SOFTWARE TOOL ARTICLE

IsoAligner: dynamic mapping of amino acid positions across protein isoforms [version 1; peer review: awaiting peer review]

Jacob Hanimann[1], Holger Moch[1], Martin Zoche[1], Abdullah Kahraman[1,2]

[1]Department of Pathology and Molecular Pathology, University Hospital Zurich, Zurich, Zurich, 8091, Switzerland
[2]Swiss Institute of Bioinformatics, Lausanne, Lausanne, 1015, Switzerland

v1 **First published:** 31 Mar 2022, **11**:382
https://doi.org/10.12688/f1000research.76154.1
Latest published: 31 Mar 2022, **11**:382
https://doi.org/10.12688/f1000research.76154.1

Abstract

Aligning protein isoform sequences is often performed in cancer diagnostics to homogenise mutation annotations from different diagnostic assays. However, most alignment tools are fitted for homologous sequences, leading often to alignments of non-identical exonic regions. Here, we present the interactive alignment webservice IsoAligner for exact mapping of exonic protein subsequences. The tool uses a customized Needleman-Wunsch algorithm including an open gap penalty combined with a gene-specific minimal exon length function and dynamically adjustable parameters. As an input, IsoAligner accepts either various gene/transcript/protein IDs from different databases (Ensembl, UniProt, RefSeq) or raw amino acid sequences. The output of IsoAligner consists of pairwise alignments and a table of mapped amino acid positions between the canonical or supplied isoform IDs and all alternative isoforms. IsoAligner's human isoform library comprises of over 1.3 million IDs mapped on over 120,000 protein sequences. IsoAligner, is a fast and interactive alignment tool for retrieving amino acids positions between different protein isoforms. Its application will allow diagnostic and precision medicine labs to detect inconsistent variant annotations between different assays and databases. Availability: This tool is available as a Webservice on www.isoaligner.org. A REST API is available for programmatic access. The source code for both services can be found at https://github.com/mtp-usz/IsoAligner.

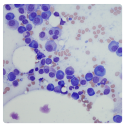
Keywords

alignment, protein isoform, amino acid sequence, protein ids, exon-mapping, amino acid position, splice-variant

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Cell & Molecular Biology gateway.

This article is included in the Bioinformatics gateway.

Corresponding author: Abdullah Kahraman (abdullah.kahraman@usz.ch)

Author roles: Hanimann J: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Moch H**: Funding Acquisition, Resources; **Zoche M**: Conceptualization, Funding Acquisition, Resources; **Kahraman A**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Hanimann J, Moch H, Zoche M and Kahraman A. **IsoAligner: dynamic mapping of amino acid positions across protein isoforms [version 1; peer review: awaiting peer review]** F1000Research 2022, **11**:382 https://doi.org/10.12688/f1000research.76154.1

First published: 31 Mar 2022, **11**:382 https://doi.org/10.12688/f1000research.76154.1

Introduction

Mapping isoform sequences to each other and identifying corresponding amino acids (AA) between isoforms is an important and prevalent task, especially in the interpretation of cancer mutations [Stephenson *et al.*, 2019]. Most functional databases like COSMIC (RRID:SCR 002260), ClinVar (RRID:SCR 006169), or gnomAD (RRID:SCR 014964) use the longest protein isoform as a reference for their annotations. However in cancer diagnostics, shorter splice variants with different amino acid positions can be chosen, leading to confusions with respect to the existence of mutations in the aforementioned functional databases. Many inconsistent variant annotations exist [Tsai *et al.*, 2021]. For example, the MET p.D1246N resistance mutation arising in lung cancers with MET exon 14 skipping events is commonly annotated as p.D1228N in diagnostic assays. While the first is annotated on the 1408 AA long NM 001127500.3, the latter is annotated on the 18 AA shorter transcript NM 000245.4. Simply looking for information on MET p.D1228N in functional databases might thus result in wrong conclusions. To find corresponding AAs in other databases, one general approach is to read off the corresponding positions from a pairwise global alignment like the Needleman-Wunsch algorithm (available for example at the EBI or SIB). However, the optimal solution to a global alignment can include the alignment of distinct exons giving the false impression that AA at these regions are corresponding to each other. To circumvent this problem, the Mirage [Nord *et al.*, 2018] software performs a computationally expensive multiple sequence alignment of the corresponding sequences in the genome.

