Using ConSurf to Detect Functionally Important Regions in RNA

Maya Rubin[1] and Nir Ben-Tal[1,2]

[1]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel
[2]Corresponding author: *bental@tauex.tau.ac.il*

The ConSurf web server (*https://consurf.tau.ac.il/*) for using evolutionary data to detect functional regions is useful for analyzing proteins. The analysis is based on the premise that functional regions, which may for example facilitate ligand binding and catalysis, often evolve slowly. The analysis requires finding enough effective, i.e., non-redundant, sufficiently remote homologs. Indeed, the ConSurf pipeline, which is based on state-of-the-art protein sequence databases and analysis tools, is highly valuable for protein analysis. ConSurf also allows evolutionary analysis of RNA, but the analysis often fails due to insufficient data, particularly the inability of the current pipeline to detect enough effective RNA homologs. This is because the RNA search tools and databases offered are not as good as those used for protein analysis. Fortunately, ConSurf also allows importing external collections of homologs in the form of a multiple sequence alignment (MSA). Leveraging this, here we describe various protocols for constructing MSAs for successful ConSurf analysis of RNA queries. We report the level of success of these protocols on an exemplary set comprising a dozen RNA molecules of diverse structure and function. © 2021 Wiley Periodicals LLC.

Basic Protocol 1: Standard ConSurf evolutionary conservation analysis of an RNA query.
Basic Protocol 2: ConSurf evolutionary conservation analysis of an RNA query with external MSA.
Support Protocol 1: Construction of an MSA for an RNA query using other online servers.
Support Protocol 2: Construction of an MSA for an RNA query using nHMMER locally

Keywords: ConSurf • evolutionary analysis • functional regions • Rate4Site • RNA sequence analysis

How to cite this article:
Rubin, M., & Ben-Tal, N. (2021). Using ConSurf to detect functionally important regions in RNA. *Current Protocols*, *1*, e270.
doi: 10.1002/cpz1.270

INTRODUCTION

In the sequences of macromolecules, the evolutionary rate per site, be it an amino acid position in a protein sequence or a nucleotide position in an RNA or DNA sequence, reflects a balance between a natural tendency of the position to mutate, i.e., 'drift', and natural selection. Rarely, the latter may lead to accelerated evolutionary rate, as for example in the ligand-recognition regions of antibodies and other components of our immune

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols e270, Volume 1
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.270

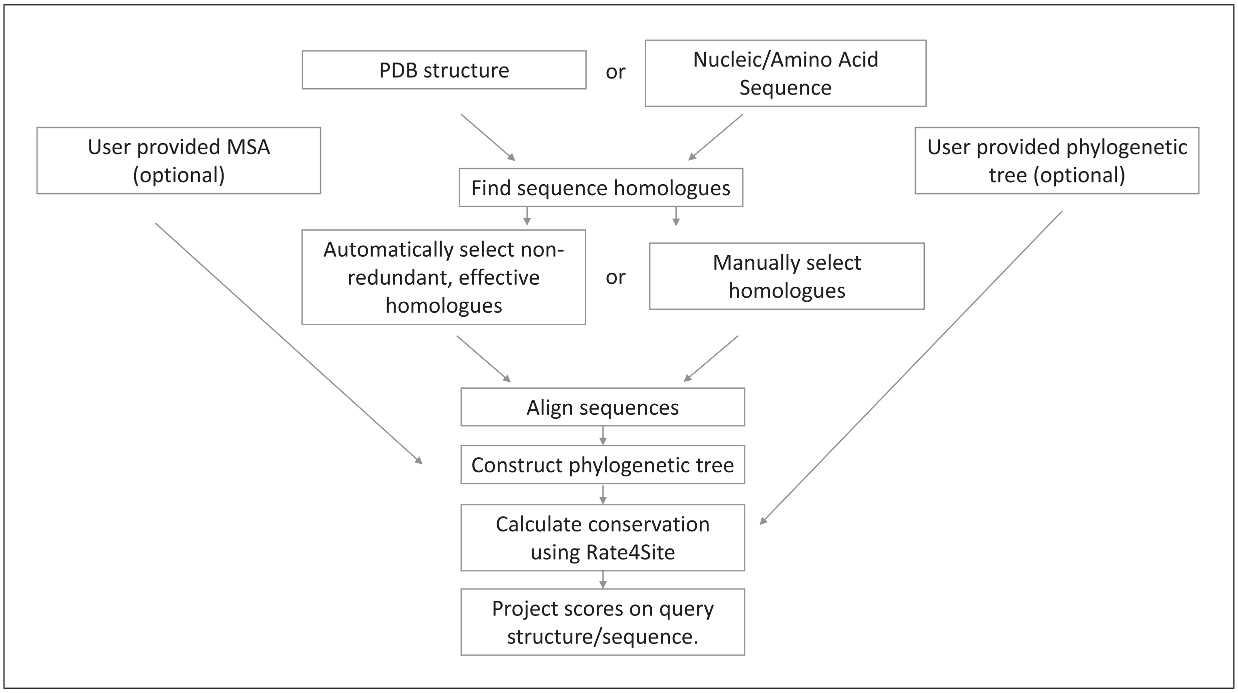

PDB structure
or
Nucleic/Amino Acid Sequence
User provided MSA (optional)
Find sequence homologues
User provided phylogenetic tree (optional)
Automatically select non-redundant, effective homologues
or
Manually select homologues
Align sequences
Construct phylogenetic tree
Calculate conservation using Rate4Site
Project scores on query structure/sequence.

