APPROVED: 20 December 2023
doi: 10.2903/sp.efsa.2024.EN-8561

Reference: OC/EFSA/GMO/2021/01

Refinement of the Risk Assessment Methodology for Open Reading Frames in GMO Applications

Daniele Urbani, Marianna Penzo, Martina Evangelisti, Marco Daniele Parenti, Alberto Del Rio

Innovamol Srl, Alma Mater Studiorum – University of Bologna

Abstract

A literature search was performed using PubMed to identify relevant studies on Open Reading Frames (ORFs) relevant to the risk assessment of GMOs. The collection of information on ORFs was steered by EFSA guidance for systematic reviews. The search queries allowed the retrieval a total of 15.484 non-redundant references. The relevance of these documents was first assessed by screening titles and abstracts against specific inclusion and exclusion criteria related to risk assessment. This was followed by full-text screening, which resulted in a total of 307 relevant documents. Information from these documents was extracted to emphasise criteria for the definition, prediction, and selection of ORFs, possibly highlighting the context of risk assessment of GMOs. Criteria such as codon identity, nucleotide composition, and mRNA secondary structure may be pertinent for developing new methods for risk assessment. However, the analysis revealed several limitations in the context of risk assessment, including the lack of structured data, diversity of application domains, paucity of information in food/feed, and the reliability of specific criteria for ORF definition, prediction and selection, among others. The analysis of prediction tools demonstrated that the generation of *de novo* experimental data or specific datasets is a critical factor. Nonetheless, certain features of ORF nucleotide sequences might prove useful in assessing the likelihood of expression of relevant ORFs for risk assessment of GMOs, but the criteria underlying this likelihood require further research and effort to be embedded in a tool. Bearing this in mind, and based on the information from the literature search, an evaluation was conducted regarding the potential of integrating various tools, their strengths and weaknesses and the challenge to integrate this knowledge into a single tool. A conceptual workflow is proposed for navigating these challenges and limitations and is presented as an attempt to integrate and streamline the tools and methods currently available.

Key words: Extensive literature search, open reading frames, ORF, GMOs, risk assessment

Question number: EFSA-Q-2023-00895
Correspondence: NIF@efsa.europa.eu

Acknowledgements: We gratefully acknowledge the expertise and contributions of Tommaso Raffaello, Antonio Fernandez Dumont, and Adrian Cesar Razquin in this external scientific report.

Suggested citation: Urbani D, Penzo M, Franco J, Evangelisti M, Parenti MD, Del Rio A, 2024. Refinement of the Risk Assessment Methodology for Open Reading Frames - Analysis in GMO Applications. 2024:EN-8561. 59 pp. doi:10.2903/sp.efsa.2024.EN-8561

ISSN: 2397-8325

Summary

The overall purpose of this project is to develop criteria for the definition and selection of ORFs relevant to risk assessment of GMOs (Objective 1) and to develop novel knowledge/methods for assessing likelihood of expression of relevant ORFs for the risk assessment of GMOs (Objective 2).

The project was developed to pursue three major tasks as explained below.

Task 1

A protocol was developed for a tailored search strategy to retrieve studies, including reviews and grey literature, pertinent to information on open reading frames (ORFs), such as definitions of ORFs and methods for assessing their likelihood of expression relevant to risk assessment. The extensive literature search was not only focused on GMOs for food-feed, import, and processing but also in areas unrelated to food safety, such as medicine. The query search allowed the retrieval of 15,484 documents, which were analysed first by reading titles and abstracts for their relevance, followed by a full-text analysis. This process resulted in the retention of 307 documents.

Task 2

The full-text reading of relevant documents allowed for the extraction of specific information, indicating the most pertinent criteria for the definition, prediction, and selection of ORFs relevant to the risk assessment of GMOs. In particular, the review of knowledge on protein expression relevant to the topic aimed to propose novel methods for assessing the likelihood of transcription and translation. The results showed that certain features of ORF nucleotide sequences affect the likelihood of gene expression, but the criteria underlying this likelihood require further research. Codon identity, nucleotide composition, and mRNA secondary structure are among the criteria that may be relevant for developing new methods for risk assessment (RA). However, the lack of structured data, diversity of application domains, and the reliability of criteria present major limitations for the development of intelligent systems to assess the likelihood of gene expression from ORF information. Furthermore, documents that explicitly consider ORFs in the context of food and feed, particularly for risk assessment, are scarce and do not address the problem of assessing the likelihood of protein expression. In addition, the reliability of criteria remains a challenge, as it is difficult to determine whether certain variables would constitute reliable input for new models or methods in the context of risk assessment. Lastly, it is pointed out that many of the prediction tools require the generation of de novo experimental data using different experimental techniques.

Task 3

Based on the information gathered in Task 2, the feasibility of using and integrating various tools was examined. The strengths and weaknesses of these tools were analysed in the context of the call, keeping into account not only the risk assessment of traditional transgenic products but also for the risk assessment of new products developed through modern genome editing techniques. It's noted that certain characteristics of ORF nucleotide sequences may be helpful in evaluating the likelihood of expression of relevant ORFs for GMOs risk assessment. However,

the criteria for this likelihood and the present limitations necessitate further investigation. Moreover, incorporating this into a unique tool is not achievable considering the diversity of existent tools, notably concerning the models and the underlying datasets, often related to some specific organisms. A conceptual workflow is proposed for navigating these challenges and limitations and is presented as an attempt to integrate and streamline the tools and methods currently available.

Table of contents

1. Introduction

1.1. Terms of reference as provided by the requestor

This contract was awarded by EFSA to a consortium with Innovamol Srl in the lead:

Contractor: Innovamol Srl

Members of the consortium are:

- Innovamol Srl, Modena, Italy
- Alma Mater Studiorum – University of Bologna, Bologna, Italy

Contract: OC/EFSA/GMO/2021/01

1.2. Background as provided by EFSA

The analysis of Open Reading Frames (ORFs) is a fundamental step in the food and feed risk assessment (RA) of genetically modified organisms (GMOs) carried out by EFSA and other risk assessment bodies around the world. ORFs are currently defined as any nucleotide sequence that contains a string of codons that is not interrupted by the presence of a stop codon in the same reading frame. According to current Regulation (EU) No 503/2013 (European Commission, 2013), all ORFs created as a result of genetic modification in plants shall be analysed using bioinformatics to predict possible similarities with known allergens or toxins.

The ORFs analysis, as performed by EFSA so far, follows the requirements laid down in EFSA guidance and other documents which were published before 2013 (EFSA Panel on Genetically Modified Organisms (GMO), 2011; European Commission, 2013; FAO, 2022). However, the knowledge on genomes and proteins has evolved by the use of ground-breaking technologies such as whole genome sequencing and omics, which together with advanced *in silico* prediction tools can better inform the risk assessment of ORFs.

Considering the scientific developments in the field of GMOs by the advances in biotechnology and genome editing applications, which can alter the genetic code of an organism without introducing foreign DNA, the definition and assessment of ORFs calls for a general revision. Such general revision of the definition and assessment of ORFs should be fully applicable not only for the risk assessment of traditional transgenic products but also for the risk assessment of new products generated by new genome editing techniques.

The assessment of ORFs generated by the genetic modification in GMOs, as part of the regulatory requirements for the molecular characterisation, is carried out *in silico* to inform food and feed risk assessment on the likelihood of peptides/proteins expression (intended and unintended) that may have similarities to known allergens and toxins. This overall approach, which is based on assumptions and general criteria defined more than 15 years ago does not consider neither the advances in the field of genetic engineering nor the availability of new *in silico*/bioinformatic tools.

The refinement in the analysis of ORFs risk assessment not only will improve the safety assessment of the GMOs applications currently in the EFSA risk assessment pipeline but will also

enhance their applicability to products developed via genome editing approaches which are likely to become more frequently applied in the near future.

1.3. Background as provided by the contractor

In molecular genetics, an open reading frame (ORF) is the part of a reading frame that has the ability to be translated, i.e. an ORF is a continuous stretch of codons that may begin with a start codon (usually AUG) and ends at a stop codon (usually UAA, UAG or UGA). The definitions of ORF have been recently reviewed by Sieber *et al.* (Sieber, Platzer and Schuster, 2018), in that ORFs differ as follows:

- Definition 1: an ORF is a sequence that has a length divisible by three and begins with a translation start codon (ATG) and ends at a stop codon.
- Definition 2: an ORF is a sequence that has a length divisible by three and is bounded by stop codons.
- Definition 3: an ORF is a sequence delimited by an acceptor and a donor splice site. Thus, it refers to a potentially translated eukaryotic internal exon. 5'- and 3'-terminal exons of a putative gene are determined at the end of the gene prediction process and are not considered for the actual ORF detection.

One common use of ORFs is as a piece of evidence to assist in gene prediction. Long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence while it should be noted that the presence of an ORF does not necessarily mean that the region is always translated. On the other hand, an increasing amount of evidence in the scientific literature indicates that several small ORF (sORF) embedded in the noncoding region of the genome seem to undergo a relatively stricter natural selection than adjacent sequences, raising the question of whether these sORFs have a capability to originate as a new gene in situ or to be integrated as a component into new genes elsewhere during evolution.

A further challenge in using ORFs to assist in gene prediction is represented by the recent advances in techniques of gene manipulation and gene engineering by which the genomes of living organisms may be modified. In fact, the artificial manipulation of gene expression may represent an important strategy for optimizing the economic traits of industrial organisms, livestock animals and crop plants. Although much previous research in plants has aimed to manipulate gene expression at the transcriptional level, the translation of mRNAs into proteins is a critical mechanism to control gene expression that provides a more immediate way to alter the cellular content of encoded proteins to maintain homeostasis (Zhang *et al.*, 2018; Si *et al.*, 2020).

Technological advances over the past decade have unravelled the remarkable complexity of RNA. The identification of small peptides encoded by long non-coding RNAs (lncRNAs) as well as regulatory functions mediated by non-coding regions of mRNAs have further complicated the understanding of the multifaceted functions of RNA. The original definition of lncRNAs concerns their low-/non- coding potential. However, accumulating evidence shows that lncRNAs have strong ribosomal associations in many species, varying from plant to animal, indicating a potential coding capacity in lncRNA sORFs. In recent years, several micro-peptides (miPEPs) derived from lncRNAs have been shown to be functional. Traditionally, RNAs could be divided

into two categories in accordance with their coding potential, that is, coding RNAs and noncoding RNAs (Fig. 1). Coding RNAs generally refers to mRNA that encodes protein to act as various components including enzymes, cell structures, and signal transductors (1, Fig.1). Noncoding RNAs act as cellular regulators without encoding proteins (3, Fig.1). However, it appears that the boundaries blur between coding RNA and noncoding RNA as some coding mRNAs can function without translating to protein via the formation of RNA secondary structure primarily derived from the untranslated region (UTR) and also from introns (2, Fig.1); finally, some lncRNAs can bind with ribosomes, and encode peptides to modulate cellular activities (4, Fig.1).

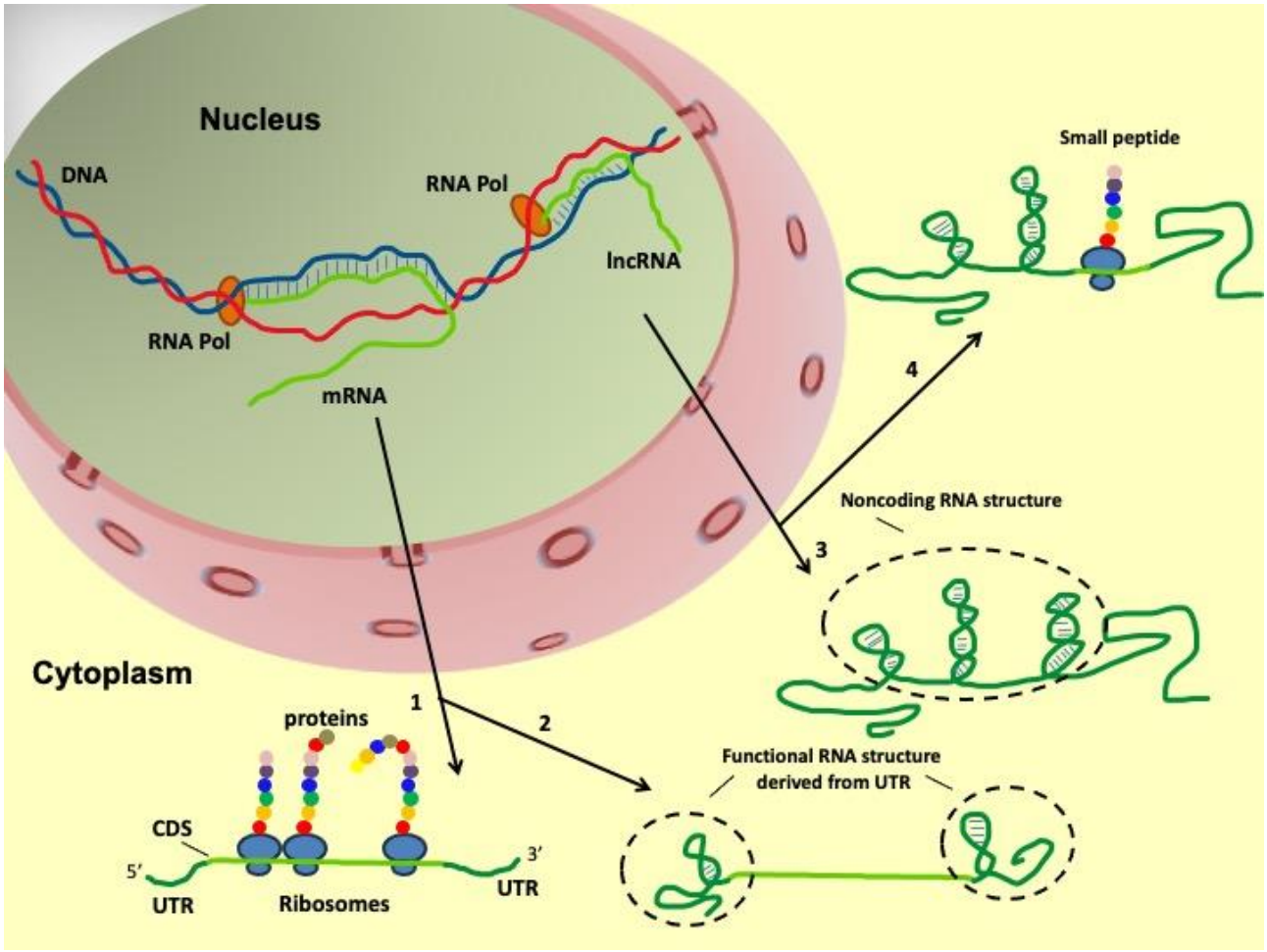

Figure 1: Coding RNAs and noncoding RNAs.

Over the past decade, several different pipelines, computational tools and environments as well as data resources have been developed to classify ncRNA coding potential by scoring conserved ORFs across diverse species. The scientific community has now a considerable choice of existing available software to suit their needs, platform preferences, and style. In addition, growing evidence illustrates the shortcomings on the current understanding of the full complexity of the proteome. Previously overlooked sORFs and their encoded microproteins have filled important gaps, exerting their function as biologically relevant regulators while the characterization of the full small proteome has potential applications in many fields. For instance, these principles have

been generalized with computational methods, by searching for homology using protein-domain databases, and by sequencing ncRNAs associated with polyribosomes as well as ribosome profiling data analysis whereby the detection of translated regions of a genome is a task for which ribosome profiling is particularly well suited. Continuous development of techniques and tools led to an improved ORFs discovery, where these can originate from bioinformatics analyses, from sequencing routines or proteomics approaches (Peeters and Menschaert, 2020; Cassidy *et al.*, 2021).

1.4. Objectives as provided by EFSA

In compliance with the tender specifications, the objectives of this procurement were:

- Objective 1: to develop criteria for the definition and selection of ORFs relevant to risk assessment of GMOs.
- Objective 2: to develop novel knowledge/methods for assessing likelihood of expression of relevant ORFs for the risk assessment of GMOs.

The activities were organized in three tasks: the first was to perform a wide-in-scope extensive literature search (ELS) on ORFs and existent methods useful for assessing their likelihood of expression relevant for risk assessment (Task 1). Second, these data were used to perform critical assessment of the results of ELS in order to develop the most relevant criteria for the definition, prediction and selection of ORFs relevant to risk assessment of GMOs as well as to review knowledge on protein expression and novel methods for assessing the likelihood of transcription and translation (Task 2). Finally, in Task 3 to develop in consultation with EFSA, novel methods for assessing the likelihood of expression of ORFs in the context of GMOs risk assessment that can be performed in an automatised manner - *in silico* tools. Such proposed novel methods should be fully applicable not only for the risk assessment of traditional transgenic products but also for the risk assessment of new products generated by the application of new genome editing techniques.

2. Data and Methodologies

2.1. Task 1: tailored search strategy to collect information on ORFs

2.1.1. Preparation of the literature search

The protocol for developing the tailored search included the definition of the following points:

- Definition of PICO/PECO. The definition of population (P), intervention/exposure (I/E), comparators (C) and outcomes (O) was essential to develop the eligibility criteria as well as inclusion and exclusion criteria at the beginning of the project.
- Definition of Boolean searches. All the database and data sources were able to support automated queries that were managed as specified below with query syntax suitable to obtain output files. Searches were performed with Boolean queries, in particular by using the Boolean operators "AND" and "OR" and parenthesis combinations involving keywords or chemical structures agreed with EFSA during the kick off meeting. The Boolean operators "NOT" was not used for queries in order to avoid automatic rejection of potentially relevant documents.

- Keywords and query syntax. The selection of keywords and the query syntax were customized and agreed with EFSA by keeping into account different block of area of interests (i.e. open reading frame(s), aspect of transcription, translation, species, expression of proteins, aspects related to the probability/likelihood of expression, *in silico* tools and methodologies used in the documents to identify ORFs protein expression).
- Definition of searches for grey literature.
- Definition of inclusion criteria.
- Definition of exclusion criteria.
- Definition of reporting methodology and summary tables.

All the above-mentioned points were agreed with EFSA before proceeding to actual searches.

