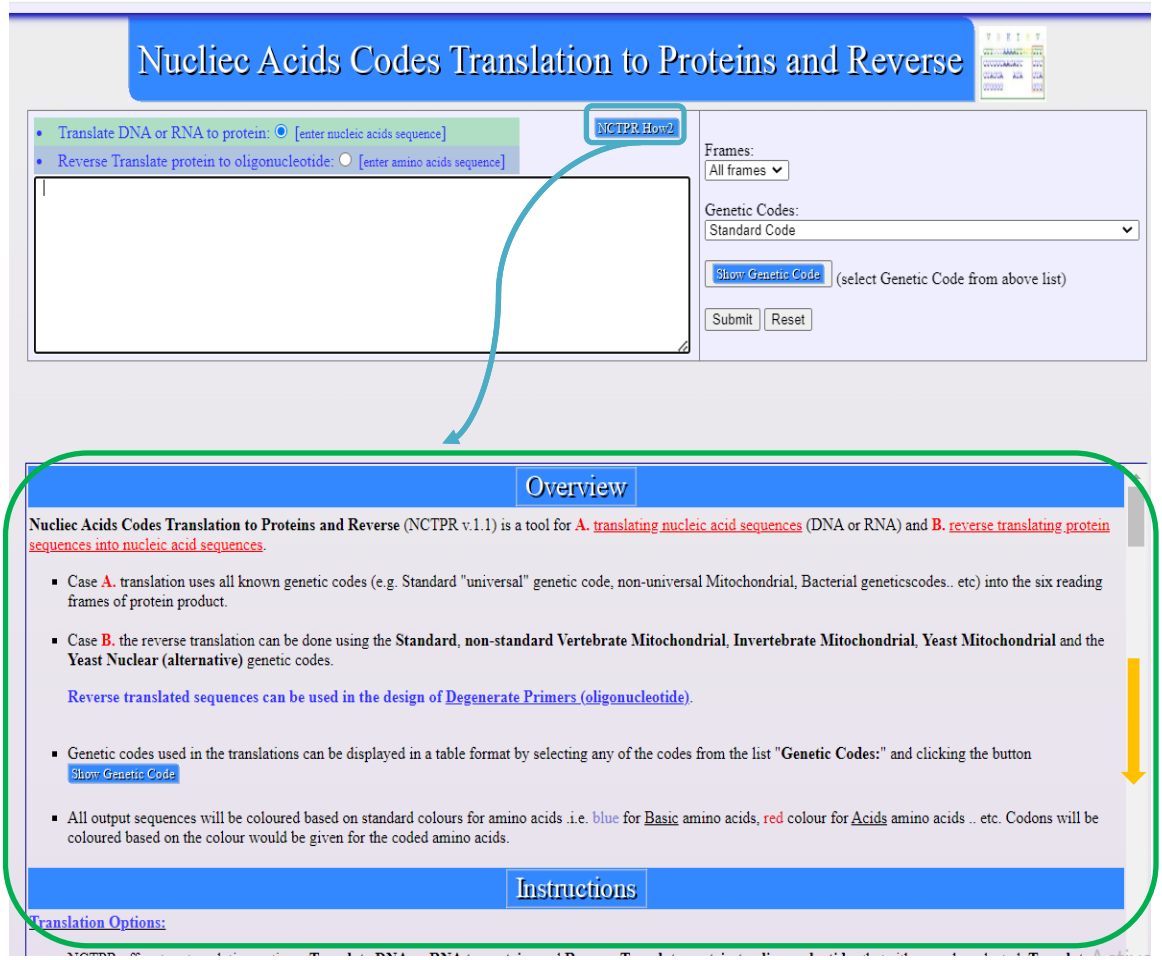

Figure 3. Screenshot shows part of the usage guidance of the NCTPR tool.

Results and Discussion

Features and Functionality

The bioinformatics tool, available at "https://bioinformatics.univ-saida.dz/nctpr/", presents a comprehensive suite of features for the translation of nucleic acid sequences and the reverse translation of protein sequences. This section provides an overview of these functionalities, emphasizing the flexibility and user-centric design of the tool. Refer above to **Figure 1** that showcase the tool's interface, which is central to the user experience.

Translation Options with Visual Interface

The tool's interface, as depicted in **Figure 1**, offers a straightforward pathway to access its powerful translation capabilities. Users are presented with the following options, each facilitating a unique aspect of sequence analysis:

- **Individual Frames (F1-F6):** Users can select any single frame for translation, providing a focused view on a specific reading frame.

- **All 3'-5' Frames and All 5'-3' Frames:** These options enable translation across all frames in the respective directions, allowing users to fully explore the coding potential of a sequence.