Figure 1 A flowchart of the analysis steps in ConSurf. Calculations can start from the sequence or structure of the query. Here we exploit the possibility to include an external alignment of homologs.

system. However, most often, natural selection limits the evolutionary rates of binding and catalytic sites, as well as of other sites that are biologically important. Thus, a slow evolutionary rate is often a clear mark of functionally important regions in protein, DNA, and RNA molecules (Capra & Singh, 2007; Del Sol, Pazos, & Valencia, 2003; Gallet, Charloteaux, Thomas, & Brasseur, 2000; Huang & Golding, 2014, 2015; Innis, 2007; Landgraf, Xenarios, & Eisenberg, 2001; Lichtarge, Bourne, & Cohen, 1996a, 1996b; Lichtarge, Yamamoto, & Cohen, 1997; Mayrose, Graur, Ben-Tal, & Pupko, 2004; Valdar, 2002). ConSurf provides a reliable and easy-to-use way to exploit this principle (Ashkenazy et al., 2016; Ashkenazy, Erez, Martz, Pupko, & Ben-Tal, 2010; Celniker et al., 2013; Mayrose et al., 2004). Starting from the user-provided sequence or structure of a query protein/RNA/DNA, ConSurf automatically collects a set of effective homologs, aligns their sequences, builds a phylogenetic tree that represents their evolutionary relationships, and estimates the evolutionary rates of the amino acid or nucleotide positions using a statistically robust evolutionary model. An outline of the ConSurf pipeline is shown in Figure 1.

While ConSurf offers a pipeline for analyzing both proteins and RNA/DNA molecules, it is most used in protein analysis and rarely with nucleotides. This is presumably because the nucleic acid analysis pipeline offered by ConSurf is frequently aborted because of failure to detect a large enough set of effective homologs to the query.

Here, we show how to improve the analysis of an RNA query by utilizing state-of-the-art sequence search tools in combination with the ConSurf pipeline. Basic Protocol 1 details an analysis that utilizes the MSA construction of ConSurf itself. This protocol, which often fails, is used mostly as a reference. Basic Protocol 2, the recommended alternative, details an analysis based on an externally constructed MSA. The Support Protocols provide guidance on constructing an MSA for an RNA query to be used with Basic Protocol 2.

STANDARD CONSURF EVOLUTIONARY CONSERVATION ANALYSIS OF AN RNA QUERY

BASIC PROTOCOL 1

This protocol provides guidance on using the ConSurf server to analyze the evolutionary conservation profile of an RNA query, given its 3D structure or nucleotide sequence.

Necessary Resources

Hardware

Computer with Internet connection, under Windows, Mac, or Linux

Software (recommended)

The PyMOL (Schrödinger, 2021), Chimera (Pettersen et al., 2004), or RasMol (Sayle & Milner-White, 1995) molecular visualizer

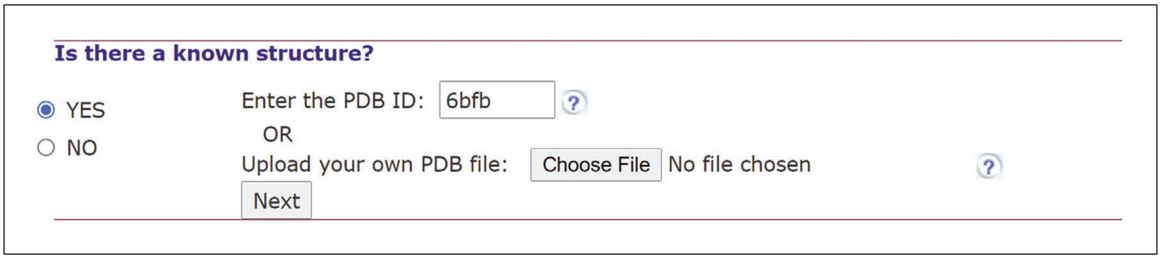
1. Upload RNA query.