2.1.2. Database

The database selected for this project was PubMed, as it was deemed more relevant for the objective of the call. Specifically, it was observed (refer to the Results section) that most journals reporting on ORFs pertain to the biomedical/genetic fields, which are comprehensively covered by PubMed. In contrast, the Web of Science (WoS) encompasses other fields (e.g., chemistry), where untargeted results were obtained. Regarding grey literature, direct access was granted to websites of various agencies and authorities both within and outside the EU. Due to the intrinsic differences in each website, tailored searches were implemented for each source. Details of the queries and the dates of the searches can be found in the Results section.

2.1.3. Software and IT tools

Raw results were collected by exporting references from PubMed in RIS format. Zotero V. 6.0.13 was used to manage references, including the creation of final RIS files for deliverables. InnoLiterature® database V.1.0, was used to merge data and perform the selection of relevance. Microsoft Word and Microsoft Excel were used to report data in text and table format.

2.1.4. Extensive literature search

The extensive literature search (ELS) was conducted following the protocol outlined previously. PubMed was systematically queried, and concurrently, a targeted search for grey literature from regulatory agencies and other authorities was performed, including EFSA, European Commission – Health & Consumer Protection Directorate-General, FSANZ (Food Standards Australia New Zealand), FDA (US Food and Drug Administration), US EPA (Environmental Protection Agency), Health Canada, China Food and Drug Administration as well as free text search on search engines that allowed to analyse possible hits from other sources. The documents retrieved from PubMed were compiled in reference management software, serving as a repository.

The repository of data was post-processed with the software functionalities to:

i. merge data from the different searches,
ii. remove duplicates,
iii. perform integrity check of each entry and,
iv. correct reference citations if needed.

Query syntax, date of performance of searches and hits obtained are reported in the result section below. The result of the post-processing returned a non-redundant and complete list of entries that were imported in the InnoLiterature® database where custom variables related to exclusion codes (EC) were added as described below.

2.1.5. Screening of titles and abstract

A screening of titles and abstracts for relevance to the risk assessment of GMOs was performed keeping into account inclusion and exclusion criteria. The experts of the team analysed references and full-text examination was preliminarily performed only in case of doubts or where a missing consensus about the relevance was present among the experts.

The work was organized in a way that two experts, plus one in case of conflicts, with different expertise[1] could review independently documents, especially in case of borderline documents. A decision on the relevance was made for each record/document if at least one expert judged it relevant as respect to the specific questions of the tender (*i.e.* inclusion and exclusion criteria). For each document, the following actions were performed:

- The relevance of each document was evaluated by checking if relevant keywords describe a real scientific relationship between ORFs and risk assessment.
- The relevance of each document will be evaluated against inclusion criteria.
- The relevance of each document will be evaluated against exclusion criteria.

In many cases, articles were deemed irrelevant and subsequently excluded without additional scrutiny. In other cases, the analysis of only the title and abstract was sufficient to judge the article as relevant. For these articles, documents were labelled as relevant and retained for further analysis as described below. At this stage, a full-text examination was performed solely in cases of uncertainty or when there was a lack of consensus about the relevance among experts. It is important to emphasise that a conservative approach was adopted for borderline documents to ensure that no potentially useful information was discarded.

2.2. Task 2: comprehensive literature search of relevant documents

Task 2 involved developing, in consultation with EFSA, the most pertinent criteria for the definition, prediction, and selection of ORFs relevant to the risk assessment of GMOs. To accomplish this, the full text of relevant documents was retrieved, critically evaluated, and data was extracted as outlined below. For records that passed the relevance screening based on titles and abstracts, or in cases when a final decision could not be made solely on the title and/or abstract, the full text of the document was obtained in pdf format. Subsequently, a critical assessment of the relevant articles was conducted by reading the full text, considering the following points:

1. The quantity of evidence. The team focused the discussion of references in terms of total number of papers screened in relation to the area of interest and keywords laid down in Task 1.

[1] Expert 1: expert in bioinformatics with at least 3 years of experience; Expert 2: expert in bioinformatics with at least 3 years of experience; Expert 3: expert in molecular biology with at least 3 years of experience

A quantification was reported in tabular form while the Discussion section describes the weight of the evidence gathered, in reference to the specific search terms.

2. The quality of the evidence. The discussion section included an assessment of the quality of the body of evidence in terms of GMOs data and/or *in silico* tools to specific search terms by considerations of study methodological quality, including advantages and drawbacks of methodologies used. In particular, the extent to which the quality of a body of evidence regarding GMOs data and/or *in silico* tools may be decreased was carefully evaluated by the reviewing team based on scientific judgements about study limitations for each main outcome. When studies provided widely different evidence (*e.g.* contradicting or unclear conclusions from different studies) the team sought explanations that were reported in the Discussion section.

3. Interpretation of the results. Chemical, biological, biotechnological and statistical significance of findings, especially for GMOs data and expression of relevant proteins was clearly extracted and explained in summarizing table, together with all assumptions made. In cases where very few relevant data were found, the characterization and reporting of the knowledge gaps was remarked as useful to support research recommendations.

4. Limitation of the reviewing process. Any limitation of the review process was reported and discussed, including amendments of the review protocol; *i.e.* No amendment of the reviewing protocol was adopted.

5. Agreements/disagreements. Agreements or disagreements with other studies or reviews were discussed and, similarly to the previous point 2), the team provided consensus reasoning and explanations in the Discussion section.

6. Complementary information. The reviewing team considered also complementing information that usually cannot be found in the scientific literature but, for instance, in commercial products (including, brochures, companies website, technical sheets, safety data sheets, marketing communications etc.) as well as analysis of existent commercial products currently used in agriculture, farming, agronomy, plant breeding and agrochemicals including information on transformation systems, characterization of the DNA inserted in the plant, inheritance and stability after transformation, protein characterization and expression, residue analytical methods etc. This information was referenced in the Discussion section whenever supported by sufficient body of evidence.

7. Further questions/hypotheses. In the view of improving EFSA's risk assessment process, a wider framework relating ORFs was considered so as secondary or complementary questions or hypotheses resulting from ELS and critical assessment were evidenced in the Discussion section.

Within Task 2, the team adopted a common annotation template to extract information in Excel that was agreed with EFSA. The parameters of the Excel template are depicted and explained in the list below.

Bibliographic information (text and URL)

- DOI: Digital object identifier of the document
- Title: Title of the document
- Abstract: Abstract of the document

Information class (checkbox)

- Definition: Checkbox labelling documents describing any criteria for the definition of ORFs
- Prediction: Checkbox labelling documents describing any prediction model or prediction tool related to ORFs in any domain, i.e. not restricted to risk assessment
- Selection: Checkbox labelling documents describing any information for the selection of criteria useful in the context of this call.

ORFs definition (textbox)

- ORF Start to Stop: Details of the start and stop sequences within the ORF
- ORF Stop to Stop: Information about consecutive stop sequences within the ORF
- Unconventional Start Codon: Details on non-traditional start codons used
- Unconventional Stop Codon: Information on non-traditional stop codons used
- Condition for Expression of ORFs: Factors influencing ORF expression (e.g., promoter, ribosome folding, codon usage, etc.)
- Organism: Organism in which the ORF is found or studied
- ORF vs. sORFs: Comparison or relationship between ORFs and short open reading frames (sORFs)

ORFs prediction (textbox)

- Method Used: Specific method or algorithm used for ORF prediction
- Input Data: Types and sources of data used as input for prediction
- Output Data: Results or outcomes of the prediction method
- Criteria and Settings: Standards and specific configurations used in the prediction process
- Applicability Domain: Context or scenarios where the prediction method is applicable
- Pros: Advantages or strengths of the prediction method
- Cons: Disadvantages or limitations of the prediction method

ORFs selection (textbox)

- Codon Identity: Specificity of codons within the ORF
- Choice and Bias Optimization: Methods for optimizing selection based on certain criteria
- mRNA Secondary Structure: Consideration of the mRNA secondary structure in the selection process
- Nucleotide Composition: Analysis of nucleotide make-up within the ORF
- Frequency/Infrequency of Codons: Assessment of codon occurrence within the ORF
- Presence of Inserts: Information regarding additional inserted sequences within the ORF
- Usage of Synthetic ORFs to Explore Critical Sequence Features: Techniques using synthetic ORFs for feature analysis
- Presence of Splice Site: Details about splicing within the ORF
- AU Composition within the 5' End of an ORF: Specifics of AU nucleotide composition at the 5' end
- RNA Post-Transcriptional Modifications: Information on modifications after RNA transcription

2.3. Task 3: novel methods for assessing the likelihood of expression of ORFs

Task 3 centred on the development, in collaboration with EFSA, of novel methods to assess the likelihood of expression of ORFs within the framework of GMOs' risk assessment.

To streamline this process, the data extraction table was streamlined to simplify data analysis and its subsequent reporting. Specifically, high-level information was manually organised by consolidating similar concepts, using a clear numbering system. This systematic rearrangement significantly enhanced the clarity and coherence of the data, further empowering the team to conceptualise a new model, as detailed below.

Following this data organization, a detailed analysis was conducted on the conditions and factors influencing ORF expression. This manual analysis derived information directly from the data table provided in Annex I.

In establishing these new methods, a rigorous examination of existing tools was essential. Each tool highlighted in the ELS of Task 1 and Task 2 was searched for on the web. Tools that no longer existed or were inaccessible were catalogued in Annex II. Conversely, tools available for download or through a web-based interface were tested using a designated test sequence. Each tool was evaluated based on user-friendliness, result turnaround time, easiness of inputting data and retrieving outputs, and an overall assessment of result reliability. Comprehensive findings are presented in the Annex II. However, this report mainly focuses on tools that directly align with the primary aim of the call, specifically, predicting the coding potential of ORFs.

During the evaluation of tools and based on ELS, it was observed that various methods to assess ORF expression have distinct strengths and weaknesses. The main limitations encountered to tackle the primary aim of the call were evidenced and discussed.

Based on the evaluations and analyses, and considering the current limitations evidenced, a conceptual framework was developed to establish the likelihood of ORF expression in risk assessment. This framework incorporates the idea that analytical tools for determining ORF likelihood of expression can be used provided that useful data set concerning the expression or non-expression of ORF will become available for training and validation.

3. Results

3.1. Task 1: Execution of the tailored search strategy to collect information on ORFs.

3.1.1. PICO/PECO definition

The definition of population (P), intervention/exposure (I/E), comparators (C) and outcomes (O) was essential to develop the eligibility criteria as well as inclusion and exclusion criteria at the beginning of the project. The following table describes the adopted definition.

Table 1: PICO and PECO definition.

	DESCRIPTION
P	Populations (P) will identify any living organism.
I/E	Interventions and exposure (I and E) will identify any intervention and/or exposure to which the P is exposed by means of ORFs involvement.
C	Comparators (C) will identify control or reference group in experimental studies or documents not exposed to I or E and information on regulatory documents. For the scope of this tender, we propose the use of common non-GMO products.
O	Outcomes (O) will identify any allergenic and/or immunogenic effects resulting from i/e in any different routes of exposure (e.g. inhalation, dermal, ingestion, intravenous and other parenteral administration routes)

3.1.2. Keyword and query syntax for PubMed

According to technical meetings held with EFSA, the process of defining the keywords essential for the ELS was meticulous and iterative. Recognizing the complexity of the subject, particularly the proliferation of bioinformatic tools developed in recent decades, the project team conducted several trial queries. These preliminary efforts were instrumental in highlighting the diverse landscape of existing literature and the challenges posed by the multifaceted nature of gene expression and computational tools.

After careful evaluation of these trial queries, the team devised a strategic query scheme where logically defined topics were embedded in searches in a way that aligned with the project's objectives, balancing the need to be comprehensive with the practicalities of making the project feasible.

The refined approach led to the development of the following three different queries:

- 1st query: documents related to aspects related to gene expression that have been published after 2011;
- 2nd query: identical to query 1 but limited to reviews published before 2010;
- 3rd query: documents related to bioinformatics, computational and *in silico* tool without years limitation.

The structure of searches is graphically depicted in the Figure 2 below.

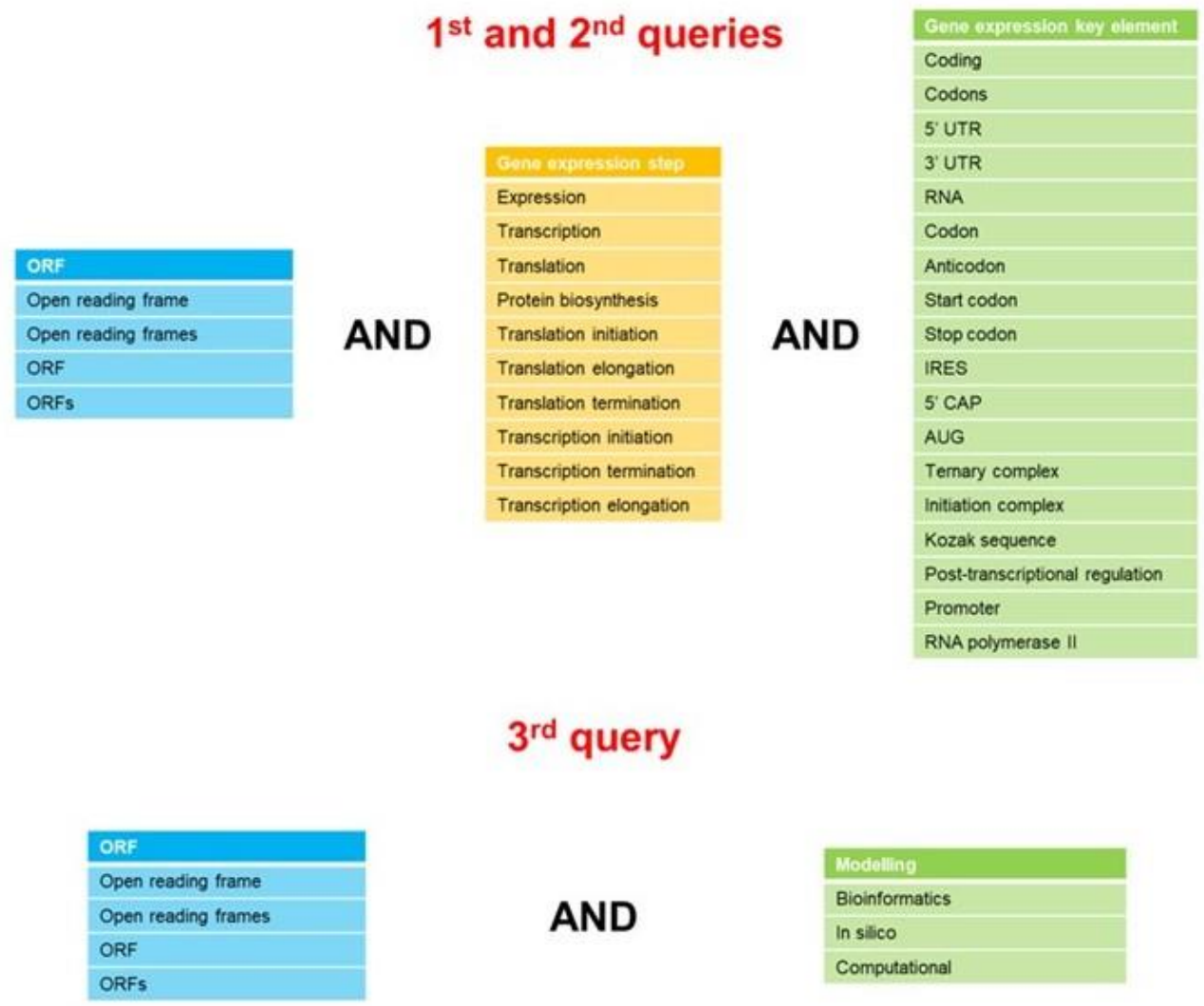

Figure 2: Graphical representation of query searches

The Table 2, as presented below, delineates the specific queries that were systematically formulated and applied to PubMed in order to retrieve the relevant documents. These queries embody the refined search strategy that was developed in collaboration with EFSA, capturing the core aspects of ORFs encompassing gene expression, bioinformatics, and computational tools necessary, consistently to tender specifications.

Table 2: Query searches in PubMed.

QUERY N°	PUBMED QURIES	YEARS COVERED

1	("open reading frame" OR "open reading frames" OR "ORF" OR "ORFs") AND ((expression OR transcription OR translation OR protein biosynthesis OR translation initiation OR translation elongation OR translation termination OR transcription termination OR transcription elongation) AND (coding OR codons OR 5′ UTR OR 3′ UTR OR RNA OR codon OR anticodon OR start codon OR stop codon OR IRES OR 5′ CAP OR AUG OR ternary complex OR initiation complex OR Kozak sequence OR post-transcriptional regulation OR promoter OR RNA polymerase II))	2011-2022
2	("open reading frame" OR "open reading frames" OR "ORF" OR "ORFs") AND ((expression OR transcription OR translation OR protein biosynthesis OR translation initiation OR translation elongation OR translation termination OR transcription termination OR transcription elongation) AND (coding OR codons OR 5′ UTR OR 3′ UTR OR RNA OR codon OR anticodon OR start codon OR stop codon OR IRES OR 5′ CAP OR AUG OR ternary complex OR initiation complex OR Kozak sequence OR post-transcriptional regulation OR promoter OR RNA polymerase II))	Before 2011 (reviews only)
3	("open reading frame" OR "open reading frames" OR "ORF" OR "ORFs") AND (bioinformatic OR *in silico* OR computational)	All

The Table 3 below presents the queries applied to the Web of Science (WoS) to retrieve documents for this study. This inclusion of WoS queries was undertaken as part of a pilot study to justify the choice of databases used. Despite the fact that only PubMed was selected as the primary document database due to its focus on relevant fields, a broader exploration was conducted with WoS to ensure that no pertinent information was overlooked. WoS, however, covers an extensive range of topics that often extend beyond the scope of this call. Consequently, additional verification was performed to ascertain the relevance of WoS documents, the details of which are further elucidated in the subsequent paragraphs.

Table 3: Web of Science queries.