- **All Six Frames Simultaneously:** For a comprehensive analysis, this option translates all possible frames at once, a critical feature for exhaustive genomic studies.

The **Frames** here refer to the Open Reading Frames (ORFs) that cover all the reading possibilities available for stretch of a DNA sequence.

Selection of Genetic Codes

The selection process for these features is facilitated by HTML "**select**" tags shown as the option **Frames**, which when used, ensures a seamless and intuitive user experience. As shown above, **Figure 1**, also illustrates the tool's ability to adapt to various genetic codes, crucial for accurate translation across different organisms. The dropdown menu allows users to choose from a list of genetic codes, including but not limited to:

- Standard Code
- Vertebrate Mitochondrial
- Invertebrate Mitochondrial
- ...and many more, each corresponding to the specialized codon usage of different organisms or organelles.

An example for the use of the this default option, "**Translate DNA or RNA to protein:**" illustrates the translation of the DNA sequence, **Figure 4**, to all the 6 possible ORFs **Figure 5**.

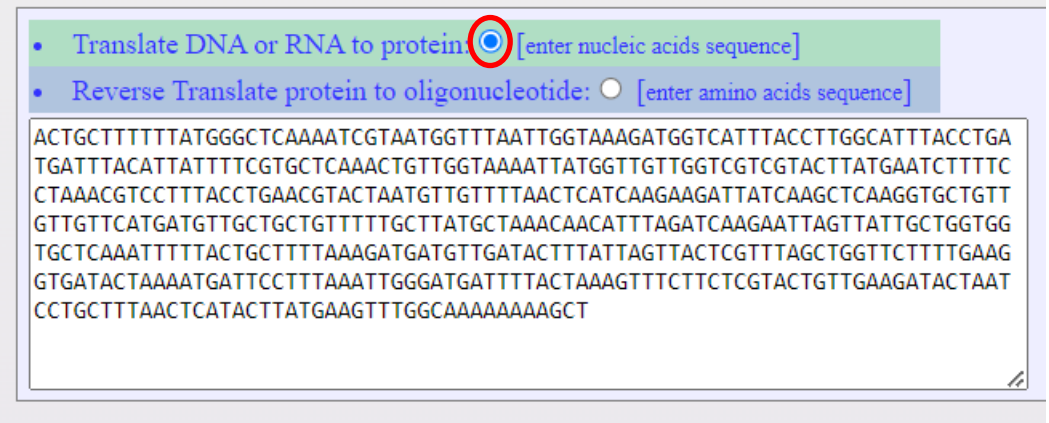

Figure 4. Screenshot show the Input text area of the NTCPR where a DNA/RNA sequence is typed in or pasted in. The translation to protein is selected by default (as encircled in red).

Figure 5. Screenshot show the output of NTCPR for the translation of the DNA sequence, in Figure 4, to all ORF frames.

In this example case, the sequence in the first frame (>Frame 1) is considered the correct translation as it continuous sequence and uninterrupted by stop codons, denoted by the asterisk symbol (*), as compared to the other sequences in the rest of frames.

Reverse Translation Feature

A key innovation of the tool is its reverse translation functionality and this feature is designed for many reasons including synthetic biology applications where a protein sequence must be reverse-engineered into a nucleic acid sequence. By considering the degeneracy of the genetic code, the tool can generate all possible DNA sequences that encode for a given protein, offering invaluable insights for genetic design and manipulation.

The interface is streamlined for ease of use, ensuring that both novices and experienced researchers can navigate and utilize the tool with minimal training. The design is responsive, providing a consistent experience across various devices and screen sizes.

In an example for the use of the option "**Reverse Translate protein to oligonucleotide:**" is illustrated below where the protein sequence, **Figure 6**, is reverse translated into alternative DNA sequences based on the number of codons each amino acid may have, see **Figure 7**.

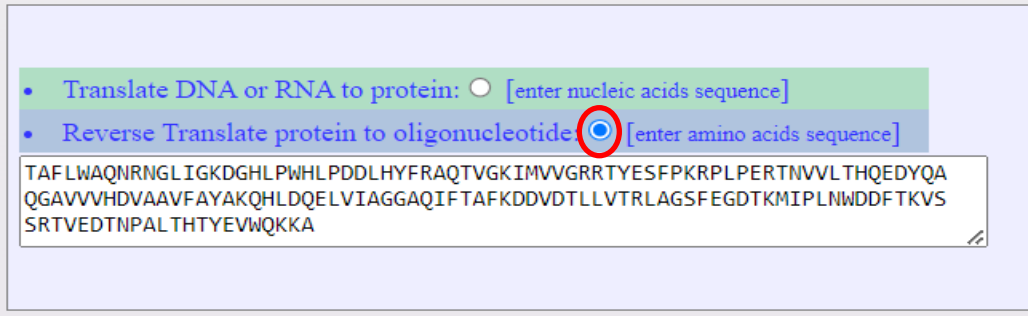

Figure 6. Screenshot show an example of a protein sequence as type in or pasted in input text area of the NTCPR. The reverse translation to DNA sequences is selected by user (shown encircled in red).