   Enter the ConSurf server web page (https://consurf.tau.ac.il/) and select the Nucleotides option. You will be redirected to a page asking if your query has a known structure. Selecting 'YES' will allow you to enter the PDB ID of the molecule (Fig. 2); press "Next" and indicate the chain of interest if needed (Fig. 3). Alternatively, you may upload the coordinate file, in PDB format. If 'NO' is selected the server will proceed to ask for an MSA. Selecting 'NO' again will allow you to enter the query sequence, in FASTA format (Fig. 4).

2. Select setting for the construction of a multiple sequence alignment (MSA). See Figure 5.

   The server will ask if you wish to upload an MSA. Select NO, at which point ConSurf will allow you to select the homology search method and nucleotide database, as well as other parameters for generating an MSA:

   Homolog search algorithm*–A choice between nBLAST (default) (Altschul, Gish, Miller, Myers, & Lipman, 1990; https://blast.ncbi.nlm.nih.gov/Blast.cgi) and nHMMER (Eddy, 2009; http://hmmer.org/).*



Is there a known structure?

● YES

○ NO

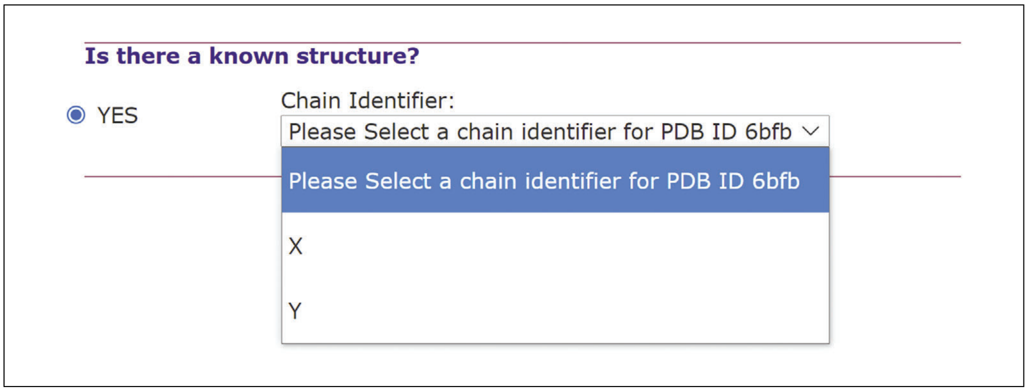
Enter the PDB ID: 6bfb

OR

Upload your own PDB file: Choose File No file chosen

Next

Figure 2 Entering the PDB ID of the structure of the RNA query.



Is there a known structure?

● YES

Chain Identifier:

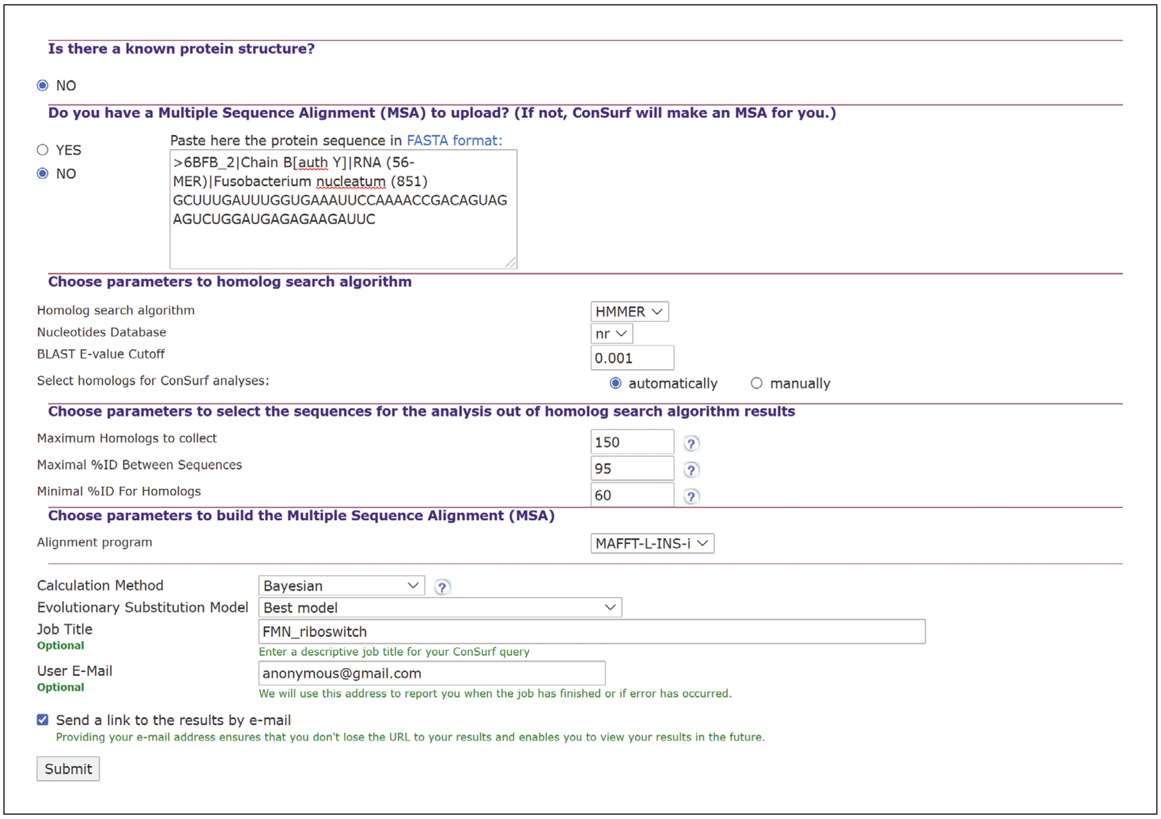
Please Select a chain identifier for PDB ID 6bfb