QUERY N°	WEB OF SCIENCE QUERIES	YEARS COVERED
1	ALL=(("open reading frame" OR "open reading frames" OR "ORF" OR "ORFs") AND ((expression OR transcription OR translation OR protein biosynthesis OR translation initiation OR translation elongation OR translation termination OR transcription termination OR transcription elongation) AND (coding OR codons OR 5′ UTR OR 3′ UTR OR RNA OR codon OR anticodon OR start codon OR stop codon OR IRES OR 5′	2011-2022

	CAP OR AUG OR ternary complex OR initiation complex OR Kozak sequence OR post-transcriptional regulation OR promoter OR RNA polymerase II)))	
2	ALL=(("open reading frame" OR "open reading frames" OR "ORF" OR "ORFs") AND ((expression OR transcription OR translation OR protein biosynthesis OR translation initiation OR translation elongation OR translation termination OR transcription termination OR transcription elongation) AND (coding OR codons OR 5' UTR OR 3' UTR OR RNA OR codon OR anticodon OR start codon OR stop codon OR IRES OR 5' CAP OR AUG OR ternary complex OR initiation complex OR Kozak sequence OR post-transcriptional regulation OR promoter OR RNA polymerase II)))	Before 2011 (reviews only)
3	ALL=(("open reading frame" OR "open reading frames" OR "ORF" OR "ORFs") AND (bioinformatic OR *in silico* OR computational))	All

3.1.3. Other sources

Grey literature searches were performed by accessing websites of agencies and authorities in EU and outside EU and by searching guidance, regulations and scientific opinions, as detailed in the methodology section. Because of the intrinsic differences in each website, a tailored search was applied for each source. Relevant documents were searched by using the internal search engine of the website and different combinations of the search terms as described above. The table below depict the list of grey literature results.

Table 4: Identified grey literature documents.

DOCUMENT	ABSTRACT/RESUME	URL
FOOD STANDARDS AUSTRALIA NEW ZEALAND. RISK ASSESSMENT REPORT	Listex P100 bacteriophage preparation (hereafter referred to as the P100 preparation) is proposed for use on non-liquid ready-to-eat food products1 for the purpose of reducing numbers of Listeria monocytogenes. The types of food where the P100 preparation may be used include all ready-to-eat non-liquid food products. The Applicant proposes the P100 preparation would be used in combination with good hygienic practices (GHP) currently applied in food processing to control contamination of food with L. monocytogenes. It is intended to complement existing GHPs, not as a replacement for GHP. It is designed for use as a spray or dip for targeted application on food products and not as a surface disinfectant or general bactericide within the processing facility. The stated purpose and technological function of the P100 preparation may be consistently achieved when process validation has been undertaken for each food product, under	URL

	commercial conditions, and when the defined protocols are followed. FSANZ has assessed the safety and the proposed technological function of the P100 preparation. In doing so, the efficacy (ability to reduce L. monocytogenes on contaminated food) and the ongoing technological function (ability to continuously limit growth of L. monocytogenes) under proposed use has been assessed. FSANZ has concluded that the P100 preparation is safe, effective and has no ongoing technological function when used under commercial conditions in non-liquid ready-to-eat foods. To achieve continued efficacy and safety of use, it is important that detailed user instructions are provided and followed on the usage and disposal of the product. Treated products are not expected to re-enter the processing facility.	
EFSA STATEMENT. RISK ASSESSMENT OF NEW SEQUENCING INFORMATION ON GENETICALLY MODIFIED CARNATION FLO-40689-6	The GMO Panel has previously assessed genetically modified (GM) carnation FLO-40689-6 and concluded that there is no scientific reason to consider that the import, distribution and retailing in the EU of carnation FLO-40689-6 cut flowers for ornamental use will cause any adverse effects on human health or the environment. On 27 October 2017, the European Commission requested EFSA to analyse new nucleic acid sequencing data and updated bioinformatics data for carnation FLO-40689-6 and to indicate whether the conclusions of the GMO Panel on the previously assessed GM carnation FLO-40689-6 remain valid. The new sequencing data indicated the correction of one nucleotide compared to the sequencing data originally provided. The new sequence was corrected by removal of one nucleotide from the polylinker region in locus 1. The removal of this base pair reported in the new nucleic acid sequencing data for carnation FLO-40689-6 has been already present in the original plant material used for the risk assessment. Thus, with the exception of bioinformatics analyses, the studies performed for the risk assessment of GM carnation FLO-40689-6 remain valid. The new sequencing data and the bioinformatic analyses performed on the new sequence, did not give rise to safety issues. Therefore, EFSA concludes that the original risk assessment of carnation FLO-40689-6 remains valid.	URL
EUROPEAN COMMISSION. GUIDANCE DOCUMENT FOR THE RISK ASSESSMENT OF GENETICALLY	Prepared for the Scientific Steering Committee by The Joint Working Group on Novel Foods and GMOs. Composed of members of the Scientific Committees on Plants, Food and Animal Nutrition. Abstract: This document is for the use of risk assessors and notifiers1 who intend to apply for the commercial release of genetically modified plants and derived cultivars under existing	URL

MODIFIED PLANTS AND DERIVED FOOD AND FEED. 6-7 MARCH 2003 PLANTS	Community legislation (Directive 2001/18/EC [Ref. 1]) and/or for the commercial authorisation of genetically modified (GM) food or feed, i.e. food or feed containing, consisting of or produced from genetically modified plants (Regulation (EC)258/97 on Novel foods [Ref. 2]; Proposal for a Regulation on GM food and feed [Ref. 3&4]). This document does not cover genetically modified animals, or micro-organisms (including micro-organisms intended for use under containment conditions which are regulated by Directive 90/219/EEC [Ref. 5], as amended by Directive 98/81/EC [Ref. 6]), or medicinal products for human or animal use (which are regulated by Regulation 93/2309/EEC [Ref. 7]). The environmental assessment of GM plants used to produce medicinal products or other non-food products (e.g. cotton fibres, flowers) is covered in this document but additional guidance may be required, for example for long lived species such as trees. Issues such as containment or risk management are not within the scope of this document and thus the post-market monitoring of GM crops and derived food and feed is not addressed specifically.	
ORF HUNTER	ORFhunteR: an accurate approach for the automatic identification and annotation of open reading frames in human mRNA molecules. Abstract: The ORFhunteR package is a R and C++ library for an automatic identification and annotation of open reading frames (ORFs) in a large set of RNA molecules. It efficiently implements the machine learning model based on vectorization of nucleotide sequences and the random forest classification algorithm. The ORFhunteR package consists of a set of functions written in the R language in conjunction with C++. The efficiency of the package was confirmed by the examples of the analysis of RNA molecules from the NCBI RefSeq and Ensembl databases. The package can be used in basic and applied biomedical research related to the study of the transcriptome of normal as well as altered (for example, cancer) human cells.	URL
ANALYZE ORF	analyzeORF: Prediction of Transcript Open Reading Frame	URL

The limited number of cases sourced from the grey literature requires elucidation as there are reasons substantiating this outcome. Firstly, the vast majority of documents addressing ORFs are predominantly located within the scientific literature, making them more accessible through the ELS. This inherently reduces the potential volume of unique documents found exclusively in the grey literature. Secondly, when it comes to prediction tools associated with ORFs, they are typically published in peer-reviewed scientific journals rather than in grey literature. Lastly, the subject of ORFs in the context of risk assessment is infrequently addressed by regulatory documents. National and international authorities have, so far, dedicated limited specific attention to ORFs, further narrowing the potential pool of relevant grey literature. Taken

together, these factors effectively explain the limited number of grey literature sources related to the subject of interest.

3.1.4. Inclusion criteria

As a general rule, articles describing criteria for the definition and selection of ORFs relevant to risk assessment of GMOs and novel knowledge/methods for assessing likelihood of expression of relevant ORFs for the risk assessment of GMOs were considered relevant, even if not involved in the area of food/feed risk assessment.

3.1.5. Exclusion criteria

To allow accounting in an efficient and unambiguous manner the reasons for excluding a particular document, a scheme to encore exclusion criteria (EC) was devised and is depicted in Table 5. Exclusion codes represents a codification of the reasons explaining why a specific document needed to be excluded.

Table 5: Applicable exclusion codes.

EXCLUSION CODES	DESCRIPTION
EC1	Document with a computational analysis not related to ORF prediction or assessment of likelihood of expression.
EC2	Document describing ORF methods with no potential application to EFSA remits or the topic of the call.
EC3	Document with general description or historical obsolete description of ORFs.
EC4	Document with the functional study of a specific gene with information that cannot be generalized to the objectives of this contract.
EC5	Any other documents that cannot be categorized in inclusion criteria and cannot be excluded with exclusion codes. Retracted documents were assigned this EC.

In the context of this study, it should be highlighted that multiple exclusion codes can potentially be attributed to a single document due to overlapping definitions. However, for the precision and rigour required in this ELS, the identification of even a single exclusion code is sufficient to exclude a document from subsequent analyses.

3.1.6. Results of the extensive literature search

Table 6 and Table 7 shows the results of the query searches obtained with PubMed and WoS.

Table 6: Results of query searches with PubMed.

QUERY N°	YEAR	N° hits
#1	2011 - 2022	9584
#2	1983 - 2010	410 (reviews only)
#3	All	7381

Total with duplicates	17375
Grey literature hits	5
Total without duplicates and grey literature	15484

Table 7: Results of query searches with Web of Science.

QUERY N°	YEAR	N° hits
#1	2011 - 2022	7553
#2	1983 - 2010	334 (reviews only)
#3	All	2337
Total with duplicates		10224
Total without duplicates		9656

As expected, the results from the query searches in both PubMed and Web of Science databases indicate a significant volume of literature dedicated to topic of the call. Specifically, query #3, which lacked temporal constraints, yielded a substantial number of hits due to the swift progression of research and development in new bioinformatic and modelling tools concerning ORFs. The considerable number of documents sourced from the last decade (search #1) further underscores the growing significance and dynamism of research on ORFs. The increased emphasis on these areas is indicative of their relevance to current scientific research and advancements in the field.

It was noted that WoS resulted in 7533 documents. This is due to the fact that WoS covers a wide range of topics in life sciences and biomedical sciences but also topics that are not relevant for this call including engineering, social sciences, arts and humanities while PubMed covers only medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences, which are all relevant for this call. Despite this consideration, and in order to ensure that no relevant document was discarded from WoS-specific documents, we performed a manual inspection of 5% random documents obtained in WoS (377). The result of this manual inspection showed that WoS did not provide any additional relevant document. Therefore, the query search for the ELS was considered only on the basis of PubMed results (Table 6).

3.1.7. Screening of relevance by title and abstract

The screening of relevance was performed using the methodology described above. All documents were processed, and it was found that 2.0% of them were relevant, corresponding to 307 documents.

3.2. Task 2. Comprehensive literature search of relevant documents

3.2.1. Full-text analysis and data extraction

Reading the full text of pertinent documents enabled the extraction of specific details, highlighting the most salient criteria for the definition, prediction, and selection of ORFs relevant to the risk assessment of GMOs. Specifically, the existing knowledge on protein expression, pertinent to the subject, was examined to suggest innovative methods for evaluating the

likelihood of transcription and translation. The relevant content was systematically categorised as presented in the Excel file of Annex I, consolidating the information into the following sections:

- Bibliographic information (columns A-C)
- Document classification (columns D-G)
- High-level information (columns H-AD)

The Bibliographic Information section includes the title, abstract, and DOI of the identified pertinent documents.

Considering the diverse nature of the documents, with regard to their topics and applicability domains, a qualitative method was employed within the document classification section to categorise them in relation to the tender specifications. In particular, the column D provides a numerical classification of importance as per the following scheme:

1. Highly relevant documents: These sources provide the most accurate insight into the potential for uncovering novel methods to define, predict, and select ORF criteria for risk assessment applications.
2. Moderately relevant documents: While these documents provide information on evaluating new methodological criteria, they necessitate integration with other tools or methodologies due to their inherent limitations.
3. Limited relevance documents: Although these documents contain some pertinent information, other sources align more closely with the primary objective.

Columns E-G include checkboxes, facilitating easy filtering of documents in Excel based on their classification concerning definition, prediction, and selection criteria.

The High-level Information section has been structured as illustrated in the Excel file, derived from the details specified in the methodology section. Only columns of high-level information deemed pertinent to definition, prediction, and selection were incorporated. It is anticipated that in Task 3, this high-level information section (columns H-AD) was further refined to better consolidate similar concepts within the same document, as delineated subsequently.

The following sections provide an encompassing overview of the extant information concerning the definition, prediction, and selection of ORFs, as defined in the tender specifications and as resulting from ELS. A more detailed discourse on their applicability and relevance in relation to further tools and methods is elaborated upon in the results of Task 3.

3.2.2. Enhanced organization of the summary table

Due to the complexity of the topics sourced from the ELS, we structured the summary table (Annex I) to better organize concepts and enhance comprehension of the main methodologies, tools, and their respective pros and cons. As evident in Annex I, the information from columns H-AD was organized using a numbering system: 1), 2), 3), etc. With this notation, each number within each document denotes similar concepts or rationales found. The organization of data was then effectively used to represent the information described in the following sections.

3.2.3. Discussion of existent information relating to the definition

An ORF, as related to genomics, can be defined as a portion of a DNA sequence that has a length divisible by three and begins with a translation start codon (ATG) and ends at a stop codon or, as a sequence that has a length divisible by three and is bounded by stop codons or, as a sequence delimited by an acceptor and a donor splice site thus referring to potentially translated eukaryotic internal exon (Sieber, Platzer and Schuster, 2018). The term "open reading frame" can be misleading because what is being read is the RNA code, *i.e.* read by the ribosomes in order to make a protein, while the term "open" indicates that sequence is open for ribosomes to keep reading the RNA code and add other amino acids to the polypeptide sequence. A stop codon ends the ORF so that the longer an ORF, the more likely is to be part of a gene which is coding for a protein. One of the main steps in gene-finding is determining which ORFs encode for a protein, and which ones occur by chance alone. Over the last 30 years, three fundamental methods were adopted for identifying ORFs in genomic sequences:

(1) sequencing randomly selected cDNA clones and aligning the sequences to their genomic sources;

(2) finding ORFs that could produce proteins similar to proteins that are already in databases; and

(3) finding ORFs *de novo*, without reference to cDNA sequences or their conceptual translations.

In more recent years, it was found that specific workflows can identify noncoding transcripts that can potentially translate intronic, intergenic and several other classes of ORFs (Fickett, 1994; Erady, Puntambekar and Prabakaran, 2020). A pipeline focused on genes coding for transcription factors was shown to increase isoform detection by an order of magnitude when compared to unenriched samples. Here, an isoform refers to the different versions of a protein that can be produced from a single gene due to alternative splicing of the mRNA transcript and enriched samples refer to samples coding for transcription factors, to facilitate their detection and analysis. Thus, this study suggested that it is possible to also identify ORFs from transcripts (Sheynkman *et al.*, 2020). Mounting evidence showed also that computational, genomic, and proteomic approaches could be used to allow faster detection and characterization of ORFs, including small ORFs (sORFs) and this is of particular relevance considering that sORFs have the translational potential to produce peptides, thus playing essential roles in various biological processes (Ma *et al.*, 2016; Wang *et al.*, 2021). It was also found that ORFs phylogenetic analysis can correctly identify the evolutionary relationships between members of Norovirus, suggesting that phylogenetics can be a valuable way to identify ORFs (Hung and Lin, 2013). Interestingly, small proteins encoded by ORFs shorter than 50 codons, i.e. sORFs, are often overlooked by annotation engines and are difficult to characterize using traditional biochemical techniques. In this context, experimental techniques such as ribosome profiling have been highlighted to have a potential to empirically improve the annotations of genomes (Vazquez-Laslop *et al.*, 2022). Notably, the primary focus of this call introduces an added layer of intricacy. It centres on discerning whether the probability of a particular gene's expression is associated with a specific ORF. More specifically, this focus aims to establish the criteria that link the gene to the ORF, enabling more accurate and reliable predictions regarding gene expression and ORF association.

The expression of ORFs is known to be influenced by a variety of factors, including promoter activity, mRNA folding, codon usage, and the presence of upstream ORFs, amongst others. These factors can significantly impact the likelihood of ORF expression (Kozak, 1996; Tomita, Shimizu

and Brutlag, 1996; Couso and Patraquim, 2017; Ong *et al.*, 2022; Sinha *et al.*, 2022). For instance, an ORF that is highly likely to be expressed due to favourable conditions such as an active promoter, optimal codon usage, and absence of inhibitory upstream ORFs (uORFs), might be more readily detected and accounted for in predictive models. Conversely, an ORF with a low likelihood of expression, perhaps due to a weak promoter, suboptimal codon usage, or the presence of inhibitory uORFs, might be less likely to be detected, potentially leading to underestimations in predictive models. Therefore, understanding these factors and how they influence ORF expression is crucial for improving the accuracy of predictions, irrespective of the potential functional or pathological effects of the protein that the ORF may encode. In the subsequent sections, a discussion of the most salient points relating to ORF definition and conditions and factors affecting ORF expression is provided.

3.2.3.1. ORF start to stop

The identification and characterization of ORFs, specifically uORFs and sORFs, represent a rapidly evolving area in molecular biology and genomics. ORFs are potentially translatable sequences beginning with a start codon and concluding with a stop codon. uORFs, a subset of these sequences, have traditionally been understood as translational repressors. They begin with a start codon, undergo translation, and ultimately terminate with a stop codon, leading to ribosome dissociation. This can inhibit downstream protein-coding region translation and may activate other pathways, including nonsense-mediated decay (NMD). NMD is a cellular mechanism that identifies and degrades mRNA molecules containing premature stop codons, preventing the translation of potentially harmful truncated proteins. In fact, the mRNA molecule contains features that influence the translation of its protein-coding regions, one of which is uORF. However, uORFs are not limited to repressive roles. Recent studies have shown that these sequences can influence protein synthesis in diverse ways. For instance, an uORF's translation can affect the ribosome's ability to locate the start codon of the primary ORF. Additionally, peptides encoded by uORFs can have regulatory effects (Chugunova *et al.*, 2018). Furthermore, uORFs are prevalent in the genome. Approximately 40% of mammalian 5' untranslated regions (UTRs) harbour uORFs. Mutations within these uORFs can have significant consequences, leading to genetic disorders and diseases. Notably, some peptides encoded by the 5'-UTRs have functions beyond regulating translation, serving as functional molecules within the cell (Chugunova *et al.*, 2018).

As computational and experimental capabilities advance, particularly with techniques such as ribosome profiling, the roles of uORFs became clearer (Spealman, Naik and McManus, 2021). A prime challenge in identifying uORFs, especially using ribosome profiling data, is the procedure's noise, the addition of potential translation initiation sites when including non-canonical start codons, and the scarcity of molecularly validated uORFs. To address this, machine learning tools such as uORF-seqr were introduced, combining ribosome profiling with RNA-seq data and transcript-aware genome annotations to detect statistically significant AUG and near-cognate codon uORFs (Spealman, Naik and McManus, 2021).