Here, we introduce IsoAligner, the first web-service for effortless, fast, dynamic and interactive positional mapping of AA between isoform sequences. The tool applies a gene-specific minimal exon length function integrated into a Needleman-Wunsch algorithm to identify false-positive correspondences between amino acids of different isoforms. IsoAligner is simple, interactive, supports simultaneously gene/transcript/protein IDs from ENSEMBL (RRID:SCR 002344), UniProt (RRID:SCR 002380), RefSeq (RRID:SCR 003496), HGNC (RRID:SCR 002827) and UCSC (RRID:SCR 011624), and returns a ready-to-use mapping table between AA positions of protein isoforms. The source code for IsoAligner is available from GitHub and is archived with Zenodo (Hanimann & Kahraman, 2022).

Methods

Alignment approach

The challenge of aligning protein isoforms can be described as matching identical exons. The IsoAligner algorithm exploits this elementary characteristics of isoforms and applies custom parameters to the established Needleman-Wunsch algorithm followed by an evaluation of all subalignment lengths. Subalignments that do not meet the gene-specific minimal exon length, are discarded and marked as false-positive correspondences. The default global alignment parameters have been selected to support island-like solutions of the alignment with an heuristically predefined open gap penalty score. Gap extensions are not penalized (match: 1, mismatch: -2, open gap: -1, gap extend: 0). However, the user has the possibility to interactively change and adjust these parameters.

Implementation

The IsoAligner software is written in Python v3.8 (RRID: SCR 008394). The alignment algorithm is based on the *align. globalms* function of the pairwise2 module from the Bio package (RRID:SCR 007173) and the website is built with streamlit. A REST API for programmatic access runs on the flask framework.

Human Isoform Library

The Human Isoform Library forms the core of IsoAligner. The library is a comprehensive reference database comprising +1.3 million gene/transcript/protein IDs mapped on 120k protein sequences from multiple sequence reference databases namely Ensembl [Howe *et al.*, 2021], UniProt [The Uniprot Consortium, 2021], RefSeq [O'Leary *et al.*, 2016], HGNC [Tweedie *et al.*, 2021] and UCSC [Damian Smedley *et al.*, 2015]. The integration of the different databases was carried out by pairing IDs of protein sequences to each other by using the Biomart (RRID:SCR 002987) mapping tool and comparing raw amino acid sequences. The individual minimal exon length was required to be at least three AA and extracted from Ensembl's GTF file (v104). For custom sequences provided by the user, we set the minimum exon length to 12 AA corresponding to the median length of all shortest exons in a gene. Using our adapted Needleman-Wunsch alignment approach we were able to map for the whole human isoform library with 106k alignments and a total of 40.5 million perfect AA matches (dataframe available online). However, we could also identify 862,136 false-positively aligned amino acids positions that could have resulted in a wrong amino acid position in an alternative isoform. Ultimately, our human isoform library provides a clean positional mapping table for corresponding AA for all alternative protein isoforms.

Operation

Since the front-end of IsoAligner is built with streamlit, it is compatible with following browsers:

- Google Chrome (version 98 or newer)
- Firefox (version 97 or newer)
- Microsoft Edge (version 98 or newer)
- Safari (version 14 or newer)

***Website*.** The input text field on www.isoaligner.org accepts various gene and protein IDs as well as custom sequence pairs (see Figure 1). Gene and protein IDs from different databases can be mixed and searched simultaneously. The workflow is as follows:

Quick Start: Click on 'Show Example' and then 'Search and Align' to get a overview.

- Enter either one isoform ID per gene or two isoform IDs per gene or a list of genes names or two raw