Please Select a chain identifier for PDB ID 6bfb

X

Y

Figure 3 Selecting the query RNA chain.



Is there a known protein structure?

● NO

Do you have a Multiple Sequence Alignment (MSA) to upload? (If not, ConSurf will make an MSA for you.)

○ YES
● NO

Paste here the protein sequence in FASTA format:

```
>6BFB_2|Chain B[auth Y]|RNA (56-
MER)|Fusobacterium nucleatum (851)
GCUUUGAUUUGGUGAAAUUCCAAAACCGACAGUAG
AGUCUGGAUGAGAGAAGAUUC
```

Choose parameters to homolog search algorithm

Homolog search algorithm	HMMER
Nucleotides Database	nr
BLAST E-value Cutoff	0.001
Select homologs for ConSurf analyses:	● automatically ○ manually

Choose parameters to select the sequences for the analysis out of homolog search algorithm results

Maximum Homologs to collect	150
Maximal %ID Between Sequences	95
Minimal %ID For Homologs	60

Choose parameters to build the Multiple Sequence Alignment (MSA)

Alignment program	MAFFT-L-INS-i

Calculation Method	Bayesian
Evolutionary Substitution Model	Best model
Job Title Optional	FMN_riboswitch Enter a descriptive job title for your ConSurf query
User E-Mail Optional	anonymous@gmail.com We will use this address to report you when the job has finished or if error has occurred.

☑ Send a link to the results by e-mail
Providing your e-mail address ensures that you don't lose the URL to your results and enables you to view your results in the future.

Figure 4 Entering the query sequence in FASTA format.

Nucleotides Database–*Currently there is only one nucleotide database that can be searched, nr NCBI Nucleotide Home page; see Internet Resources).*

BLAST E-value Cutoff–*This value is set at 0.001 by default and can be increased up to <1. Increasing the value may help in finding more homologs (but also potentially non-homologs).*

Select homologs for ConSurf analyses–*There are two options to choose from, "automatically" or "manually." Selecting "manually" will eventually require the user to mark sequences from the hits list, which could be valuable for users who are very familiar with their query. We highly recommend selecting "automatically."*