On the other hand, long noncoding RNAs (lncRNAs) resemble mRNAs, and it has been a matter of debate to what extent these transcripts code for proteins. Although many of these RNAs are considered noncoding, a significant portion may undergo translation. However, the resulting peptides from these translations are often unstable or non-functional (Housman and Ulitsky, 2016). Nevertheless, the development of the translation initiation sequencing (TI-seq) has

allowed researchers to map translation initiations globally, revealing complex translational patterns. The toolkit, Ribo-TISH, has been instrumental in detecting and comparing translation initiation from TI-seq data, revealing novel ORFs in several regions including lncRNAs (Zhang *et al.*, 2017). Nevertheless, the identification of genuine lncRNAs remains a challenge. Current methods can only identify a fraction of full-length protein-coding transcripts in humans. This limitation often leads to the misclassification of protein-coding transcripts as lncRNAs. To address this, tools like lncScore have been developed. lncScore, in particular, can differentiate lncRNAs from mRNAs, even those that are partial-length (Zhao, Song and Wang, 2016).

The importance of predicting ORFs using stop codon frequencies has been stressed, especially in relation to the GC content (Pohl, Thei\betaen and Schuster, 2012). For instance, the work by Pohl *et al.* elucidates that in non-coding DNA, a stop codon sequence is predicted approximately in every 21st trinucleotide. Conversely, the frequency of stop codons in coding sequences deviates from this rate. This deviation is due to the longer median length of protein-coding ORFs in bacteria and eukaryotes. The stop codon frequency in these coding sequences, within the relevant reading frame, significantly diverges from the background frequency of the corresponding trinucleotides. This distinction is crucial for detecting coding ORFs and the relevant reading frames. Traditional methods of gene prediction based on stop codon frequencies were built upon the assumption of a 50% GC content. However, the research presented a method that can describe the influence of varying GC content on determining the threshold lengths of potential coding ORFs. In addition, the study highlights that the utility of ORF prediction based on stop codon frequency is considerably efficient in genomes with low GC content, such as *Rickettsia prowazekii*. For the purpose of their study, an ORF is a sequence stretch divisible by three, starting with the codon 5′ -AUG-3′ and ending with one of the stop codons (5′ -UAG-3′ , 5′ -UGA-3′ , or 5′ -UAA-3′), with no internal stop codons.

The study of uORFs has also expanded beyond the traditional AUG start codon. Ribosome profiling has identified small open reading frames (sORFs) that begin with non-AUG start codons. These non-canonical start codons are found in both prokaryotic and eukaryotic genomes. Their presence in sORFs, compared to main ORFs, suggests potential roles in controlling expression levels and responding to specific cellular conditions (Cao and Slavoff, 2020)

In another study (Suenaga *et al.*, 2022), the ORF dominance score was developed to differentiate between coding and non-coding RNAs. This score correlates with translation efficiency and is specifically defined as the ratio of the longest ORF to the total length of all putative ORFs. The study points out that ORFs initiate at AUG and terminate at any of the three established stop codons, UAA, UAG, or UGA, in a 5' to 3' direction within an RNA sequence. Sequences commencing at AUG and finishing at the 3'-terminus of RNA without the aforementioned stop codons aren't considered as ORFs. Interestingly, a novel approach to gene prediction in metagenomics using a convolutional neural network (CNN) was presented by Al-Aijlan *et al* (Al-Ajlan and El Allali, 2019). CNN-MGP, the program developed in this study, predicts genes from raw DNA sequences, eliminating the need for manual annotation. ORFs, in this context, are sequences starting with a start codon (ATG, CTG, GTG, or TTG) followed by multiple codons and ending with one of the stop codons (TAG, TAA, or TGA). ORFs that are missing either or both start and stop codons are labelled as incomplete.

Lastly, the article (Yang *et al.*, 2023) by Yang *et al.* highlights the role of uORFs in the regulation of autophagy-related protein translation where uORFs are identified as brief coding sequences with start codons located in the 5' untranslated region (5' UTR) of eukaryotic mRNAs. The study emphasizes that the translation events that initiate at uORFs might either terminate before the main coding sequence (CDS) or might partially overlap with the CDS, contingent on the position of the associated stop codon. In related research, a study by Xiang et al. (Xiang *et al.*, 2023) elaborates on the regulation of start-codon selection through pervasive downstream RNA hairpins. It underscores how these structures, found downstream of uAUGs, dynamically dictate the initiation of translation, a crucial mechanism for organisms to adapt to changing conditions. This additional layer of regulation enhances the understanding of uORFs' impact on translational control, offering a broader perspective on the interplay between uORFs and the coding sequences they precede.

To summarise, the identification of start and stop codons, as well as the detection of upstream open reading frames (uORFs), play a significant role in delineating the putative coding region of genes. This is of greatest significance in molecular biology and genetics as it allows for the differentiation between coding and non-coding sequences. The significance of these codons lies in their crucial role in ensuring precise gene identification and annotation within genomic sequences. In addition to delineating coding regions, various factors including the surrounding elements of codons, secondary structures of mRNA, and codon usage exert an influence on the processes of translation initiation, elongation, and termination. The comprehensive examination of mRNA context, encompassing uORFs and regulatory motifs, enhances the comprehension of the start-to-stop context, emphasising its importance in fundamental biological processes and potential new approaches for GMO risk assessment.

3.2.3.2. ORF stop to stop

The current predominant understanding of ORFs is largely defined by the start to stop criterion, as this is the approach most textbooks advocate, as seen in the previous point. Historically, this can be attributed to the first fully sequenced genomes being primarily prokaryotic, excluding viruses. Given the simpler gene structure in prokaryotes due to the absence of splicing, this definition became more widespread in academic literature and teachings (Sieber, Platzer and Schuster, 2018). However, when delving deeper into the stop-to-stop definition, its utilization appears to be comparatively sparse. Notably, while some guidance documents propose the stop-to-stop criterion, its practical relevance can be questioned. This is because ORFs (defined from stop to stop) longer than 225 bp are expected to appear randomly every kb on a single DNA strand. Therefore, using ORF size as a sole criterion might not be effective in accurately pinpointing protein-coding regions. Such challenges with short and sparse vertebrate coding regions have spurred the advent of innovative statistical methods, aiming to better estimate the coding potential of diverse genomic subsequence (Claverie, 1997; Claverie, Poirot and Lopez, 1997). Conclusively, while the nuances between these definitions and their implications on the risk assessment of GMOs have been discussed, there remains a need for further research to holistically understand the role or significance of the stop-to-stop definition in relation to ORF delineation and its consequent impact on GMOs risk assessment.

3.2.3.3. Unconventional start codon

Unconventional start codon definitions refer to non-standard codons that are used to initiate protein synthesis. These codons are not recognized by standard translation machinery and require specialized ribosomes or modification of the translation process. Examples of unconventional start codons include those found in some bacteria, archaea, and viruses. In the context of defining ORFs, unconventional start codons can dramatically reshape the understanding of protein synthesis. In fact, contrary to the widely held belief that eukaryotic mRNAs typically contain a single translation start site encoding a unique functional protein product, emerging research suggests a more intricate landscape. Eukaryotic ribosomes can recognize multiple alternative translation start sites, diversifying the range of encoded proteins from a single mRNA sequence (Kochetov, 2008). This ability to initiate translation from alternative sites is no longer limited to the conventional AUG start codon. Sometimes, non-AUG codons are utilized, though their efficient use typically demands additional signals, such as a perfect context or specifically positioned downstream secondary structures. Factors such as the surrounding nucleotide context can influence the efficiency of these non-standard initiation events (Kochetov, 2008). The term 're-initiation' refers to the ability of the translation machinery to begin translating a new ORF downstream of an already translated uORF. This process is intriguing as it is influenced by the length of the uORF. For instance, effective re-initiation of translation was observed following a uORF consisting of 10 to 12 codons. However, this re-initiation efficiency diminished when the uORF length surpassed this range (Kochetov, 2008). The ability to re-initiate translation can significantly broaden the array of functional proteins encoded by a single mRNA sequence, thus enriching the functional diversity encapsulated within mRNAs.

An additional layer of complexity arises when considering non-canonical start codons. Ribosome profiling has unveiled the presence of numerous novel coding sequences termed small sORFs. Interestingly, these sORFs often initiate at non-AUG codons, challenging traditional assumptions about gene annotations (Cao and Slavoff, 2020). The frequency of such unconventional translation initiation varies across organisms. For instance, in bacteria, initiation at GUG, UUG, CUG, and AUU codons has been observed. Similarly, mammalian cells display a remarkable abundance of non-AUG initiation events, with CUG emerging as a dominant near-cognate initiation codon (Cao and Slavoff, 2020).

It should be noted that, nowadays, machine learning is aiding the discovery of these unconventional translation events. For example, a novel algorithm, uORF-seqr (see also 3.2.3.1), has been developed to identify statistically significant AUG and near-cognate codon uORFs in yeast, leveraging ribosome profiling data (Spealman, Naik and McManus, 2021). Near-cognate codons (NCCs) stand out as particularly interesting because their usage can surge under stress conditions. Such changes in NCC utilization might stem from stress-induced modifications of translation initiation factors, underscoring the dynamic nature of translation under varying environmental contexts (Spealman, Naik and McManus, 2021).

In conclusion, the classical view of translation initiation in eukaryotic mRNAs is evolving. The discovery of alternative translation start sites, both within and outside of the canonical AUG codon, signifies an expanded coding potential, offering a reservoir of functional diversity. Exploring these mechanisms can furnish invaluable insights into gene expression regulation,

evolutionary constraints, and the complex world of protein isoforms. Such understanding is crucial, especially when predicting gene expression in contexts like GMOs risk assessments.

3.2.3.4. Unconventional stop codon and stop codon

Upon examination of the existing literature, no specific information regarding the role of unconventional stop codons in the context of defining ORFs was found. Conventional stop codons (UAA, UAG, and UGA) are well-established as the primary signals for translation termination. However, the existing body of literature does not provide information regarding the existence or impact of unconventional stop codons on ORF definition. Given the lack of information on unconventional stop codons, it is difficult to assess their relevance in the context of GMOs risk assessment for assessing the likelihood of expression of ORFs.

3.2.3.5. Condition for expression of ORFs

In the existing literature, the conditions for expression of ORFs is a crucial information. A plethora of factors can influence the expression of ORFs, including promoter sequences, transcription factors, regulatory elements, and environmental conditions. Promoter sequences initiate transcription, while transcription factors and regulatory elements contribute to the modulation of gene expression. Environmental conditions, such as temperature, nutrient availability, and stress, can impact the expression of ORFs by modulating cellular signalling pathways and gene regulation mechanisms. Post-transcriptional modifications, codon usage bias, and ribosome binding sites are other factors that can influence translation efficiency and, as a result, ORFs expression. The stability, folding, and accessibility of the mRNA molecule are influenced by these elements, affecting the rate at which ribosomes can bind and initiate translation. Furthermore, the discovery of alternative splicing or overlapping ORFs can compound the intricacy of gene regulation, as these events can generate multiple transcript isoforms or protein products from a single genetic locus.

A recent shift in understanding reveals that contrary to the previous belief that eukaryotic mRNAs typically contain a single translation start site, they can have multiple alternative translation start sites. This insight suggests an even more intricate regulatory system. The number of experimentally verified examples of alternative translation is growing, and their frequent occurrence, supported by computational evaluations, points to their functional significance in the broader eukaryotic proteome (Kochetov, 2008). The nucleotide context surrounding the start codon is crucial in influencing translation initiation. For instance, in mammalian translation systems, the consensus sequence GCCRCCAUGG surrounding the start AUG codon corresponds to what is known as the "perfect context" — a sequence that all ribosomes recognize as a Translation Initiation Site (TIS). The importance of the positions near the start codon varies among organisms. In plants, for example, maize and tobacco cells have shown that GCCAUGGC and RAAAUGGC are the most efficient TISs. On the other hand, in yeast cells, AUG recognition efficiency can vary significantly based on the nucleotide at position -3 (Kochetov, 2008). Further intricacies, like the presence of a stable secondary structure located 13–17 nucleotides downstream of the start codon, can delay ribosome movement, potentially facilitating the recognition of TIS in weaker contexts.

Kozak analysed (Kozak, 1996) the complexities of interpreting cDNA sequences. In most vertebrate mRNA 5' noncoding sequences possess an elevated G + C content, leading to extensive base-pairing that could significantly lower translational efficiency. In fact, many

vertebrate mRNAs seem to have 5' noncoding sequences that curtail translation. Another interesting revelation is the potential pitfalls in relying solely on in vitro translation systems for the expression of cDNAs. These systems may sometimes initiate translation from an internal ATG codon, resulting in discrepancies in the derived polypeptide. Emphasis is also placed on the consideration of Mg^{2+} concentrations, as they can influence the selection of the ATG codon utilized. Additionally, the GC-rich leader sequences in vertebrate mRNAs can inhibit translation in reticulocyte lysates, offering an avenue to boost protein synthesis in vitro by substituting the GC-rich 5' UTR with a less structured leader sequence. The study concludes with a recommendation for rigorous testing to ensure the accurate interpretation and translation of cDNA sequences.

Interestingly, in a comprehensive computer analysis of the GenBank database by Tomita *et al.*, a correlation between splicing sites and their positions in reading frames was examined (Tomita, Shimizu and Brutlag, 1996). Investigating splicing site locations across various eukaryote taxonomic groups, the study found introns predominantly interrupt reading frames at codon boundaries rather than within codons. This pattern supports the exon shuffling hypothesis, suggesting that exons ending at codon boundaries can seamlessly concatenate without causing frame shifts, offering an evolutionary advantage. Conversely, when introns do interrupt within codons, they are more frequently located between the second and third bases of the codon, although the rationale remains unclear. Additionally, exons and pairs of adjacent exons exhibited lengths that are multiples of 3 more often than expected. These observations, which echo findings by Long, Rosenberg, and Gilbert (Long, Rosenberg and Gilbert, 1995), underscore the potential evolutionary preference for introns to be positioned in ways that preserve the integrity of coding sequences, further emphasizing the evolutionary implications of exon shuffling.

With the progression of biotechnological methodologies, the understanding and capability to manipulate ORFs is constantly evolving through new genome editing techniques. For instance, Zhang *et al.* demonstrated that by employing genome editing techniques on endogenous uORFs in plants, the translation of mRNA from specific uORFs, pivotal to either development or antioxidant biosynthesis, can be modulated. Notably, editing the uORF of a gene vital for vitamin C biosynthesis in lettuce not only heightened oxidation stress tolerance but also augmented ascorbate levels by approximately 150% (Zhang *et al.*, 2018). In a complementary study, Si *et al.* provided a detailed protocol, using CRISPR-Cas9, for the nuanced adjustment of gene translation in plants by targeting endogenous uORFs. Their methodology offers an efficient way to generate transgene-free uORF mutants, furthering the refinement in gene function analysis and crop trait enhancement (Si *et al.*, 2020). These insights are important, especially considering the broader applications of genome editing in GMOs risk assessment.

Additionally, the intricate web of proteomic diversity is further complicated by the discovery of circular RNAs (circRNAs), primarily viewed as non-coding entities. However, recent findings suggest a potential translational capability in animal circRNAs. This proposition gains momentum when considering plant-pathogenic circRNAs like viroids, which, despite their long-standing status as non-coding agents, have now been identified to possess ORFs potentially encoding peptides. Such revelations underscore the transformative potential of circRNAs, hinting at their ability to serve as non-canonical translatable transcripts under specific cellular conditions (Marquez-Molins *et al.*, 2021). More recently, in another study by Sinha et al, the potential of circular circRNAs to encode proteins was explored (Sinha *et al.*, 2022). Unlike typical mRNAs,

circRNAs are covalently closed noncoding RNA molecules, formed through a unique process termed back-splicing. Despite lacking the conventional 5' and 3' ends, certain circRNAs exhibit evidence of translational capacity. Key markers for this include the presence of ORFs, internal ribosomal entry sites (IRES), and N6-methyladenosine (m6A) modifications. While eukaryotic mRNA translation usually necessitates a 5' cap and 3' poly-A tail, circRNAs utilize cap-independent translation initiation mechanisms. Specifically, IRES sequences can initiate translation without the 5' cap, while the m6A modification can recruit specific proteins to begin the translation process. Recent research has suggested that a significant subset of circRNAs, especially those associated with polyribosomes, have the potential for translation, pointing to a broader coding potential than previously believed.

It is important to note that, considering these numerous factors, it becomes vital to account for the conditions for expression of ORFs when evaluating their expression in the context of GMOs risk assessment. However, the inherent complexity, compounded by limited data and multiple variables, poses challenges for the development of predictive algorithms and comprehensive models. Despite these challenges, it's essential to incorporate as many of these variables as possible.

3.2.3.6. Organisms

The diversity and intricacy of organisms play a pivotal role in the identification and understanding of ORFs, as unique genetic elements, regulatory mechanisms, and evolutionary narratives intrinsic to different organisms can greatly influence ORF identification and functionality. While the genetic code is generally consistent across life forms, certain organisms or organelles exhibit variations, such as alternative start and stop codons or differing patterns of codon usage bias. Such variations can hinder the precise prediction of ORFs and their resulting protein products. Recent research has shed light on the emerging significance of sORFs found across all genomes. Historically overlooked due to challenges in confirming their translational status, advancements in computational biology, proteomics, and high-throughput analyses have unveiled potentially coding sORFs in numerous organisms. Particularly, these sORFs encode functional peptides whose cellular roles are yet to be fully grasped (Andrews and Rothnagel, 2014). uORFs have been discerned as modulators of ribosome access to the subsequent coding sequence, affecting its translation in various ways. Despite being short and not sharing the reading frame of the downstream coding sequence, they are widespread in eukaryotes, from vertebrates to fungi.

Moreover, sORFs of 100 codons or fewer, commonly omitted from proteome annotations, are proving to be more essential than previously believed. For instance, the *Drosophila melanogaster* transcriptome encompasses myriad actively translated sORFs, producing peptides with still-unknown functionalities. These sORFs exist in various functional capacities, from inert DNA sequences to transcribed and translated cis-regulators and peptides with the potential to regulate membrane-associated proteins (Couso and Patraquim, 2017). The presence of such sORFs in model organisms like flies, mice, and humans, can offer invaluable insights into peptide biology pertaining to development, physiology, and human diseases.