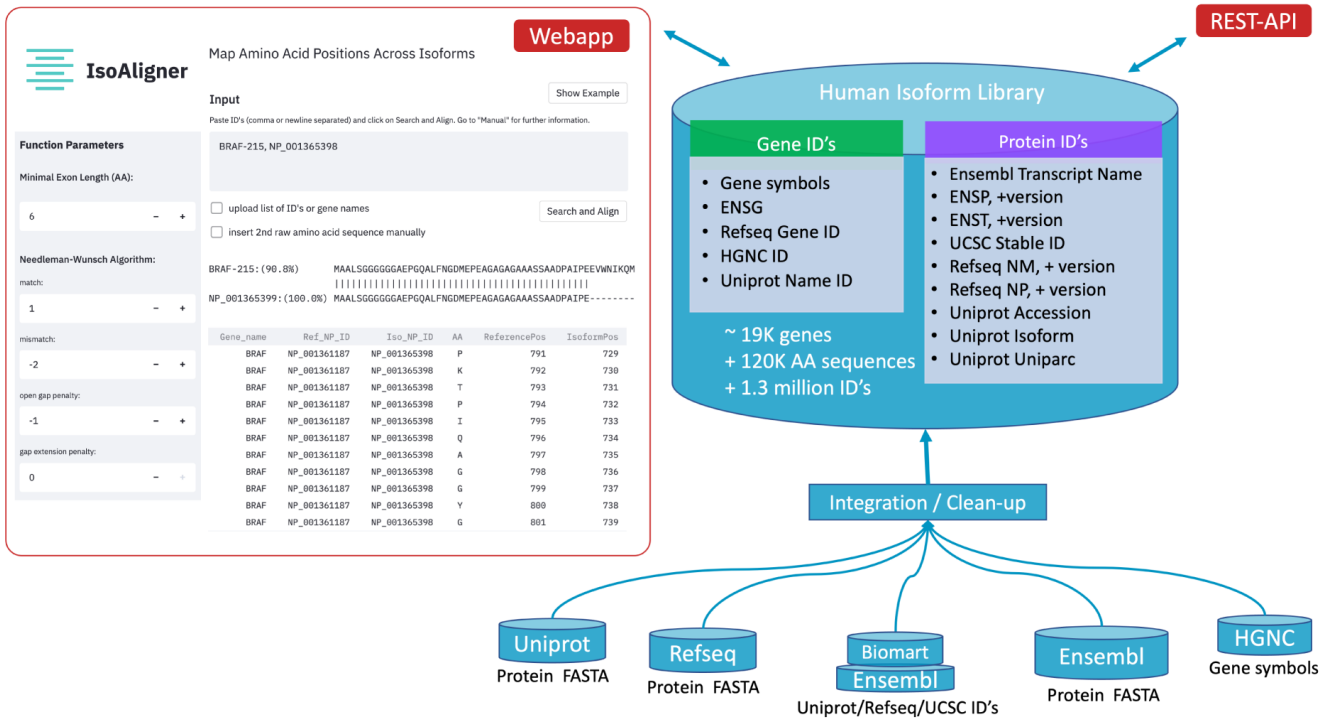


Figure 1. Overview of IsoAligner: the project structure from back-end generation of the human isoform library to the front-end user interaction.

amino acid sequences. The input can be tab, comma or whitespace separated. Click 'Search and Align' or 'Align' to compute alignments.

- Information to the chosen reference sequence and its alignments against all other isoforms can be displayed by using corresponding drop-down buttons.
- Further down on the page, the computed mapping table is shown. On the left sidebar of the application, the function parameters of the Needleman-Wunsch algorithm and the minimal exon length function are displayed. Changing the values of the parameters, instantly updates the alignment visualisation and mapping table.
- The mapping table can be filtered using the 'Filter table for exact value' input field and pressing enter.
- The entirety of the mapping table can additionally be complemented with associated isoform IDs and be downloaded as a csv or tsv file.

Further information on how to use the IsoAligner can be found at the "Manual & About" section on the left sidebar of IsoAligner's website.

REST API. The REST API is built using Flask v1.1 and is accessible through the URL www.isoaligner.org/api. Currently, a get method for isoform IDs called "map" is available for the retrieval of mapping tables between corresponding amino acid positions as well as the method "align" to retrieve the alignment of two raw protein sequences.

The **resource "map"** gives access to the human isoform library, computes alignments with specified parameters and retrieves whole mapping tables in json format. The only required parameter is id1, to provide the Isoform ID of interest. Additional parameters are:

- *Parameter: id1*

  ID of any type (Ensembl, Refseq, Uniprot, UCSC) to access the isoforms of a gene of the human isoform library. To define the reference protein sequence against which all other splice variants will be aligned, a specific isoform identifier should be used. Otherwise, the longest isoform is automatically chosen as the reference canonical sequence.

- Request example: *www.isoaligner.org/api/map?id1=EGFR-201*

  Response: Entirety of a mapping table in json format for EGFR with EGFR-201 defined as the reference sequence aligned against all other isoforms of the human isoform library.

- *Parameter: id2*

Specific isoform ID (Ensembl, Refseq, Uniprot, UCSC) to use as the alternative splice variant to align with the reference sequence of id1.

Request example: *www.isoaligner.org/api/map?id1=EGFR-201&id2=EGFR-207*

Response: mapping table in json format for EGFR-201 aligned against EGFR-207.