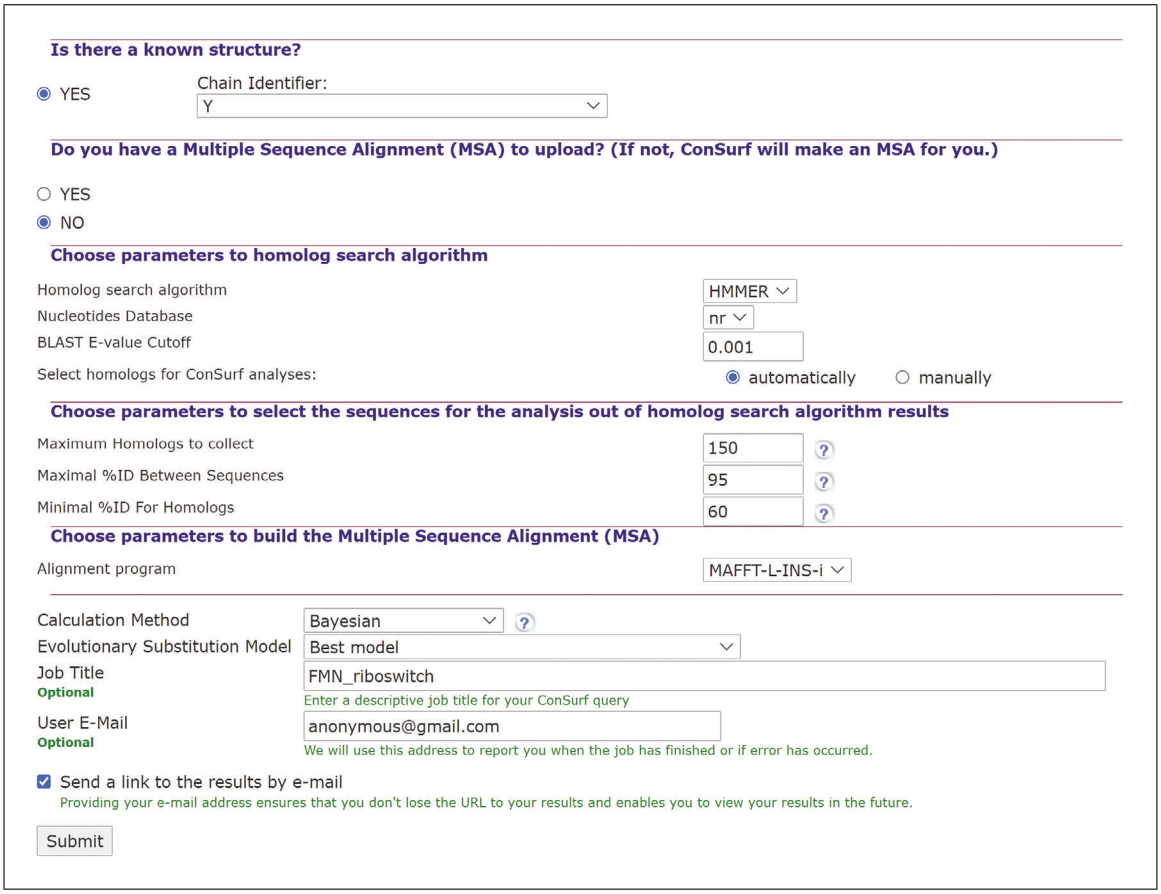
- *"automatically": The user is asked to indicate the maximum number of homologs (150 is the default; selecting more than 300 would significantly slow the calculation), as well as the maximum and minimum sequence ID percentages (95 and 60 by default). In ConSurf, the hits (coming from the nBLAST or nHMMER search) are sorted by their E-values in an ascending order, based on the principle that the lower the E-value, the more likely the hit is to be a true homolog. When selecting "automatically," a predetermined number of hits are sampled evenly from the sorted list to create the final list of homologs of the query protein. The user is also asked to choose between three methods for multiply align the selected homologs: MAFFT-L-INS-i (default), PRANK, or CLUSTALW.*

3. Selecting analysis methods (Fig. 5)

   The Calculation Method can be Bayesian (Default) or Maximum Likelihood. For the Evolutionary Substitution Model there is a choice between "T92 model (Tamura, 1992)," "GTR: General Time Reversible," "JC69 model (Juke & Cantor, 1969)," "HKY85 model (Hasegawa, Kishino, & Yano, 1985)," and "Best model" (default).

   We recommend using the default settings.

4. Run job entry (Fig. 5).



Is there a known structure?

● YES   Chain Identifier: Y

Do you have a Multiple Sequence Alignment (MSA) to upload? (If not, ConSurf will make an MSA for you.)

○ YES

● NO

Choose parameters to homolog search algorithm

Homolog search algorithm	HMMER
Nucleotides Database	nr
BLAST E-value Cutoff	0.001
Select homologs for ConSurf analyses:	● automatically ○ manually

Choose parameters to select the sequences for the analysis out of homolog search algorithm results

Maximum Homologs to collect	150
Maximal %ID Between Sequences	95
Minimal %ID For Homologs	60

Choose parameters to build the Multiple Sequence Alignment (MSA)

Alignment program	MAFFT-L-INS-i

Calculation Method	Bayesian
Evolutionary Substitution Model	Best model
Job Title Optional	FMN_riboswitch Enter a descriptive job title for your ConSurf query
User E-Mail Optional	anonymous@gmail.com We will use this address to report you when the job has finished or if error has occurred.

☑ Send a link to the results by e-mail

Providing your e-mail address ensures that you don't lose the URL to your results and enables you to view your results in the future.

Figure 5 An analysis page of an RNA query of known structure ready to be submitted. Default settings are used in all fields, and job title and e-mail address are added, as recommended.

There is an option to give a job title as well as an e-mail address. The latter is particularly convenient because, if used, an e-mail will be sent with a link to the results once the run has finished. Each job is given a number, regardless of whether a job title was entered, and the results are kept on the server for up to 3 months.

CONSURF EVOLUTIONARY CONSERVATION ANALYSIS OF AN RNA QUERY WITH EXTERNAL MSA

BASIC PROTOCOL 2

This protocol will provide guidance to using the ConSurf server to analyze the evolutionary conservation profile of an RNA query, given its 3D structure or nucleotide sequence, using an externally provided MSA. Two Support Protocols for constructing MSA for the query are provided further below.