Unique regulatory elements, such as organism-specific promoter sequences and transcription factors, further modulate gene expression, with potential repercussions for ORF expression specific to certain taxa. These discrepancies, along with concerns like horizontal gene transfer (HGT) events that can facilitate genetic material transfer between unrelated species, emphasize

the importance of understanding the phylogenetic context of ORFs (Kochetov, 2008). Yet, despite the expanding literature, a comprehensive understanding of how these factors interplay to influence ORF functionality in diverse organisms remains a challenge for assessing GMO risks accurately.

Consequently, the variations in ORF identification and prediction, influenced by the selected organism, highlight the need for a meticulous approach in this field. Understanding the intricacies specific to each organism is essential, as they can significantly impact the accuracy of ORF expression predictions. It is vital to have specific datasets for risk assessment to ensure robust and precise model predictions.

3.2.3.7. ORF size

The size of an ORF plays a pivotal role in influencing various aspects of protein translation, stability, and function. One of the significant implications of ORF size pertains to its prediction of likelihood of expression (Andrews and Rothnagel, 2014). For instance, translation efficiency tends to vary depending on the size of an ORF. The translation of larger ORFs generally requires more resources, making it a less efficient process compared to the translation of smaller ORFs. As a result, one might expect the expression levels of larger ORFs to be lower than that of their smaller counterparts. When it comes to the stability of the product translated from an ORF, it is noteworthy that the size can be a determinant. Generally, larger proteins possess a higher degree of stability than the smaller ones, which might influence their role and impact at the functional level. Further intricacies related to ORF size are its regulatory elements. Larger ORFs might encompass internal regulatory elements like internal ribosome entry sites (IRESs) or uORFs. These elements can potentially impact the translation of the ORF. Additionally, the susceptibility to misfolding and aggregation increases with the size of the protein, which could be detrimental, inducing cellular stress and leading to diseases.

Another complexity arising from the size is the functional aspect. Larger ORFs, especially those translating to proteins, often exhibit intricate functions, housing multiple domains that might interact with various proteins or molecular structures. This makes their functional outcomes challenging to anticipate. In a more layered context, some ORFs can undergo processes like alternative splicing or might overlap with other ORFs. This overlapping can lead to the synthesis of several distinct proteins from a single gene, adding another dimension to the intricacies of predicting their expression and function.

Interestingly, the distinction between ORFs and short open reading frames (sORFs) forms a nuanced aspect of genome analysis, bearing significance for the accurate interpretation of genetic material and its implications for GMO risk assessment. Traditional ORFs, which are typically larger, have been the primary focus of bioinformatics tools, inadvertently overshadowing sORFs, which can sometimes be crucial in regulating cellular functions. Consequently, sORFs, despite their ubiquitous nature across all genomes, have been largely neglected due to the inherent challenges in verifying their translational potential (Andrews and Rothnagel, 2014). In fact, earlier genome annotations often overlooked sORFs, assuming them to carry a high likelihood of being false discoveries. However, advancements in computational biology, combined with high-throughput Ribo-seq (or ribosome profiling), have shed light on a plethora of translated sORFs (Chugunova *et al.*, 2018). These findings expanded the perceived protein-coding potential of genomes.

Recent studies have unveiled a novel category of ORFs—referred to as both sORFs and alternative ORFs (alt-ORFs). These alt-ORFs, though shorter than their traditionally annotated counterparts, challenge the existing annotation paradigms. They often co-exist on the same RNA, contrary to earlier assumptions (Brunet, Leblanc and Roucou, 2020). While definitions can vary, sORFs are generally considered to range between 10 to 100 codons and are often situated in regions annotated as non-coding. On the other hand, alt-ORFs are ORFs longer than 30 codons and can be larger than 100 codons. They commence with an ATG and can be found within any RNA, including mRNAs and non-coding RNAs (ncRNAs) (Brunet, Leblanc and Roucou, 2020). Interestingly, a comprehensive examination of sORFs across species, from flies and mice to humans, revealed their diverse functional classes. These range from mere inert DNA sequences to vital cis-regulators of translation and components of ancestral protein complexes in the cellular cytoplasm (Couso and Patraquim, 2017). Thus, the potential implications of sORFs on gene, peptide, and protein evolution cannot be understated, especially in the context of human disease, physiology, and development. One cannot emphasise enough the significance of minimum ORF length. It's an essential criterion, primarily because it filters out ORFs that are less probable to translate into functional proteins. Many genome studies, including those in *Drosophila melanogaster*, indicate a plethora of these sORFs that actively undergo translation, producing peptides of mostly unknown function (Couso and Patraquim, 2017). The median size across species for these sORFs is around 23 codons, a value expected based on the probability of encountering stop codons. Other research references suggest a length of approximately 100 nucleotides to differentiate between potential protein-coding ORFs and mere random occurrences of codons (Woodcroft, Boyd and Tyson, 2016; McNair *et al.*, 2019; Cerqueira and Vasconcelos, 2021).

In conclusion, when forecasting the likelihood of an ORF's expression, it's not only essential to assess the potential for expression and translation but also to consider the ORF size range. Specifically, we must regard the minimum and maximum lengths tied to their ability to produce proteins. Given that the median size for sORFs across species stands at about 23 codons, this can be used as a conservative benchmark to identify ORFs with a minimal probability of expression.

3.2.4. Discussion of existent information relating to the prediction

Advances in the field of bioinformatics have steered in an era of refined prediction tools and computational methodologies, tailored to discern crucial aspects of ORFs. One pivotal development in this arena was an analytical model specifically shaped for eubacterial genomes, which employed statistical properties of ORFs, such as codon composition and sequence length. This model predicted the average and maximum length, as well as the length distribution of ORFs across a wide spectrum of species, encompassing varied GC contents between 21% and 74% (Mir *et al.*, 2012). Although these models provide extensive insights, intriguing deviations have been observed, especially concerning the alternative reading frames, which surprisingly show a pronounced depletion of stop codons. Such anomalies could be underlined by a selection pressure preventing the fixation of stop codon mutations, potentially pointing to an unknown protein coding capability (Mir *et al.*, 2012).

In tandem with these analytical models, specialized software tools have flourished, streamlining codon analysis. For instance, the PCBI program quickly computes the codon bias index (CBI), a

critical measure that provides insights into gene expression levels and the evolutionary relationships between genes within and across species (Wang, Cheng and Lee, 1998). Moreover, the GCWIND tool is adept at identifying protein-coding ORFs by analysing the base compositions for each possible reading phase, honing in on the sequences that exhibit coding potential (Shields, Higgins and Sharp, 1992). A dimension that's been relatively uncharted until recently is the existence of out-of-frame alternative ORFs. The Human alternative open reading frames (HAltORF) database has emerged as a pioneering tool, showcasing the potential products of out-of-frame alternative translation initiation (ATI) in human mRNAs. This ATI mechanism, prevalent in viruses, seems to also be operational in eukaryotes, potentially augmenting the diversity of the human proteome. The HAltORF database therefore bridges a significant knowledge gap, providing a platform to explore ORFs with strong Kozak contexts and consequently, likely expression in the human transcriptome (Vanderperre, Lucier and Roucou, 2012).

Determining accurate translational start sites is crucial for understanding protein function and transcriptional regulation. Most translational start sites in genome databases are based on bioinformatics predictions, which may not always be accurate. To address this, an experimental method was developed to determine the start sites of proteins in *Mycobacterium tuberculosis* using a combination of epitope tagging and frameshift mutagenesis (Smollett *et al.*, 2009). This method revealed that proteins might start before or after the predicted sites. For example, a previously unannotated ORF upstream of Rv1955 was found to be expressed as a protein, named Rv1954A and the method proposed by the authors revealed that proteins might indeed start before or after the predicted sites and this method can be applied to any bacterial species that can undergo plasmid transformation.

In gene prediction, a common technique involves searching for stop codons. Given that 3 out of 64 trinucleotides in the standard genetic code are stop codons, there is a noticeable difference in stop codon frequency between coding and non-coding sequences. Interestingly, this difference is used for predicting protein-coding ORFs (Pohl, Thei\betaen and Schuster, 2012). Many methods assume a GC content of 50%, but many genomes deviate from this percentage. Adjustments to this method have been made and tested on bacterial genomes such as *Rickettsia prowazekii*, *Escherichia coli*, and *Caulobacter crescentus*, with especially good results on low GC content genomes. Additionally, stop codons can sometimes be read as 'sense' or undergo readthrough, depending on their context. A genome-wide study in *Saccharomyces cerevisiae* identified specific contexts that influence stop codon readthrough (Williams, 2004). It was observed that positions downstream of the stop codon are vital for determining termination efficiency, providing a new perspective on defining potential readthrough contexts in various genomes.

With the advancement of Next Generation Sequencing (NGS), the volume of sequence data has grown significantly, leading to computational challenges. Identifying ORFs in large datasets has become slower. Recent tool like OrfM were developed to overcome these limitations. OrfM uses the Aho-Corasick algorithm to quickly identify ORFs in sequences and it is claimed to be accurate as other tools but much faster, making it ideal for analysing extensive datasets produced by platforms like Illumina (Woodcroft, Boyd and Tyson, 2016).

As ORF prediction tools and methodologies evolve, their applications and implications, especially in the context of GMO risk assessment in food and feed, need further exploration. The

advancements discussed below provide valuable and detailed insights into ORF prediction and its broader applications.

The sections below, provide a more comprehensive analysis of ORF prediction. These subsections explore the various methodologies, input and output data, criteria, settings, and the applicability domain, with a particular focus on the context of GMO risk assessment. Through the process of dissecting the methodologies employed and thoroughly examining the data utilised, the objective is to enhance the level of transparency and deepen the understanding of ORF prediction paradigm that plays a crucial role in the advancement of GMO risk assessment.

3.2.4.1. Methods for predictions

Various methods have been developed and employed to predict ORFs, each with its strengths and limitations. Among the most recurring methods are ab initio gene prediction, homology-based methods, and hybrid approaches.

Ab initio gene prediction involves computational algorithms that identify ORFs solely based on the intrinsic features of genomic sequences. These algorithms utilize statistical models, such as hidden Markov models (HMMs), to recognize specific patterns, including start and stop codons, splice sites, and coding/non-coding regions. Some widely used ab initio gene prediction tools include GENSCAN, Augustus, and FGENESH. Additionally, popular ORF predictors like Glimmer, GeneMark, Prodigal, and FgeneSB, employ different approaches such as interpolated Markov model, hidden Markov model, and log-likelihood to pinpoint the genic regions in genomes (Kumar *et al.*, 2016). Although these tools can predict ORFs without prior knowledge of gene structure, they can be sensitive to the choice of training data and may not perform well for organisms with distinct genomic features.

Homology-based methods, on the other hand, rely on the comparison of genomic sequences with known gene sequences from closely related organisms. These methods, which include tools like BLAST and FASTA, search for conserved regions, enabling the identification of putative ORFs based on sequence similarity. While homology-based methods can provide more accurate predictions when homologous sequences are available, they may be limited by the completeness of reference databases and may not detect species-specific or novel genes.

Hybrid approaches combine ab initio gene prediction with homology-based methods to improve the accuracy of ORF prediction. These methods integrate information from multiple sources, such as sequence similarity, gene expression data, and functional annotations. Examples of hybrid gene prediction tools include Maker and JIGSAW. One such method integrates ribosome profiling data with computational strategies to discern 'real' translation from noise, capturing features of translating protein-coding ORFs (Pauli, Valen and Schier, 2015). By leveraging the strengths of both ab initio and homology-based methods, hybrid approaches can offer more accurate predictions, especially in instances where individual methods might have limitations.

Beyond the methods mentioned above, there are other approaches for ORF prediction. These may include machine learning-based methods that utilize features like codon usage bias, sequence motifs, and other sequence properties to predict ORFs. An overview of these tools is provided below but is also detailed in Annex I.

Neural network techniques predicted solely on the basis of single codon frequencies calculated by counting the number of the 64 possible different codons in an ORF of a specified length (Farber, Lapedes and Sirotkin, 1992). In a study to correct over-annotation in microbial genomes, an enhanced graphical representation, I-TN, was introduced for the *Amsacta moorei entomopoxvirus* genome, leading to accurate reannotation of protein-coding genes (Yu and Sun, 2010). Concurrently, the rise of human genomic sequences has been addressed using the Virtual Transcribed Sequence project combined with GENSCAN, identifying novel human genes not yet cataloged in EST databases (Miyajima, Burge and Saito, 2000). The realm of long noncoding RNAs (lncRNAs) has been scrutinized to discern between protein-coding and noncoding transcripts. Multiple tools like Coding Potential Calculator (CPC), Ribo-seq, polysomal fractionation, and Mass spectrometry (MS) have illuminated the translation aspects of these lncRNAs, suggesting that while many undergo translation, only a fraction yield functional peptides (Housman and Ulitsky, 2016). To tackle sORFs prediction in microbial genomes, SmORFinder was developed. Incorporating profile hidden Markov models and deep learning, it offers enhanced prediction and annotation capabilities, and its findings are accessible via a public web portal (Durrant and Bhatt, 2021). Another tool named SYNCOD is a program developed for identifying protein-coding regions and it relies on the conservative evolutionary properties of coding regions, utilizing BLASTN alignment (Rogozin, D'Angelo and Milanesi, 1999). An improved method, the Alternative Spectral Rotation measure, enhances the prediction of protein-coding regions in rice DNA (Jin, 2004). Ribo-TISH is another toolkit designed for analysing translation initiation sequencing (TI-seq) data, offering the potential to predict novel ORFs (Zhang *et al.*, 2017). lncScore is an alignment-free tool that differentiate lncRNAs from mRNAs (Zhao, Song and Wang, 2016). A proteogenomic method was automated for maize, uncovering novel protein-coding genes in its genome (Castellana *et al.*, 2014). GeneScan utilizes Fourier techniques to detect the three-base periodicity in genomic sequences, aiding in recognizing coding regions (Tiwari *et al.*, 1997). ORFLine, a computational pipeline, is adept at identifying and classifying sORFs, providing insights into potential secreted proteins from lymphocytes (Hu *et al.*, 2021). It should be noted that all these methods may require extensive training datasets and might not be as universally applicable as ab initio, homology-based, or hybrid methods.

In the context of GMO risk assessment, the accurate prediction of ORFs is essential for assessing the likelihood of expression of transgenes or other genetic elements. While existing methods have been successfully applied across some organisms and contexts, there remains room for improvement, especially in addressing challenges posed by diverse organisms, novel genes, and the integration of multiple data types. It is essential to recognize the challenge to combine organisms and context in one single tool. This is primarily because such a combination would necessitate a comprehensive dataset for training, modelling and validation that cannot be readily generalized to different organisms.

3.2.4.2. Input data for prediction tools

In the context of predicting open reading frames (ORFs), various input data types are used to improve the accuracy and reliability of prediction methods. Some of the most common input data include genomic sequences, transcriptomic data, and proteomic data, while other less common data types may also be utilized to enhance the prediction process.

Genomic sequences are the primary source of information for ORF prediction (Miyajima, Burge and Saito, 2000). These sequences, which represent an organism's complete genetic

information, serve as the basis for identifying coding regions and potential ORFs. The quality of genomic sequence data is critical to the success of ORF prediction, as errors or gaps in the sequence can lead to false predictions or missed ORFs.

Transcriptomic data, which encompass the entire set of RNA transcripts in a given cell or tissue, provide valuable information about gene expression patterns and can aid in the identification of ORFs. Transcriptomic data can be obtained through methods such as RNA sequencing (RNA-seq) and, still nowadays, microarrays (Kiniry, Michel and Baranov, 2020; Hu *et al.*, 2021). The integration of transcriptomic data with genomic sequences helps to refine ORF predictions by identifying expressed regions and providing evidence for the existence of functional genes.

Proteomic data, which represent the entire set of proteins expressed in a specific cell or tissue, can also be used to predict ORFs. Mass spectrometry-based techniques are commonly employed to identify and quantify proteins, providing experimental evidence for the translation of ORFs into functional proteins. By comparing proteomic data with predicted ORFs, researchers can validate and improve the accuracy of ORF prediction methods (Castellana *et al.*, 2014).

In addition to these common data types, other sources of information may be used to support ORF prediction. These can include functional annotations, such as gene ontology terms or protein domain information, which can help to assign putative functions to predicted ORFs. Comparative genomics data, which involve the analysis of genomic sequences from multiple related species, can be employed to identify conserved coding regions and improve ORF predictions. For example, almost all ab initio predictors rely on a training set of ORFs from which it generates a model of protein coding genes (Kumar *et al.*, 2016). Moreover, epigenomic data, such as histone modification patterns and DNA methylation profiles, can provide insights into the regulatory mechanisms governing gene expression and inform ORF prediction efforts.

In the context of GMO risk assessment, it should be noted that the integration of diverse input data types is expected to improve the performance of ORF prediction methods and enhance the overall reliability of risk assessment efforts, including the identification of new data sources and integrative approaches that are needed to address the challenges posed by diverse organisms, novel genes, and the complex nature of gene expression regulation.

3.2.4.3. Output data

The most common output data include predicted ORF coordinates, amino acid sequences, and functional annotations. Additional output data may encompass expression levels, protein domains, and evolutionary conservation scores, among others (Marhon and Kremer, 2011; Housman and Ulitsky, 2016; Spealman, Naik and McManus, 2021).

Predicted ORF coordinates are a fundamental output of ORF prediction methods. These coordinates define the genomic locations of the predicted ORFs, including start and stop codon positions, reading frames, and strand information. This information is crucial for subsequent analyses, such as gene expression studies, functional characterization, and comparative genomics.

Amino acid sequences, derived from the translation of the predicted ORFs, represent another common output data type. These sequences provide insights into the potential protein products of the ORFs and can be used for various downstream analyses, such as protein structure

prediction, functional annotation, and identification of conserved domains (Rogozin, D'Angelo and Milanesi, 1999). Functional annotations are often generated for predicted ORFs to assign putative biological roles to the resulting proteins. These annotations may be based on sequence homology to known proteins or domains, as well as other bioinformatics resources, such as Gene Ontology (GO) terms or KEGG pathway information. Functional annotations can help to prioritize ORFs for further experimental validation and provide insights into the potential impact of the ORFs on the host organism.