- *Parameter: pos*

  In case of setting id1 and id2 in the request, a single corresponding AA positions on the alternative isoform sequence can be retrieved.

  Request example: *www.isoaligner.org/api/map?id1=EGFR-201&id2=EGFR-207&pos=1038*

  Response: 993

- *Parameter: min_ex_len*

  The alignment parameter for the minimal exon length (consecutive AAs) is gene-specific per default and can be manually defined as follows:

  Request example: *request:www.isoaligner.org/api/map?id1=EGFR-201&id2=EGFR-207&min ex len=23*

- *Parameter: df_ids*

  Sequence database IDs to be included in the mapping table. Per default, the mapping table consists of the same type of IDs sent with the request. Available options are: [ensembl, refseq, uniprot, ucsc, hgnc].

  Request example: *request:www.isoaligner.org/api/map?id1=EGFR-201&id2=EGFR-207 &df ids=[ensembl,uniprot]*

- *Parameter: match*

  Needleman-Wunsch alignment parameter to reward matches. This value must be ≥0.

- *Parameter: mismatch*

  Needleman-Wunsch alignment parameter to penalize mismatches. This value must be ≤0.

- *Parameter: open_gap*

  Needleman-Wunsch alignment parameter to penalize opening a gap. This value must be ≤0.

- *Parameter: gap_open*

  Needleman-Wunsch alignment parameter to penalize extending a gap. This value must be ≤0.

With the **resource "align"**, one can align two raw amino acid sequences sent with the request and retrieve a mapping table in json format. The required parameters are seq1 and seq2. All alignment parameters: min ex len, match, mismatch, open gap, gap open are also applicable to this resource.

- *Parameters: seq1 and seq2*

  Reference and alternative raw amino acid sequences. Must be at least 7 AA's long, for example:

  Request: *www.isoaligner.org/api/align?seq1=CRSSWTAAMELSAEYLREKLQRDLEAEHVE&seq2=YLREKLQRDLEAEHVEVEDTTLNRCSCSFRVLVVSAKFEGKPLLQRH*

  Response: mapping table in the json format.

Use cases

IsoAligner helps to transfer annotated mutational data from one transcript to another when working with different isoform database IDs. For example, the identification of the MET p.D1246N and p.D1228N resistance mutations in different transcripts as discussed in the introduction can be easily identified with IsoAligner. First, the user needs to paste one of the transcripts IDs, for example the RefSeq ID NM 000245.4 of the p.D1228N annotation into the "Input" text field at the top of the website (see Figure 2).

Clicking the "Search and Align" button, runs the IsoAligner algorithm and returns simple statistics on the transcript. In this case, there are 8 human isoform entries for the MET gene in the IsoAligner database. The Refseq ID given as the input (NM 000245.4) is automatically mapped to the ensemble transcript name (MET-202). Additional information such as the isoform sequence, various gene attributes and isoform IDs can be found by clicking the "View details about this Isoform Entry" drop-down menu (see Figure 3).

Pairwise sequence alignments between all transcripts and the query transcripts can be found by clicking on the drop-down menu "View Alignment Visualisations" (see Figure 4). The alignments update immediately, when ever any alignment parameters on the left-hand sidebar is adjusted.

Further down, the user can find the "Mapped Amino Acid Positions" table that lists the corresponding amino acid positions between all MET isoforms in the IsoAligner database and the query isoform NM 000245.4. Typing the amino acid position of the resistance mutation 1228 into the "Filter table for exact value" text field and pressing enter shows all mapped amino acid position to position 1228 (see Figure 5). The first listed position is the corresponding amino acid in the canonical transcript MET-201 with the RefSeq ID NM 001127500.3. The last column provides the information that the corresponding amino acid position of 1228 in the canonical transcript is 1246. Note, that a corresponding amino acid to a shorter isoform with a RefSeq ID NM 001324402.2 is also shown, mapping the position 1228 to 798, while the third hit corresponds to a map between position 1210 in the query transcript and 1228 in the canonical transcript.