Necessary Resources

Hardware

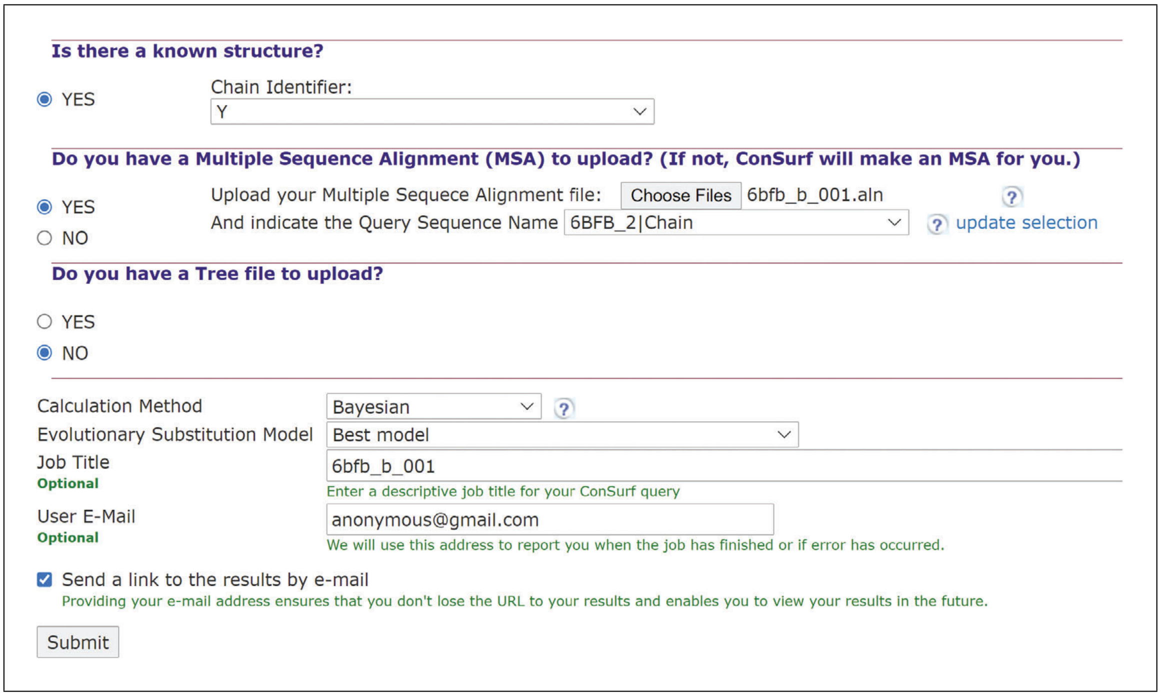
Computer with Internet connection, under Windows, Mac, or Linux.

Software (recommended)

The PyMOL, Chimera or RasMol molecular visualizers

1. Upload RNA query.

Enter the ConSurf server web page (https://consurf.tau.ac.il/) and select the Nucleotides option. You will be redirected to a page asking if your query has a known structure. Selecting 'YES' will allow you to enter the PDB ID of the molecule (Fig. 2); press "Next" and indicate the chain of interest (Fig. 3). Alternatively, you may upload the coordinate file, in PDB format. If 'NOTE' is selected, the server will proceed to ask for MSA.



Is there a known structure?

● YES

Chain Identifier: Y

Do you have a Multiple Sequence Alignment (MSA) to upload? (If not, ConSurf will make an MSA for you.)

● YES
○ NO

Upload your Multiple Sequece Alignment file: Choose Files 6bfb_b_001.aln ?

And indicate the Query Sequence Name 6BFB_2|Chain ? update selection

Do you have a Tree file to upload?

○ YES
● NO

Calculation Method: Bayesian ?

Evolutionary Substitution Model: Best model

Job Title (Optional): 6bfb_b_001
Enter a descriptive job title for your ConSurf query

User E-Mail (Optional): anonymous@gmail.com
We will use this address to report you when the job has finished or if error has occurred.

☑ Send a link to the results by e-mail
Providing your e-mail address ensures that you don't lose the URL to your results and enables you to view your results in the future.

Submit

Figure 6 Uploading an external MSA to an RNA query of known structure.

2. Upload a multiple sequence alignment (MSA).

   The server will ask if you wish to upload an MSA. Select YES and choose the MSA file from your computer. Once uploaded, choose your query sequence from the "Indicate the Query Sequence Name" menu and click on "update selection" to continue (Fig. 6).

   Make sure that the format is one supported by ConSurf: NBRF/PIR, EMBL/SwissProt, Pearson (FASTA), GDE, Clustal, GCG/MSF, and RSF format. Additionally, in all the sequences, U must be replaced with T for compatibility with the server. Also, the name of the file uploaded must only contain dashes and underscores as gap symbols, no whitespaces.

   In the Support Protocols provided, you will find two different ways of constructing an MSA for an RNA query which can then be uploaded to ConSurf.