Additional output data types may include expression levels, protein domains, and evolutionary conservation scores. Expression levels, derived from transcriptomic data, can provide information on the relative abundance of predicted ORFs and their potential biological relevance. Protein domains, identified through sequence similarity to known domain databases, can help to infer the molecular functions and biological processes in which the predicted ORFs may be involved. Evolutionary conservation scores, calculated through comparative genomics approaches, can provide insights into the functional importance of the predicted ORFs, as conserved sequences are more likely to be functionally relevant.

In the context of GMO risk assessment, accurate and comprehensive output data types can be integrated to provide a more holistic understanding of the potential impact of the ORFs on the host organism and the environment.

3.2.4.4. Criteria and settings for ORF prediction

A variety of criteria and settings are employed to optimize the accuracy and sensitivity of the prediction process. Some of the most frequently used criteria include minimum ORF length, start and stop codon usage, sequence context, and the presence of specific regulatory elements. Additional settings may involve the choice of genetic code, the selection of appropriate reference databases, and the integration of experimental data.

Minimum ORF length is a crucial criterion, as it filters out short ORFs that are less likely to encode functional proteins. While the threshold for minimum ORF length may vary depending on the organism or specific research question, it is considered that small ORFs have a length between 36 and 300 nucleotides. In addition, different tools consider approximately 100 nucleotides the length to differentiate between putative protein-coding ORFs and random occurrences of start and stop codons (Woodcroft, Boyd and Tyson, 2016; McNair *et al.*, 2019; Cerqueira and Vasconcelos, 2021).

Start and stop codon usage is another important criterion in ORF prediction. Although the canonical start codon is AUG (coding for methionine), some organisms utilize alternative start codons. Similarly, the three canonical stop codons (UAA, UAG, and UGA) may be subject to variations. Accurate prediction of ORFs requires accounting for the usage of both canonical and non-canonical start and stop codons in the target organism. Sequence context, such as the presence of Kozak consensus sequences and Shine-Dalgarno sequences, plays a role in translation initiation and is often considered in ORF prediction (Andrews and Rothnagel, 2014). By incorporating information on these regulatory elements, the ORF prediction process can better identify translation initiation sites and improve the accuracy of the predicted ORFs.

The choice of genetic code is another setting that must be carefully considered in ORF prediction. Different organisms may use alternative genetic codes, and selecting the appropriate code for

the target organism is essential for accurate translation of the predicted ORFs into amino acid sequences. Reference databases, such as known protein-coding genes or functional domains, are often utilized in ORF prediction algorithms to identify homologous sequences or conserved features (Cerqueira and Vasconcelos, 2021). The selection of appropriate reference databases can greatly impact the sensitivity and specificity of the prediction process.

Integration of experimental data, such as transcriptomic or proteomic data, can further refine ORF prediction by providing additional evidence for the expression or functionality of the predicted ORFs. By incorporating these data types, researchers can prioritize the most biologically relevant ORFs for further validation and risk assessment.

3.2.4.5. Applicability domain in the context of GMO risk assessment

The applicability domain of ORF prediction tools is an important consideration in the context of GMO risk assessment, as it determines the extent to which a specific method can reliably predict ORFs across various organisms, genomic sequences, and conditions. Several factors influence the applicability domain of ORF prediction methods, including the genetic code, sequence complexity, organism type, and data availability.

The genetic code employed by an organism is a significant factor in determining the applicability domain of ORF prediction methods. Although the standard genetic code is used by the majority of organisms, variations exist in some species, such as in mitochondrial genomes and certain bacteria (Yu and Sun, 2010; Kumar *et al.*, 2016). The applicability of a particular ORF prediction method is affected by its ability to account for these variations and accurately predict ORFs in organisms with non-standard genetic codes.

Sequence complexity, including features such as GC content, repeated regions, and the presence of overlapping or nested genes, can impact the performance of ORF prediction tools. Accurate ORF prediction in regions with high sequence complexity requires algorithms capable of handling these intricacies and distinguishing between true ORFs and artifacts arising from complex sequence features.

Organism type is another factor influencing the applicability domain of ORF prediction methods. Prediction tools may be developed with a focus on specific organisms or groups of organisms, such as prokaryotes or eukaryotes. The performance of these tools may be diminished when applied to organisms outside their intended scope. Therefore, selecting a prediction method tailored to the target organism is essential for reliable ORF prediction (Zhao, Song and Wang, 2016).

Data availability, including the availability of reference genomes, transcriptomic data, and proteomic data, can also impact the applicability domain of ORF prediction methods. The accuracy of ORF prediction can be improved by integrating these data types, which provide additional evidence for the expression and functionality of predicted ORFs. However, the availability of these data varies across organisms and genomic regions, which may constrain the applicability of certain ORF prediction tools.

In the context of GMO risk assessment, the applicability domain of ORF prediction tools is a critical consideration for assessing the likelihood of expression of ORFs and their potential impacts on the host organism and the environment. While many ORF prediction tools are

available, their applicability domain may be limited by factors such as genetic code, sequence complexity, organism type, and data availability.

3.2.5. Discussion of existent information relating to the selection

According to the current Regulation (EU) No 503/20132 on the risk assessment of genetically modified plants, all ORFs created as a result of genetic modification in plants need to be analysed using bioinformatic tools to predict possible similarities with known allergens or toxins. Knowing that the analysis of ORF is a fundamental step in the food and feed risk assessment of GMO, the ELS focused on the possibility to introduce new criteria, or combine existing ones, that are relevant to determine the likelihood of peptides/proteins synthesis (intended and unintended) starting from the information of a specific ORF.

The existing literature, unfortunately, provided limited relevant information directly associated with risk assessment. Nevertheless, by means of the ELS and the evaluation based on the specific guidelines outlined in section 2.2, a series of criteria emerged as possible indicators for the advancement of novel risk assessment methods in the context of ORFs. The evaluation involved examining relevant documents and assessing their methodological rigour, as well as evaluating the quality and quantity of evidence. The ensuing list resulted in a synthesis of research findings. Interestingly, several criteria have been identified, highlighting their potential significance and providing a roadmap for enhancing the research with a more comprehensive understanding and ability to navigate the complexities of ORFs in GMO risk assessment. These criteria included:

- Codon identity, codon choice, and codon bias optimization: Recent ribosome profiling and proteomic studies have discovered many novel coding sequences known as sORFs. These genes have features like non-AUG start codons, making them distinct and challenging to identify using traditional genomic tools. Their involvement in critical biological processes emphasizes the importance of understanding this class of genes (Cao and Slavoff, 2020).
- mRNA secondary structure (e.g., masking of ribosome binding site, secondary structures encoded entirely within the 5′ end of the ORF, and the absence of secondary structure in the 5′ end of ORFs): Alternative splicing plays a significant role in protein diversity, with several subtypes identified like exon skipping, intron retention, and alternative splice sites. This splicing complexity can influence the mRNA secondary structure, having implications for translation efficiency (Pohl *et al.*, 2013).
- Frequency/infrequency of codons: The presence of stop codons can be a tool for identifying potential coding regions. The frequency of these stop codons can deviate based on the GC content, which can impact the length thresholds of potentially coding ORFs (Pohl, Thei\betaen and Schuster, 2012).
- Presence of intron insertion patterns and presence of splice site: Computer analyses found that reading frames often get interrupted by introns at codon boundaries rather than within codons. This pattern is advantageous evolutionarily as it prevents frame shifts (Tomita, Shimizu and Brutlag, 1996).
- AU composition within the 5′ end of an ORF: Bioinformatic analysis of ORF sequences in various bacterial genomes highlighted regional trends in nucleotide sequences influencing protein expression levels. It was found that protein expression is greatly reliant on high AU content in the 5' region of the ORF. This composition affects ribosomal functions including initiation, elongation, and termination phases (Allert, Cox and Hellinga, 2010).

- RNA post-transcriptional modifications: Transcriptomes of higher organisms consist of numerous non-coding RNAs (ncRNAs) that have regulatory roles in gene expression and other biological processes. These RNA modifications and the presence of sense-antisense pairs indicate their significance in functional diversification across genomes (Suzuki and Hayashizaki, 2004).

3.3. Task 3: Novel methods for assessing the likelihood of expression of ORFs

This section undertakes an examination of the viability of utilising and incorporating different tools, as outlined in Task 2, specifically in the risk assessment domain keeping into account both traditional transgenic techniques and modern genome-editing techniques. In the following sections, the examination of available prediction tools will be laid out, followed by a discussion on the advantages and drawbacks of different approaches in this context. Then, it will be pointed out the transition into the exploration of future challenges concerning the assessment of ORF expression likelihood within GMOs risk assessment. Lastly, a conceptual framework aimed at addressing the assessment of ORF expression likelihood in risk assessment will be articulated, laying a possible outlook for the development of automated methodologies in the future.

3.3.1. Testing of the available prediction tools

To address the objectives of the tender and evaluate the potential application of existing tools within the framework of ORF risk assessment, a systematic testing of these tools was conducted, as detailed in the method section. Several criteria informed the selection of tools for potential inclusion in the final workflow.

The primary criterion was the added value a tool might offer, as described in their respective publications. Indeed, it was found that different tools lack a of clear added value for the scope of the call and were not explored further. For a detailed list of these tools, the reader is invited to refer to Annex II. Among the different reasons of exclusion it is possible to list:

- Not relevance: Several tools are based on alignment of sequences and are not informative in the context of the coding potential of ORFs.
- Non-availability: Some tools, published several years ago, are no longer accessible. However, the possibility of extracting valuable details about the algorithm or method from their respective publications could be considered.
- Performance issues: Some tools may operate at reduced speeds, but if they introduced new features or methods, they could be still considered. The potential reasons for performance issues were investigated, such as the need for a more advanced computational system or parallel processing.
- Results display: Challenges related to the presentation of results are considered minor. With appropriate parsing techniques, the display of results could be optimised.
- System requirements: Some system requirements can be addressed with solutions such as using virtual machines for tools designed for Linux.
- Large databases: With appropriate computational systems, managing large databases could be achievable.

In contrast, the tools below showed potential benefits upon evaluation for the scope of this tender. As reported also in Task 2, methods for determining the likelihood of ORF expression can generally be divided into three distinct categories:

- coding potential tools
- machine learning methods
- mathematical methods

Each of these approaches provides unique capabilities that, when leveraged correctly, could significantly enrich our understanding of genomic data and potential ORF expression. Machine learning techniques, notably Convolutional Neural Networks (CNNs), can be highly advantageous as they extract crucial characteristics from raw data, thereby enhancing the accuracy of ORF prediction. There are computational tools available that are reference-free, allowing for exploration of novel or less well-characterized genomes without needing prior knowledge of the composition of protein-encoding genes. Further, advanced multivariate analyses often prove more effective than the common e-value cutoff approach, bolstering the precision of ORF predictions. Certain tools also exhibit robustness and speed, allowing for real-time assessment of the coding potential of transcripts. Some can distinguish between different types of RNAs with sufficient accuracy and others can be automated, thereby simplifying the ORF prediction and identification process.

However, there are limitations to consider. For example, data quality and availability can pose challenges, such as the limited availability of well-aligned genomes for multiple species or the absence of benchmarkable gold standard translated ORF sets. Genomes with high GC content pose further challenges due to fewer stop codons and more alternate start codons, while the quality and depth of sequencing can sometimes lead to false-positive results. Computational limitations, such as the intensive computation required to process and analyse large amounts of data, overfitting and high computational cost associated with training CNNs, difficulties with long DNA sequences containing multiple exons and introns, and the increased computation time required for certain methods, can also impact the reliability and efficacy of the tools. Methodological challenges include the presence of footprints in a genomic region that may not necessarily signify translation, and difficulties in detecting potential peptides from the translation of annotated lncRNAs using classical Mass Spectrometry (MS) design. Similarly, biological and experimental limitations may include the need for experimental validation to confirm *in silico* determinations of the expression of transcripts with sORFs, particularly those from intergenic regions. Finally, software and tool limitations can include differences in predictions made by various tools for predicting translated ORFs, and limitations of certain software in predicting sORFs. The training set used can significantly influence the sensitivity and specificity of gene prediction, and over-training the ANN can decrease the global error of the training set.

Considering these observations and the extensive testing of various tools (see Annex II), three computational tools emerged as the most valuable for the objective of this call:

- CPAT (Coding Potential Assessment Tool)
- RNA samba
- Coding Potential Calculator 2

These tools and their application are detailed as follows.

Coding Potential Assessment Tool (CPAT)

The CPAT is a bioinformatics software tool used to assess the protein-coding potential of a given RNA sequence. This tool is designed to rapidly and distinguish coding from noncoding RNA sequences, a critical step in functional genomics studies. CPAT employs a logistic regression model, using four sequence features frequently associated with coding potential. These are: open reading frame (ORF) size, ORF coverage, Fickett TESTCODE statistic (an indicator of coding potential based on nucleotide arrangement), and hexamer usage bias. The logistic regression model is trained on these four features to provide a scoring system for the coding potential of a given sequence. CPAT can build species-specific models for coding potential prediction. It comes with pre-computed logistic regression models for several species, but it also provides flexibility to the user to train the model on a new set of coding and noncoding transcripts for a specific species. CPAT is high-speed performance and minimal computational resource requirements make it an effective tool also for large-scale transcriptome data. This tool can be accessed from Web server for Coding Potential Assessing Tool (CPAT) (bcm.edu) and was tested within the scope of this call with some example sequences. A sequence can be submitted in FASTA format and in a straightforward manner with the setup displayed and the string ">name of the sequence" must be appended before the nucleotide sequence to start the prediction. Results are displayed in the dialog box and be opened with any plain text file reader. It should be emphasized that the tool is efficient and user-friendly. In addition, it should be noted that the tool initially comes with a limited number of species in its database. However, as previously mentioned, there is flexibility to incorporate custom datasets for expanded usability. The creation of these datasets can pose challenges, particularly in ensuring they are sufficiently extensive for robust model training and internally consistent (e.g., derived from the same organism), which is crucial for the tool's effectiveness and accuracy in predictions.

RNA Sequence and Motif Base Assessment (RNA samba)

RNA samba is a computational tool utilized for the prediction of coding potential in transcripts. It applies a machine learning approach using a Random Forest classifier. The tool leverages sequence-derived features, such as k-mer frequency and positional nucleotide frequencies, as well as other motifs identified within the sequences. RNA samba integrates a multitude of sequence and motif-based features. Additionally, it's built on machine learning, which means its predictions could be improved as it processes more data. However, like other tools, the performance of RNA samba is closely related to the quality and representativeness of the training data. Currently, RNA samba primarily supports human and mouse species data. However, its usability for other organisms can be enhanced by incorporating custom datasets, though the process may pose challenges in ensuring the datasets are sufficiently comprehensive and internally consistent for effective utilization. This tool can be accessed from RNAsamba: coding potential calculator for transcript sequences (unicamp.br). It requires that a FASTA file is uploaded as input and, after uploading, the sequence is submitted. Few minutes are needed to perform the calculation and the system will provide a link for downloading the results. This can be done in csv format and opened with any text reader The RNAsamba tool is both efficient and user-friendly.

Coding Potential Calculator 2 (CPC2)

CPC2 is a bioinformatics tool designed to evaluate the protein-coding potential of a transcript. It makes use of multiple sequence features and a support vector machine (SVM) classifier to predict the coding potential of a given RNA sequence. These features include ORF size, ORF coverage, Fickett TESTCODE statistic, isoelectric point, and more. Unlike its predecessor, CPC, CPC2 does not require a sequence alignment, which makes it faster and more suitable for large datasets. The tool has been trained with different datasets. It can be used for a broad range of sequences, including long non-coding RNAs (lncRNAs), and sequences from novel or less well-characterized genomes. CPC2 is species-neutral, making it useful for non-model organism transcriptomes. CPC2 has a user-friendly interface and can be used online or installed locally. The results, which include the calculated coding potential scores and detailed information about the predicted ORFs, can be downloaded for further analysis. This tool can be accessed from CPC2 @ CBI, PKU (gao-lab.org). FASTA sequence need to be pasted into the box and, once the sequence is processed, output are shown in a distinct tab. Results depict an output with a coding probability for the standard sequence, and a non-coding probability for the reverse sequence. A comprehensive analysis of the result can be accessed through the "View" hyperlink.

3.3.2. Advantages and drawbacks of different approaches

Predictive techniques for ORFs in genomic sequences have seen various developments over the years, resulting in an assortment of methods. These differ widely in their features and capabilities, notably in the domains of accuracy, efficiency, and their areas of application. An in-depth exploration into these methods provides insights into the inherent advantages they offer for ORF prediction. The following sections will delve into the advantages and limitations of both computational and experimental methods. These distinct approaches have historically been treated as separate paradigms. However, in recent times, there has been a noticeable trend towards amalgamating computational and experimental techniques to harness the strengths of both, leading to more holistic methodologies. It is prudent to note, based on the evidence gathered, that while a range of tools currently exist for ORF prediction, they typically address specific aspects or challenges. Evidence from the ELS suggests that there remains a gap in the provision of a holistic tool adept at assessing the likelihood of gene expression, particularly from ORF data, in the context of GMOs risk assessment.

3.3.2.1. Computational methods

In the realm of computational methods for predicting ORFs in genomic sequences, three predominant approaches have been developed. Firstly, *ab initio* methods rely exclusively on the intrinsic features of genomic sequences to identify ORFs. These algorithms make use of statistical models like hidden Markov models (HMMs), support vector machines (SVMs) or artificial neural network (ANN) to detect sequence features associated with coding regions. An intrinsic advantage of this approach is its independence from previous experimental data, making it particularly useful for novel or lesser-known organisms. However, its effectiveness can be constrained by the precision of the statistical models applied. Homology-based methods, on the other hand, pivot on sequence similarity between the sequence in question and known coding sequences from related organisms. Notably, these methods can tap into an existing body of knowledge on gene structures and functions, offering insights into potential roles of predicted ORFs. By leveraging the conservation of function and structure present in homologous sequences, these methods are better poised to predict orthologous genes with accuracy. Lastly, integrative methods encompass a holistic approach by amalgamating multiple lines of evidence,

such as those from ab initio, homology-based, and other approaches, notably with combination of also experimental data. This synergy not only ensures better accuracy but also allows researchers to identify ORFs which might be overlooked when using a singular approach.