A "Download Table" button is available to retrieve the entirety of the table in tsv or csv format. Consider clicking the checkbox "Select all columns" to get additional database IDs for the

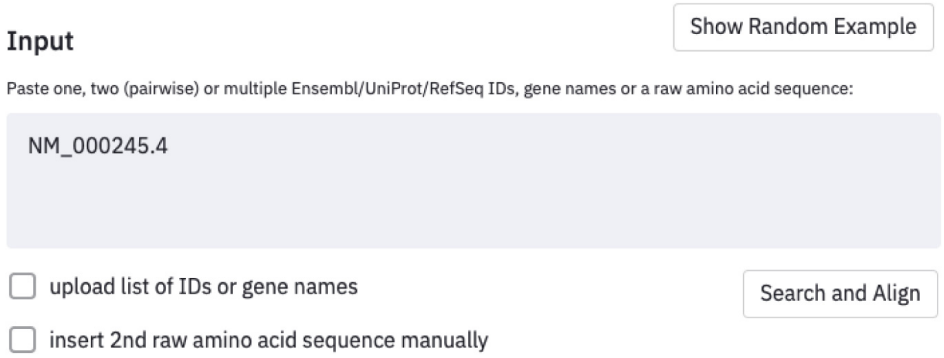


Figure 2. **Example input to investigate the position of MET D1228N on NM 000245.4 across isoforms.**

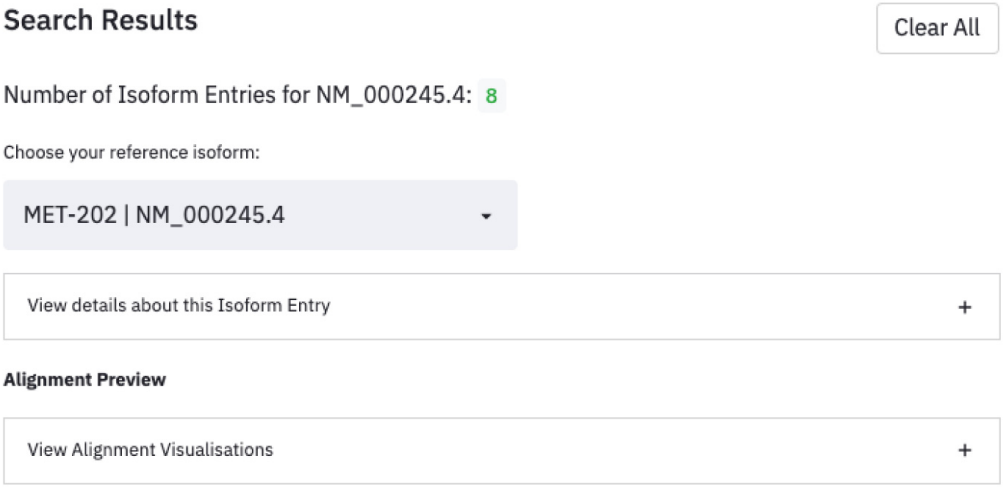


Figure 3. **Various information related to the sequence, gene attributes and isoform IDs can be found in IsoAligner.**

Figure 4. **Pairwise alignment between the query transcript NM 000245.4 and the canonical transcript NM 001127500.3, highlighting the corresponding amino acid positions D1228N and D1246N.**