3. Tree upload and analysis methods.

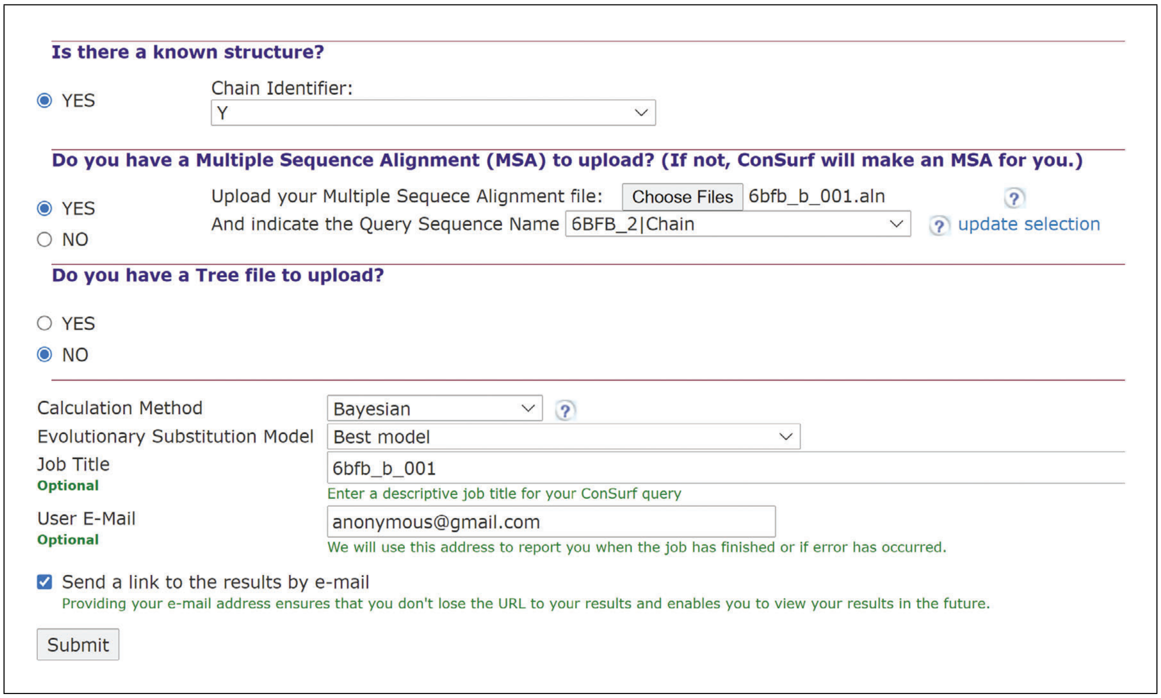
   Select NO when asked for a tree upload, at which point the server will provide the options to select the parameters for the evolutionary analysis (Fig. 7).

   The Calculation Method can be Bayesian (Default) or Maximum Likelihood (Pupko, Bell, Mayrose, Glaser, & Ben-Tal, 2002). For the Evolutionary Substitution Model, there is a choice between T92 model (Tamura, 1992), GTR: General Time Reversible (Tavare, 1986), JC69 model (Jukes & Cantor, 1969), HKY85 model (Hasegawa et al., 1985), and "Best model" (default).

   We recommend using the default settings.

4. Run job entry.

   There is an option to give a job title as well as an e-mail address (Fig. 7). The latter is particularly convenient because, if used, an e-mail will be sent with a link to the results once the run has finished. Each job is given a number, regardless of whether or not a job title was entered, and the results are kept on the server for up to 3 months.



Is there a known structure?

● YES Chain Identifier: Y

Do you have a Multiple Sequence Alignment (MSA) to upload? (If not, ConSurf will make an MSA for you.)

● YES Upload your Multiple Sequece Alignment file: Choose Files 6bfb_b_001.aln ?

○ NO And indicate the Query Sequence Name 6BFB_2|Chain ? update selection

Do you have a Tree file to upload?

○ YES

● NO

Calculation Method: Bayesian ?

Evolutionary Substitution Model: Best model

Job Title *Optional*: 6bfb_b_001

Enter a descriptive job title for your ConSurf query

User E-Mail *Optional*: anonymous@gmail.com

We will use this address to report you when the job has finished or if error has occurred.

☑ Send a link to the results by e-mail

Providing your e-mail address ensures that you don't lose the URL to your results and enables you to view your results in the future.

Submit

Figure 7 An analysis page with an external MSA ready to be submitted. Default settings are used in all fields, and job title and E-mail address are added, as recommended.

SUPPORT PROTOCOL 1

CONSTRUCTION OF AN MSA FOR AN RNA QUERY USING OTHER ONLINE SERVERS

Starting from a nucleotide query, ConSurf currently provides only searches in the NCBI NT database (marked as nr for non-redundant; NCBI Resource Coordinators, 2018). Unfortunately, this limited search often does not yield a large enough set of effective homologs. To overcome this obstacle, it is recommended to provide an external MSA for your RNA query. The following is a description of a protocol to construct a ConSurf-compatible MSA.