Main advantages of computational methods: The advancements in computational tools have significantly bolstered the accuracy of ORF prediction. For instance, machine learning techniques, including convolutional neural networks (CNNs), hold the capacity to derive meaningful characteristics straight from raw data. Furthermore, certain computational tools, being reference-free, don't hinge on prior knowledge of protein-encoding genes, though their applicability domain need to be validated. Nevertheless, this feature is invaluable when dealing with lesser-known genomes. By leveraging advanced multivariate analysis, tools can also exhibit enhanced sensitivity over the more rudimentary e-value cutoff. Their robust nature, combined with speed and user-friendly interfaces, empowers researchers to assess the coding potential of several numbers of transcripts in real-time. Moreover, some automation capabilities offered by certain tools can streamline the ORF prediction process.

Main drawbacks of computational methods: Computational methods are not devoid of challenges. In terms of data quality and availability, limitations include coverage of well-aligned genomes for different species and the absence of benchmark sets of translated ORFs. The peculiarities of high GC content genomes present additional challenges. However, the quality of deep sequencing can inadvertently lead to false positives. From a computational viewpoint, the sheer volume of data necessitates intensive computation. There are also concerns over overfitting and the high computational costs associated with training advanced models like CNNs. Some methods, especially those that engage recursive algorithms, are notorious for their extended computation times. Methodological limitations include the difficulty of discerning true signals from noise, such as the translation of annotated lncRNAs. Furthermore, there are inherent difficulties in spotting non-AUG initiated ORFs or those with a length of fewer than 30 codons. Biological and experimental limitations point towards the need for experimental validation. There are also issues with the software and tools themselves, as anticipated in the previous section. Predictions can vary between tools, and certain software might have intrinsic limitations when it comes to predicting specific types of ORFs or handling unique genomic features. Additionally, the training set used can influence prediction outcomes. Over-training, especially in models like the artificial neural network (ANN), can diminish the efficacy of the predictions and reduce the applicability domain to a single organism. Lastly, setting limitations dictate that the choice of parameters during prediction will invariably depend on the specific context and biological query at hand.

3.3.2.2. Experimental methods

In the context of this tender, while the emphasis primarily rests on computational methods, there is an acknowledgment of the value brought by experimental methods, evident also from the results of ELS seen in Task 1 and Task 2. These methodologies have been divided into two main categories for discussion: Transcriptome-based methods and Proteome-based methods.

Transcriptome-based Methods: Using transcriptomic data, primarily sourced from RNA-seq, these methods target ORFs by identifying the transcribed regions within genomes. The benefits of this approach are manifold. They offer a holistic view of gene activity, highlighting ORFs that are in the process of transcription and translation. This perspective aids in refining genome

annotations, enhancing the accuracy of ORF identification. Moreover, these techniques prove effective even without a specific reference genome sequence, facilitating the exploration of lesser-studied human genomes. As the quality of genome and transcriptome annotations improves, the accuracy of predicting functional ORFs also sees an enhancement. These methods complement findings from proteomics data (see below), endorsing the existence of projected ORFs. They also offer solutions to some challenges posed by proteomics, especially when determining the exact boundary of novel genes. Their versatility is further showcased in their ability to study gene expression across a spectrum of cells and tissues. On the other hands, transcriptome-based methods present also challenges. The complexities of the human genome pose potential barriers. For instance, some regions, due to their complex alignment or lack of sequencing, can impede gene expression profiling. The chosen training gene set plays a pivotal role. If this set fails to capture the breadth of organism genes, the results might be biased. Varying attributes among genes, such as GC content and codon usage, have the potential to mislead predictions and, in some cases, results can be ambiguous and influenced by factors like RNA sample quality and experimental conditions. Contaminants, particularly ribosomal RNAs in sequenced data, introduce further complexities. Using RiboSeq data to detect translation presents its challenges, leading to potential misinterpretations. Furthermore, there is recognition that not all sequences in a ribosome profiling cDNA library originate genuinely, adding a potential layer of inaccuracy. On a logistical note, these methods often demand significant resources, be it time, funds, or technical expertise.

Proteome-based Methods: When exploring proteome-based techniques, mass spectrometry-based proteomics data is at the forefront. This approach delves into the protein products of projected ORFs, underscoring their functional relevance. A hallmark of these methods is the direct identification of translated ORFs, encouraging confidence in their functional role. These methods not only refine genome annotation but also address challenges in frame identification that are commonly faced in transcriptomics-reliant annotation processes. Furthermore, they validate the transition of predicted ORFs into tangible proteins. However, proteome-based methods are not devoid of challenges. Proteomics studies, while rich in potential, are not primarily designed for genome annotation. Detecting peptides, especially those resulting from the translation of annotated lncRNAs using standard mass spectrometry design, can be intricate and an inherent limitation of mass spectrometry is its sensitivity, which can sometimes omit genes expressed at more subtle levels, leading to potential inaccuracies in predictions.

As advancements in the field of genomics continue, there is a visible convergence of computational and experimental methods. It's evident that these experimental techniques can work in tandem with computational ones. For those assessing the likelihood of expression in GMO risk assessment, this integrated perspective should be given due consideration to achieve comprehensive and precise insights.

3.3.3. Future challenges for ORF for assessing the likelihood of expression of ORFs in the context of GMOs risk assessment

From the ELS, it can be observed that different aspects currently hinder the development of models to assess the likelihood of ORF expression. Simultaneously, these aspects discussed below represent future challenges not only for risk assessment in GMO but also for the broader

scientific community. The sections below are meant to provide a detailed description of the encountered challenges.

Absence of specific data sets

A main hurdle in creating new tools and methods to determine the likelihood of ORF expression and subsequent protein production based on ORF information is the absence of organised data. In fact, while current research on ORFs covers various facets of this issue, data sources frequently differ and are not presented in a consistent format. This void of organised and standardised data, such as on plants or specific organisms, poses difficulties in creating and confirming computer-based models for predicting expression of ORFs. For example, Pauli *et al.* (Pauli, Valen and Schier, 2015) touches upon the challenges and potential in pinpointing coding RNAs and small peptides. Notably, the data they used does not come from a single, consistent source, making the research hard to repeat or expand on. Kumar *et al.* (Kumar *et al.*, 2016) also underscores the problem of data arrangement. They mention that while continuous advancements in sequencing technologies have led to many microbial genome sequencing projects, the prokaryotic genomes sequenced so far are not evenly spread across their evolutionary tree. Such irregular data distribution further hampers the creation of thorough and precise ORF prediction models.

Most importantly, and within the scope and aim of this call, the real gap in specific data sets pertains to the missing details regarding unintended ORFs that might have relevance for food and feed safety in GMOs, rather than the ORFs that are known to be expressed. To address this shortcoming, future studies should concentrate on creating standardised data of ORFs expression in GMOs. This data should ideally be gathered in database that would ease the creation and validation of computer-based models for the prediction of ORF expression.

Diversity of application domain

The diversity of application domains in the field of ORFs further complicates the development of a universal tool for ORF prediction. The literature reveals that the practical applications of novel methods in the field of ORFs span a wide range of aspects, from understanding the mechanisms of viruses to identifying markers of human diseases. This diversity makes it challenging to develop a one-size-fits-all solution for ORF prediction. For example, Pauli *et al.* (Pauli, Valen and Schier, 2015) discusses the application of ORF prediction in the context of zebrafish annotation studies. The authors highlight the use of computational approaches guided by ribosome profiling to identify coding RNAs and small peptides. However, the specificities of zebrafish biology and the techniques used in this study may not be directly applicable to other organisms or contexts. Similarly, the study by Kochetov (Kochetov, 2008) explores the possibility of recognizing several alternative translation start sites in eukaryotic mRNAs. While this research has significant implications for understanding the complexity of eukaryotic gene expression, the methods and findings may not be directly applicable to prokaryotic organisms or to the prediction of ORFs in the context of GMO risk assessment. Another example is the study of Kumar *et al.* (Kumar *et al.*, 2016) that applies ORF prediction to the study of rare taxonomic phyla. While this research has the potential to uncover a wealth of new protein-coding genes, the specificities of these rare organisms may limit the applicability of the methods and findings to other contexts.

The diversity of application domains in the field of ORFs is therefore a significant challenge. Future research should aim to develop flexible and adaptable tools for ORF prediction that can accommodate the specificities of different organisms and contexts.

Lack of information in food/feed

The ELS evidenced no specific studies examining ORFs within the scope of food and feed, especially in risk assessment. This gap does not stem from challenges in assessing the expression of intentionally introduced ORFs, as their expression can be fairly established. Instead, the root of the issue in risk assessment centres on evaluating those hypothetical ORFs which might be unintentionally expressed and could bear potential safety implications. Looking ahead, research endeavours should target this specific area, directing efforts towards devising methods and tools for the prediction of these unintended ORFs, in tune with the unique requirements of GMO risk assessment in relation to food and feed.

Reliability of criteria

Despite putative criteria useful for the definition, prediction, and selection of ORFs relevant to risk assessment being listed, a major limitation is related to the fact that it is challenging to determine whether some variables would constitute reliable input for new models or methods in the context of risk assessment. This aspect is also linked to the lack of structured data in the field. Similarly, the study by Kiniry *et al.* (Kiniry, Michel and Baranov, 2020) discusses the complexity and challenges associated with analysing ribosome profiling data. The paper highlights that while ribosome profiling generates a wealth of data, the processing and analysis of this data require intensive computation and the signal produced is far more complex than standard RNA-seq. This suggests that the reliability of criteria used in these computational methods may be influenced by the complexity of the data and the computational resources available.

Prediction tools associated with experimental data

Different documents describe possible effective models or tools that could be used in the context of risk assessment. However, these often require the generation of de novo data with different experimental techniques such as metabolomics, proteomics, phylogenetics, ribosome profiling, etc. For instance, Kiniry *et al.* (Kiniry, Michel and Baranov, 2020) describes computational methods and tools developed to analyse ribosome profiling data and Kumar *et al.* (Kumar *et al.*, 2016) discusses the use of proteogenomic to annotate protein coding genes on a genome-wide scale, particularly in rare taxonomic phyla.

Lack of integrated tool and ready-to-use methodologies

The current landscape of ORF prediction and assessment in risk analysis is fragmented, with no single tool or methodology that can comprehensively address all the needs in this area. This lack of integrated tools and ready-to-use methodologies is a significant barrier to progress in the field. While there are various tools and methods available for specific aspects of ORF prediction and assessment, they often operate in isolation, each focusing on a particular aspect of the problem. This approach can lead to gaps in the overall analysis, as no single tool can provide a complete picture of the likelihood of ORF expression. For example, some tools may be adept at predicting ORFs based on sequence data, but they may not take into account other important

factors such as the regulatory elements that influence gene expression, or the biological context in which the ORF is found. Other tools may focus on the experimental data such as metabolomics or proteomics, but they may not be equipped to integrate this data with computational predictions to provide a comprehensive assessment of ORF expression likelihood. Furthermore, many of the available tools and methods require a high level of expertise to use effectively, which can be a barrier for researchers who are not specialists in this area. There is a need for more user-friendly tools that can be used by a wider range of researchers, including those who may not have a deep background in bioinformatics or computational biology. In addition, the lack of standardized formats for data and results can make it difficult to compare and integrate outputs from different tools and methods. This lack of standardization can also hinder the development of comprehensive methodologies that incorporate multiple types of data and analysis. Thus, there is a pressing need for the development of integrated tools and methodologies that can provide a comprehensive assessment of the likelihood of ORF expression in risk assessment. Such tools and methodologies should be user-friendly, incorporate multiple types of data and analysis, and adhere to standardized formats to facilitate comparison and integration of results.

3.3.4. Conceptual framework for assessing the likelihood of expression of ORFs in risk assessment

As previously discussed, the complexity of ORF prediction and expression analysis necessitates a comprehensive yet adaptable framework. Within this call, it is proposed a conceptual workflow for navigating the challenges and limitations inherent to ORF research, particularly in the risk assessment context. This framework is presented as an attempt to integrate and streamline the tools and methods currently available to the scientific community. The practical usefulness of such an approach is yet to be understood.

Figure 3 illustrates the workflow beginning with a sequence of interest. Initially, this sequence is analysed with well-known programmes such as Genscan and GeneMark. As described in the preceding section, these tools excel at predicting ORFs and genes within genomic sequences, laying the groundwork for further analysis. Depending on the availability of additional experimental data, the workflow diverges into two distinct paths following this initial evaluation. In cases where sequences lack associated datasets, the workflow emphasises the importance of coding potential tools, particularly CPC2, as described in the previous section. In contrast, for sequences that stand to benefit from existing datasets, machine learning models such as RNA samba become relevant. In addition to machine learning, mathematical tools like CPAT can also be utilised. In fact, a deterministic analysis grounded in mathematical logic may provide a more concrete perspective, especially when supplemented with experimental data. Aware of the dynamic nature of ORF research, the design of the workflow places special emphasis on coding potential prediction tools, recognising their central role and rigorous testing.

While the framework provides an integrated approach, it is essential to underscore that its comprehensive applicability has not undergone extensive testing across diverse data scenarios. However, as the domain advances and as newer tools and datasets surface and undergo refinement, this framework could act as an initial guide for ORF prediction within the realm of risk assessment.

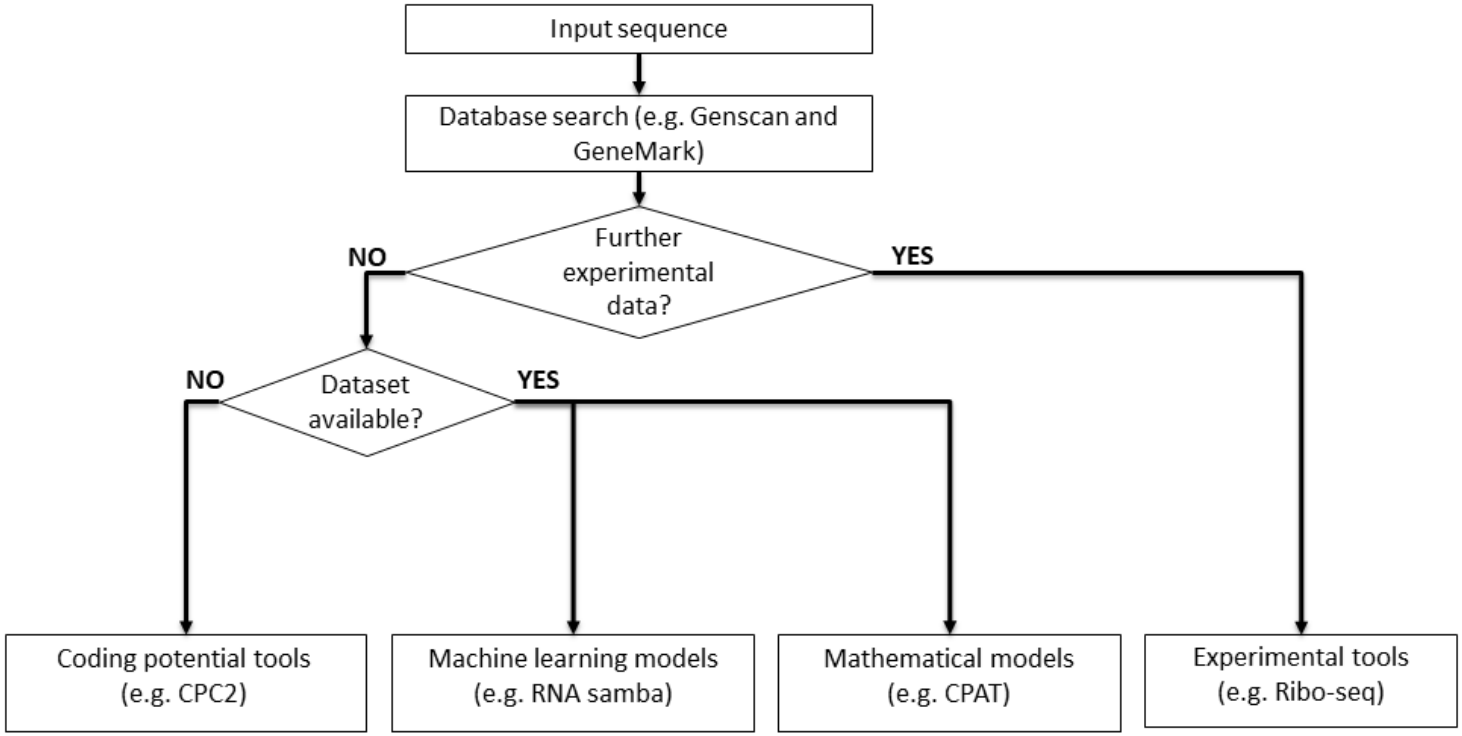

Figure 3: Conceptual framework for assessing the likelihood of expression of ORFs in risk assessment

4. Conclusion

This report partially adheres to the objectives set out by EFSA, embodying the development of criteria for the definition and selection of ORFs pivotal to the risk assessment of GMOs (Objective 1) and the construction of novel knowledge/methods to assess the likelihood of relevant ORF expression (Objective 2). These objectives were dissected into three defined tasks, each laying a critical foundation in the pursuit of refined risk assessment strategies.

Task 1. A systematic search protocol was established to retrieve studies, including reviews and grey literature, that provide information on ORFs. This encompasses their definitions and methods to assess their probability of expression related to risk assessment. The extensive literature search covered areas such as GMOs for food, feed, import, and processing, and extended to areas outside of food safety, like medicine. From the search, 15,484 documents were retrieved. The analysis of titles and abstracts and, subsequently, the full text was conducted, leading to the selection of 307 documents.

Task 2. The in-depth analysis of these documents provided detailed information, highlighting the most significant criteria for the definition, prediction, and selection of ORFs essential for GMO risk assessment. This analysis specifically focused on protein expression related to the primary objective, with an aim to introduce novel methods for evaluating the likelihood of transcription and translation. Findings indicated that certain characteristics of ORF nucleotide sequences influence the likelihood of gene expression. However, the criteria determining this likelihood require further research. Factors such as codon identity, nucleotide composition, and mRNA secondary structure were identified as potentially relevant for developing new risk assessment methodologies. However, challenges remain: the lack of structured data, the diversity of application domains, and the reliability of these criteria pose significant barriers to developing systems that can predict gene expression likelihood from ORF data. Furthermore, documents that directly address ORFs in the context of risk assessment are limited.