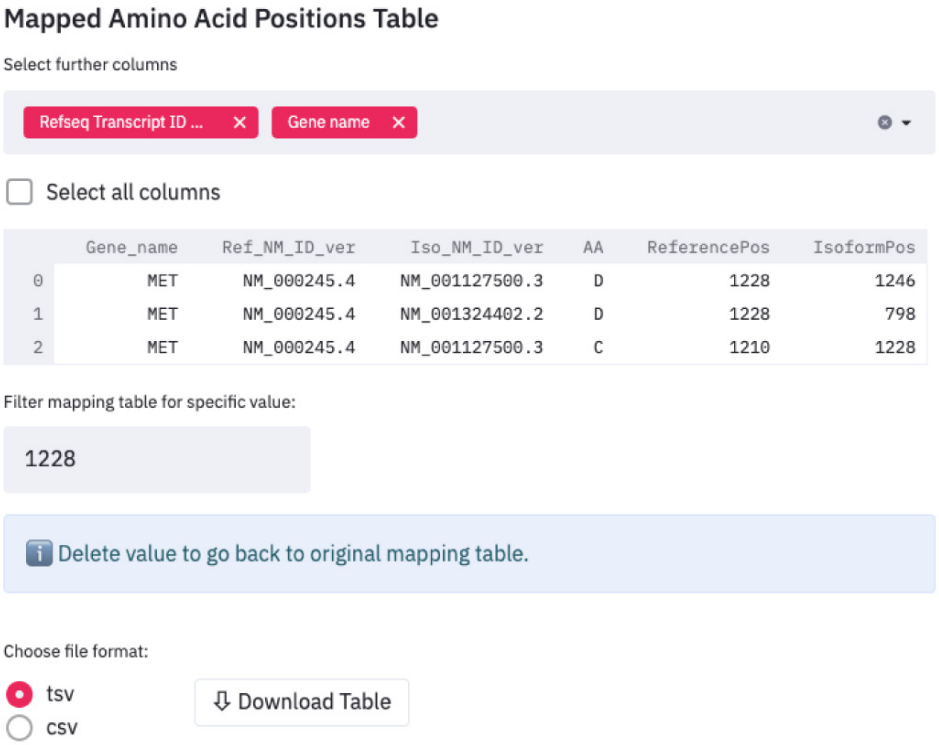


Figure 5. Mapping table listing corresponding amino acid positions between all MET isoforms in the IsoAligner database.

transcripts. Alternatively, users might send an API request https://www.isoaligner.org/api/map?id1=NM_000245.4 to the REST API and retrieve a json file representation of the mapping table.

The data used in this use case can be found as *Underlying data* (Hanimann & Kahraman, 2022).

Conclusion

IsoAligner is a fast and interactive protein isoform alignment webservice that uses a customised Needleman- Wunsch algorithm to specifically align protein alternatively spliced isoforms. The comprehensive library of IsoAligner comprises 1.3 million IDs and 120k protein sequences of 19k human genes which allows rapid positional mapping of amino acids across isoform IDs from Ensembl, RefSeq, UCSC and UniProt.

Data availability

Zenodo: IsoAligner: dynamic mapping of amino acid positions across protein isoforms. https://doi.org/10.5281/zenodo.6354488 (Hanimann & Kahraman, 2022).

This project contains the following underlying data:

- Human Isoform Library Data (the datasets used to generate the Human Isoform Library)
- human isoform library v1.tsv.gz (the pre-computed mapped human isoform library)
- Example Manuscript (the input/output files from the example Use Case section).

Software availability

Webtool available at: https://www.isoaligner.org/

REST API available at: https://www.isoaligner.org/api

Source code available from: https://github.com/mtp-usz/IsoAligner

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6354488 (Hanimann & Kahraman, 2022)

License: CC0-1.0

Acknowledgements

We thank the members of the Clinical Computational Biology group at the University Hospital Zurich for their constant support and valuable inputs.

References

Hanimann J, Kahraman A: **IsoAligner: dynamic mapping of amino acid positions across protein isoforms (IsoAligner v1.2.0).** *Zenodo.* 2022.
http://www.doi.org/10.5281/zenodo.6354488

Howe KL, Achuthan P, Allen J, *et al.*: **Ensembl 2021.** *Nucleic Acids Res.* 2021; **49**(D1): D884–D891.
PubMed Abstract | Publisher Full Text | Free Full Text

Nord A, Carey K, Hornbeck P, *et al.*: **Splice-Aware Multiple Sequence Alignment of Protein Isoforms.** *ACM BCB.* 2018; **2018**: 200–210.
PubMed Abstract | Publisher Full Text | Free Full Text

O'Leary NA, Wright MW, Brister JR, *et al.*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res.* 2016; **44**(D1): D733–45.
PubMed Abstract | Publisher Full Text | Free Full Text

Smedley D, Haider S, Durinck S, *et al.*: **The BioMart community portal: an innovative alternative to large, centralized data repositories.** *Nucleic Acids Res.* 2015; **43**(W1): W589–W598.
PubMed Abstract | Publisher Full Text | Free Full Text

Stephenson JD, Laskowski RA, Nightingale A, *et al.*: **VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations.** *Bioinformatics.* 2019; **35**(22): 4854–4856.
PubMed Abstract | Publisher Full Text | Free Full Text

The Uniprot Consortium: **UniProt: the universal protein knowledgebase in 2021.** *Nucleic Acids Res.* 2021; **49**(D1): D480–D489.
PubMed Abstract | Publisher Full Text | Free Full Text

Tsai JM, Hata AN, Lennerz JK: ***MET* D1228N and D1246N are the Same Resistance Mutation in *MET* Exon 14 Skipping.** *Oncologist.* 2021; **26**(12): e2297–e2301.
PubMed Abstract | Publisher Full Text | Free Full Text

Tweedie S, Braschi B, Gray K, *et al.*: **Genenames.org: the HGNC and VGNC resources in 2021.** *Nucleic Acids Res.* 2021; **49**(D1): D939–D946.
PubMed Abstract | Publisher Full Text | Free Full Text