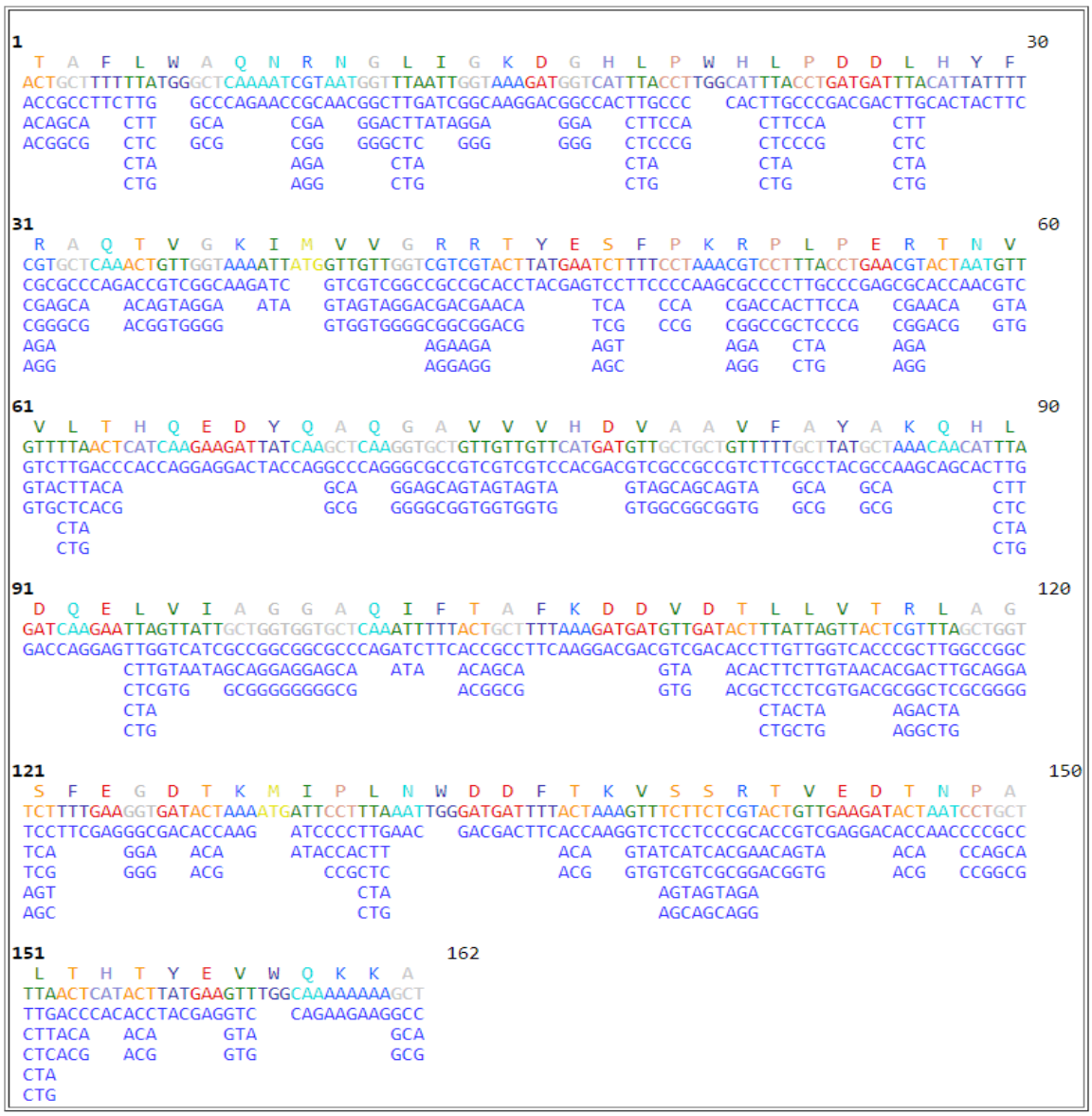

Figure 7. Screenshot show the DNA sequences generated by the reverse translation of the protein sequence seen above in Figure 6.

In this example case, all of the alternative sequences may be of importance to research being sought by the users for learning and research undertaking in synthetic biology and genetic engineering.