Task 3. Based on the information from Task 2, an evaluation was conducted regarding the potential of integrating various tools. The strengths and weaknesses of these tools were assessed in the context of the project's objectives, considering both the risk assessment of traditional transgenic products and products derived from modern genome editing techniques. It was observed that certain characteristics of ORF nucleotide sequences may be relevant in determining the likelihood of expression of significant ORFs for GMO risk assessment. However, further exploration is needed to clarify these criteria and understand the existing limitations. Integrating this knowledge into a single tool is challenging given the diversity of existing tools, especially with respect to their models and associated datasets, which often pertain to specific organisms. A conceptual workflow is proposed for navigating the challenges and limitations inherent to ORF research, particularly in the risk assessment context. This framework is presented as an attempt to integrate and streamline the tools and methods currently available to the scientific community.

Abbreviations

alt-ORFs	Alternative ORFs
ANN	Artificial Neural Network
ATI	Alternative translation initiation
CBI	Codon bias index
circRNAs	Circular RNAs
CPC2	Coding Potential Calculator 2
CPAT	Coding Potential Assessment Tool
dORF	Downstream open reading frame
DB	Database
EC	Exclusion criteria
EFSA	European Food Safety Authority
ELS	Extensive Literature Search
EU	European Union
GMO	Genetic modified organism
HGT	Horizontal gene transfer
HMMs	Hidden Markov models
IRES	Internal ribosomal entry sites
lncRNA	Long non-coding RNA
miPEPs	Micro-peptides
MS	Mass Spectrometry
NCCs	Near-cognate codons
NMD	Nonsense-mediated decay
ORF	Open reading frame
PECO	Populations, Exposure, Comparators, Outcomes
PICO	Populations, Interventions, Comparators, Outcomes
RA	Risk assessment

Ribo-seq	Ribosome profiling
RNA samba	RNA Sequence and Motif Base Assessment
sORF	Small open reading frame
SR	Systematic review
SVMs	Support vector machines

Annexes

Annex I. Summarizing table for data extraction relating ORFs definition (in Excel).

Annex II. List of other tested prediction tools (in Word).

References

Al-Ajlan, A. and El Allali, A. (2019) 'CNN-MGP: Convolutional Neural Networks for Metagenomics Gene Prediction', *Interdisciplinary Sciences: Computational Life Sciences*, 11(4), pp. 628–635. Available at: https://doi.org/10.1007/s12539-018-0313-4.

Allert, M., Cox, J.C. and Hellinga, H.W. (2010) 'Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames', *Journal of Molecular Biology*, 402(5), pp. 905–918. Available at: https://doi.org/10.1016/j.jmb.2010.08.010.

Andrews, S.J. and Rothnagel, J.A. (2014) 'Emerging evidence for functional peptides encoded by short open reading frames', *Nature Reviews Genetics*, 15(3), pp. 193–204. Available at: https://doi.org/10.1038/nrg3520.

Brunet, M.A., Leblanc, S. and Roucou, X. (2020) 'Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs', *Experimental Cell Research*, 393(1), p. 112057. Available at: https://doi.org/10.1016/j.yexcr.2020.112057.

Cao, X. and Slavoff, S.A. (2020) 'Non-AUG start codons: Expanding and regulating the small and alternative ORFeome', *Experimental Cell Research*, 391(1), p. 111973. Available at: https://doi.org/10.1016/j.yexcr.2020.111973.

Cassidy, L. *et al.* (2021) 'Bottom-up and top-down proteomic approaches for the identification, characterization, and quantification of the low molecular weight proteome with focus on short open reading frame-encoded peptides', *Proteomics*, (May), pp. 1–13. Available at: https://doi.org/10.1002/pmic.202100008.

Castellana, N.E. *et al.* (2014) 'An Automated Proteogenomic Method Uses Mass Spectrometry to Reveal Novel Genes in Zea mays', *Molecular & Cellular Proteomics*, 13(1), pp. 157–167. Available at: https://doi.org/10.1074/mcp.M113.031260.

Cerqueira, F.R. and Vasconcelos, A.T.R. (2021) 'OCCAM: Prediction of small ORFs in bacterial genomes by means of a target-decoy database approach and machine learning techniques', *Database*, 2020(6), pp. 1–22. Available at: https://doi.org/10.1093/database/baaa067.

Chugunova, A. *et al.* (2018) 'Mining for Small Translated ORFs', *Journal of Proteome Research*, 17(1), pp. 1–11. Available at: https://doi.org/10.1021/acs.jproteome.7b00707.

Claverie, J. (1997) 'Computational methods for the identification of genes in vertebrate genomic sequences', *Human Molecular Genetics*, 6(10), pp. 1735–1744. Available at: https://doi.org/10.1093/hmg/6.10.1735.

Claverie, J.-M., Poirot, O. and Lopez, F. (1997) 'The difficulty of identifying genes in anonymous vertebrate sequences', *Computers & Chemistry*, 21(4), pp. 203–214. Available at: https://doi.org/10.1016/S0097-8485(96)00039-3.

Couso, J.-P. and Patraquim, P. (2017) 'Classification and function of small open reading frames', *Nature Reviews Molecular Cell Biology*, 18(9), pp. 575–589. Available at: https://doi.org/10.1038/nrm.2017.58.

Durrant, M.G. and Bhatt, A.S. (2021) 'Automated Prediction and Annotation of Small Open Reading Frames in Microbial Genomes', *Cell Host & Microbe*, 29(1), pp. 121-131.e4. Available at: https://doi.org/10.1016/j.chom.2020.11.002.

EFSA Panel on Genetically Modified Organisms (GMO) (2011) 'Guidance for risk assessment of food and feed from genetically modified plants', *EFSA Journal*, 9(5). Available at: https://doi.org/10.2903/j.efsa.2011.2150.

Erady, C., Puntambekar, S. and Prabakaran, S. (2020) *Use of short-read RNA-Seq data to identify transcripts that can translate novel ORFs*. preprint. Genomics. Available at: https://doi.org/10.1101/2020.03.21.001883.

European Commission (2013) *Commission Implementing Regulation (EU) No 503/2013 of 3 April 2013 on applications for authorisation of genetically modified food and feed in accordance with Regulation (EC) No 1829/2003 of the European Parliament and of the Council and amending Commission Regulations (EC) No 641/2004 and (EC) No 1981/2006 Text with EEA relevance*, *32013R0503*. Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32013R0503.

FAO (2022) *Codex Alimentarius - international food standards*. Available at: https://www.fao.org/fao-who-codexalimentarius/en/.

Farber, R., Lapedes, A. and Sirotkin, K. (1992) 'Determination of eukaryotic protein coding regions using neural networks and information theory', *Journal of Molecular Biology*, 226(2), pp. 471–479. Available at: https://doi.org/10.1016/0022-2836(92)90961-I.

Fickett, J.W. (1994) 'Inferring genes from open reading frames', *Computers & Chemistry*, 18(3), pp. 203–205. Available at: https://doi.org/10.1016/0097-8485(94)85014-3.

Housman, G. and Ulitsky, I. (2016) 'Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs',

Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1859(1), pp. 31–40. Available at: https://doi.org/10.1016/j.bbagrm.2015.07.017.

Hu, F. *et al.* (2021) 'ORFLine: a bioinformatic pipeline to prioritize small open reading frames identifies candidate secreted small proteins from lymphocytes', *Bioinformatics*. Edited by J. Xu, 37(19), pp. 3152–3159. Available at: https://doi.org/10.1093/bioinformatics/btab339.

Hung, C.-L. and Lin, C.-Y. (2013) 'Open Reading Frame Phylogenetic Analysis on the Cloud', *International Journal of Genomics*, 2013, pp. 1–9. Available at: https://doi.org/10.1155/2013/614923.

Jin, J. (2004) 'Identification of Protein Coding Regions of Rice Genes Using Alternative Spectral Rotation Measure and Linear Discriminant Analysis', *Genomics, Proteomics & Bioinformatics*, 2(3), pp. 167–173. Available at: https://doi.org/10.1016/S1672-0229(04)02022-4.

Kiniry, S.J., Michel, A.M. and Baranov, P.V. (2020) 'Computational methods for ribosome profiling data analysis', *Wiley Interdisciplinary Reviews: RNA*, 11(3), pp. 1–22. Available at: https://doi.org/10.1002/wrna.1577.

Kochetov, A.V. (2008) 'Alternative translation start sites and hidden coding potential of eukaryotic mRNAs', *BioEssays*, 30(7), pp. 683–691. Available at: https://doi.org/10.1002/bies.20771.

Kozak, M. (1996) 'Interpreting cDNA sequences: Some insights from studies on translation', *Mammalian Genome*, 7(8), pp. 563–574. Available at: https://doi.org/10.1007/s003359900171.

Kumar, D. *et al.* (2016) 'Proteogenomics of rare taxonomic phyla: A prospective treasure trove of protein coding genes', *PROTEOMICS*, 16(2), pp. 226–240. Available at: https://doi.org/10.1002/pmic.201500263.

Livny, J. (2005) 'sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes', *Nucleic Acids Research*, 33(13), pp. 4096–4105. Available at: https://doi.org/10.1093/nar/gki715.

Long, M., Rosenberg, C. and Gilbert, W. (1995) 'Intron phase correlations and the evolution of the intron/exon structure of genes.', *Proceedings of the National Academy of Sciences*, 92(26), pp. 12495–12499. Available at: https://doi.org/10.1073/pnas.92.26.12495.

Ma, J. *et al.* (2016) 'Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides', *Analytical Chemistry*, 88(7), pp. 3967–3975. Available at: https://doi.org/10.1021/acs.analchem.6b00191.

Marhon, S.A. and Kremer, S.C. (2011) 'Gene Prediction Based on DNA Spectral Analysis: A Literature Review', *Journal of Computational Biology*, 18(4), pp. 639–676. Available at: https://doi.org/10.1089/cmb.2010.0184.

Marquez-Molins, J. *et al.* (2021) 'Might exogenous circular RNAs act as protein-coding transcripts in plants?', *RNA Biology*, 18(sup1), pp. 98–107. Available at: https://doi.org/10.1080/15476286.2021.1962670.

McNair, K. *et al.* (2019) 'PHANOTATE: a novel approach to gene identification in phage genomes', *Bioinformatics*. Edited by J. Hancock, 35(22), pp. 4537–4542. Available at: https://doi.org/10.1093/bioinformatics/btz265.

Mir, K. *et al.* (2012) 'Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes', *PLoS ONE*. Edited by J.R. Battista, 7(9), p. e45103. Available at: https://doi.org/10.1371/journal.pone.0045103.

Miyajima, N., Burge, C.B. and Saito, T. (2000) 'Computational and Experimental Analysis Identifies Many Novel Human Genes', *Biochemical and Biophysical Research Communications*, 272(3), pp. 801–807. Available at: https://doi.org/10.1006/bbrc.2000.2866.

Nissley, D.A. and O'Brien, E.P. (2014) 'Timing Is Everything: Unifying Codon Translation Rates and Nascent Proteome Behavior', *Journal of the American Chemical Society*, 136(52), pp. 17892–17898. Available at: https://doi.org/10.1021/ja510082j.

Ong, S.N. *et al.* (2022) 'Small open reading frames in plant research: from prediction to functional characterization', *3 Biotech*, 12(3), p. 76. Available at: https://doi.org/10.1007/s13205-022-03147-w.

Pauli, A., Valen, E. and Schier, A.F. (2015) 'Identifying (non-)coding RNAs and small peptides: Challenges and opportunities: Prospects & Overviews', *BioEssays*, 37(1), pp. 103–112. Available at: https://doi.org/10.1002/bies.201400103.

Peeters, M.K.R. and Menschaert, G. (2020) 'The hunt for sORFs: A multidisciplinary strategy', *Experimental Cell Research*, 391(1). Available at: https://doi.org/10.1016/j.yexcr.2020.111923.

Pohl, M. *et al.* (2013) 'Alternative splicing of mutually exclusive exons—A review', *Biosystems*, 114(1), pp. 31–38. Available at: https://doi.org/10.1016/j.biosystems.2013.07.003.

Pohl, M., Thei\betaen, G. unter and Schuster, S. (2012) 'GC content dependency of open reading frame prediction via stop codon frequencies', *Gene*, 511(2), pp. 441–446. Available at: https://doi.org/10.1016/j.gene.2012.09.031.

Ray, W.C., Munson Jr, R.S. and Daniels, C.J. (2001) '*Tricross*: using dot-plots in sequence-id space to detect uncataloged intergenic features', *Bioinformatics*, 17(12), pp. 1105–1112. Available at: https://doi.org/10.1093/bioinformatics/17.12.1105.

Rogozin, I.B., D'Angelo, D. and Milanesi, L. (1999) 'Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences', *Gene*, 226(1), pp. 129–137. Available at: https://doi.org/10.1016/s0378-1119(98)00509-5.

Sheynkman, G.M. *et al.* (2020) 'ORF Capture-Seq as a versatile method for targeted identification of full-length isoforms', *Nature Communications*, 11(1), p. 2326. Available at: https://doi.org/10.1038/s41467-020-16174-z.

Shields, D.C., Higgins, D.G. and Sharp, P.M. (1992) 'GCWIND: a microcomputer program for identifying open reading frames according to codon positional G + C content', *Bioinformatics*, 8(5), pp. 521–523. Available at: https://doi.org/10.1093/bioinformatics/8.5.521.

Si, X. *et al.* (2020) 'Manipulating gene translation in plants by CRISPR–Cas9-mediated genome editing of upstream open reading frames', *Nature Protocols*, 15(2), pp. 338–363. Available at: https://doi.org/10.1038/s41596-019-0238-3.

Sieber, P., Platzer, M. and Schuster, S. (2018) 'The Definition of Open Reading Frame Revisited', *Trends in Genetics*, 34(3), pp. 167–170. Available at: https://doi.org/10.1016/j.tig.2017.12.009.

Sinha, T. *et al.* (2022) 'Circular RNA translation, a path to hidden proteome', *WIREs RNA*, 13(1). Available at: https://doi.org/10.1002/wrna.1685.

Smollett, K.L. *et al.* (2009) 'Experimental determination of translational start sites resolves uncertainties in genomic open reading frame predictions – application to Mycobacterium tuberculosis', *Microbiology*, 155(1), pp. 186–197. Available at: https://doi.org/10.1099/mic.0.022889-0.

Spealman, P., Naik, A. and McManus, J. (2021) 'uORF-seqr: A Machine Learning-Based Approach to the Identification of Upstream Open Reading Frames in Yeast', in V.M. Labunskyy (ed.) *Ribosome Profiling*. New York, NY: Springer US (Methods in Molecular Biology), pp. 313–329. Available at: https://doi.org/10.1007/978-1-0716-1150-0_15.

Suenaga, Y. *et al.* (2022) 'Open reading frame dominance indicates protein-coding potential of RNAs', *EMBO reports*, 23(6), p. e54321. Available at: https://doi.org/10.15252/embr.202154321.

Suzuki, M. and Hayashizaki, Y. (2004) 'Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs', *BioEssays*, 26(8), pp. 833–843. Available at: https://doi.org/10.1002/bies.20084.

Tiwari, S. *et al.* (1997) 'Prediction of probable genes by Fourier analysis of genomic sequences', *Bioinformatics*, 13(3), pp. 263–270. Available at: https://doi.org/10.1093/bioinformatics/13.3.263.

Tomita, M., Shimizu, N. and Brutlag, D.L. (1996) 'Introns and reading frames: correlation between splicing sites and their codon positions', *Molecular Biology and Evolution*, 13(9), pp. 1219–1223. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a025687.

Vanderperre, B., Lucier, J.-F. and Roucou, X. (2012) 'HAltORF: a database of predicted out-of-frame alternative open reading frames in human', *Database*, 2012(0), pp. bas025–bas025. Available at: https://doi.org/10.1093/database/bas025.

Vazquez-Laslop, N. *et al.* (2022) 'Identifying Small Open Reading Frames in Prokaryotes with Ribosome Profiling', *Journal of Bacteriology*. Edited by T.M. Henkin, 204(1). Available at: https://doi.org/10.1128/jb.00294-21.

Wang, B. *et al.* (2021) 'Improved Identification of Small Open Reading Frames Encoded Peptides by Top-Down Proteomic Approaches and De Novo Sequencing', *International Journal of Molecular Sciences*, 22(11), p. 5476. Available at: https://doi.org/10.3390/ijms22115476.

Wang, T.-T., Cheng, W.-C. and Lee, B.H. (1998) 'A simple program to calculate codon bias index', *Molecular Biotechnology*, 10(2), pp. 103–106. Available at: https://doi.org/10.1007/bf02760858.

Williams, I. (2004) 'Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae', *Nucleic Acids Research*, 32(22), pp. 6605–6616. Available at: https://doi.org/10.1093/nar/gkh1004.

Woodcroft, B.J., Boyd, J.A. and Tyson, G.W. (2016) 'OrfM: a fast open reading frame predictor for metagenomic data', *Bioinformatics*, 32(17), pp. 2702–2703. Available at: https://doi.org/10.1093/bioinformatics/btw241.

Xiang, Y. *et al.* (2023) 'Pervasive downstream RNA hairpins dynamically dictate start-codon selection', *Nature*, 621(7978), pp. 423–430. Available at: https://doi.org/10.1038/s41586-023-06500-y.

Yang, Y. *et al.* (2023) 'Upstream open reading frames mediate autophagy-related protein translation', *Autophagy*, 19(2), pp. 457–473. Available at: https://doi.org/10.1080/15548627.2022.2059744.

Yin, X., Jing, Y. and Xu, H. (2019) 'Mining for missed sORF-encoded peptides', *Expert Review of Proteomics*, 16(3), pp. 257–266. Available at: https://doi.org/10.1080/14789450.2019.1571919.

Yu, J.-F. and Sun, X. (2010) 'Reannotation of protein-coding genes based on an improved graphical representation of DNA sequence', *Journal of Computational Chemistry*, 31(11), pp. 2126–2135. Available at: https://doi.org/10.1002/jcc.21500.

Zhang, H. *et al.* (2018) 'Genome editing of upstream open reading frames enables translational control in plants', *Nature Biotechnology*, 36(9), pp. 894–900. Available at: https://doi.org/10.1038/nbt.4202.

Zhang, P. *et al.* (2017) 'Genome-wide identification and differential analysis of translational initiation', *Nature Communications*, 8(1), p. 1749. Available at: https://doi.org/10.1038/s41467-017-01981-8.

Zhao, J., Song, X. and Wang, K. (2016) 'lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts', *Scientific Reports*, 6(1), p. 34838. Available at: https://doi.org/10.1038/srep34838.