Testing and Validation

To ensure the reliability and accuracy of the "Nucleic Acids Codes Translation to Proteins and Reverse" tool, rigorous testing and validation procedures were implemented. This section describes the methods used to test the tool and the outcomes that substantiate its precision and efficacy.

The validation process involved extensive testing with known genetic sequences. These sequences, which have well-documented translations, provided a robust basis for assessing the tool's performance. By inputting these sequences into the tool and analyzing the resulting translations, a direct comparison could be made with expected outcomes based on established biological knowledge.

Comparison with Established Tools

For a comprehensive validation, the tool's translations were compared with those generated by the Expasy - Translate tool, one of the most widely recognized and utilized translation tools in the bioinformatics community (Gasteiger et al., 2003). The Expasy tool is acclaimed for its accuracy and has been a reference point for sequence translation in the research field for many years.

Validation Results

The comparison revealed a 100% agreement between the translations provided by the "Nucleic Acids Codes Translation to Proteins and Reverse" tool and the Expasy - Translate tool. This congruence was observed across all tested sequences, which included a variety of complexities and reading frames. The results exhibit the tool's capability to handle intricate sequence translations with a high degree of accuracy, thereby validating its effectiveness and reliability for both academic and research purposes.

Significance of Results

The validation tests confirm that the tool meets the high standards required for accurate translation of nucleic acid sequences. Such precision is essential for applications in genomics, proteomics, and synthetic biology, where even minor discrepancies can lead to significant misinterpretations. The tool's accuracy ensures that users can trust the results for critical analyses and experimental planning.

Conclusion

The development and deployment of the "Nucleic Acids Codes Translation to Proteins and Reverse" tool represent a significant contribution in the realm of nucleic acid sequence analysis. The tool, accessible through "https://bioinformatics.univ-saida.dz/nctpr/", stands as a testament to the advancements in bioinformatics applications for both research and educational purposes at the University of Saida, Algeria.

This tool's high accuracy in translating and reverse-translating sequences, as confirmed by rigorous testing and validation against established resources like the Expasy - Translate tool (Gasteiger et al., 2003), offers invaluable support to researchers and students. Its comprehensive suite of features, including the ability to process various genetic codes and translate across all possible frames, addresses the needs of a broad spectrum of genomic research. The reverse translation feature, in particular, is critical for synthetic biology and genetic engineering, where precise DNA sequence design is paramount.

Furthermore, the tool's intuitive interface and flexibility make it an excellent resource for educational settings, particularly in enhancing the practical understanding of molecular biology for students in the fields of Biochemistry and Biotechnology. By providing hands-on experience with actual sequence data, the tool aids in bridging the gap between theoretical knowledge and real-world applications.

The "Nucleic Acids Codes Translation to Proteins and Reverse" tool not only fulfills a pivotal role in current research endeavours but also contributes significantly to the educational landscape, preparing the next generation of scientists with state-of-the-art bioinformatics tools.

References

- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell.* Garland Science.
- Andersson, S.G.E., & Kurland, C.G. (1990). Codon Preferences in Free-Living Microorganisms. Microbiological Reviews, 54(2), 198-210.
- Andrianantoandro, E., Basu, S., Karig, D.K., & Weiss, R. (2006). Synthetic Biology: New Engineering Rules for an Emerging Discipline. Molecular Systems Biology, 2, Article 2006.0028.
- Berg, J.M., Tymoczko, J.L., & Stryer, L. (2002). *Biochemistry.* W H Freeman.
- Collins, F.S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. *Science, 300*(5617), 286-290.

- Crick, F. (1958). On Protein Synthesis. *Symposia of the Society for Experimental Biology, 12*, 138-163.
- Crick, F.H.C. (1966). Codon—Anticodon Pairing: The Wobble Hypothesis. Journal of Molecular Biology, 19(2), 548-555.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research, 31*(13), 3784-3788.
- Lander, E.S. (2011). Initial Impact of the Sequencing of the Human Genome. *Nature, 470*(7333), 187-197.
- Lerdorf, R. (1995). PHP: Hypertext Preprocessor. PHP Documentation.
- Mendel, G. (1866). Experiments on Plant Hybridization. *Proceedings of the Natural History Society of Brunn, 4*, 3-47.
- Mount, D.W. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.
- Nielsen, H. (2005). Predicting Secretory Proteins with SignalP. In Methods in Molecular Biology (pp. 59-73). Humana Press.
- Osawa, S., Jukes, T.H., Watanabe, K., & Muto, A. (1992). Recent Evidence for Evolution of the Genetic Code. Microbiological Reviews, 56(1), 229-264.
- Watson, J.D., & Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature, 171*(4356), 737-738.
- Watson, J.D. (1976). Molecular Biology of the Gene. W.A. Benjamin.