Necessary Resources

Hardware

Computer with Internet connection, under Windows, Mac, or Linux

Software

Notepad++ or any other similar text editor

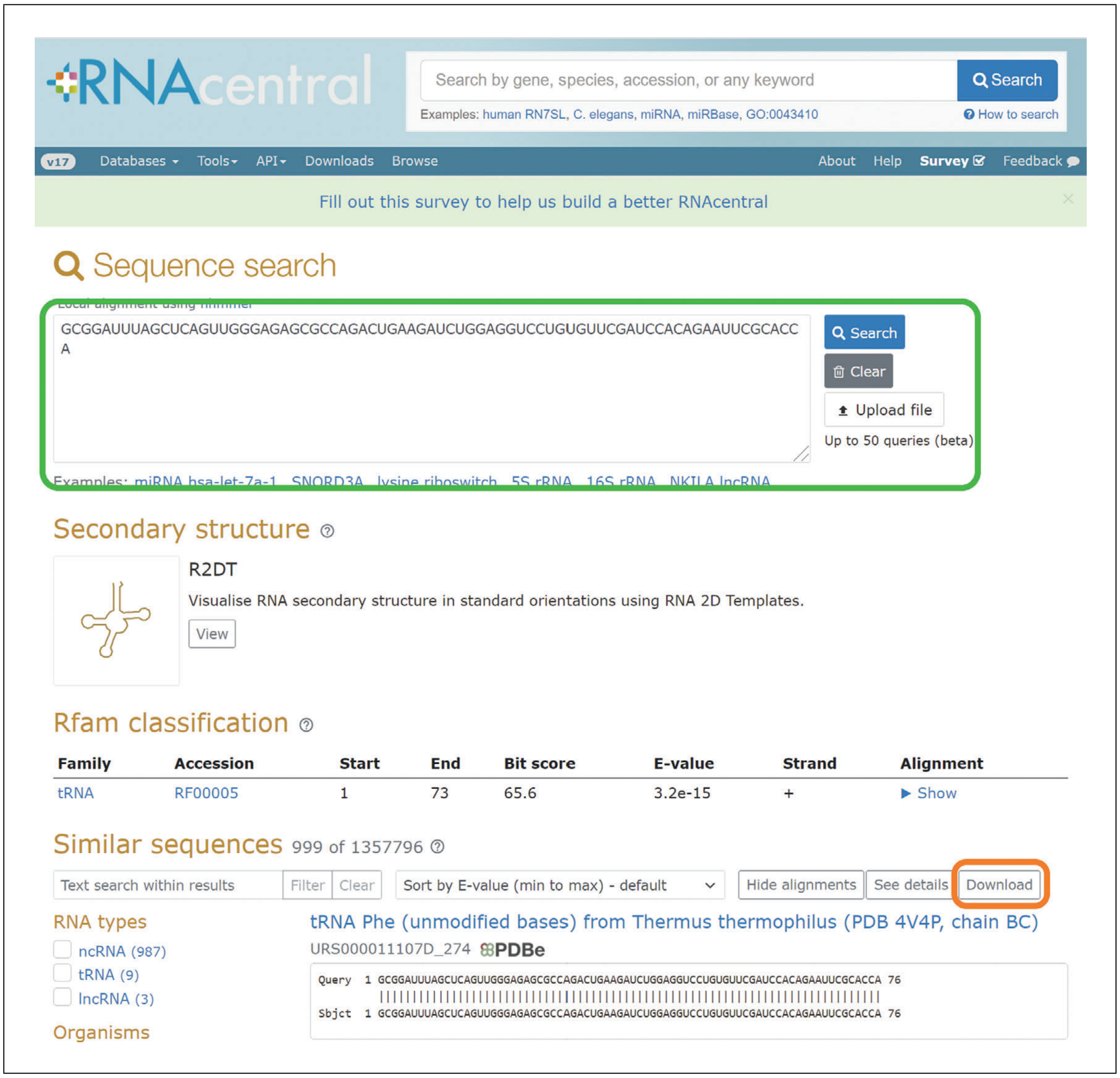
1. Enter RNACentral (*https://rnacentral.org/*), select the "Sequence search" option, enter your RNA query sequence into the search box, and hit the Search button (Fig. 8).

   The RNACentral search engine utilizes nHMMER to search a much larger database than the one ConSurf searches (RNAcentral Consortium, 2021; Wheeler & Eddy, 2013).

   The results, up to one thousand hits that are the most similar to your query, can be downloaded from the server.

2. Download the results from RNACentral by pressing the "Download" button at the top of the search results.

   A zip folder will be downloaded to your computer. When opened, there will be a folder titled "sequences" which contains three files. One of the files is a FASTA file of all the similar sequences in FASTA format, called `similar-sequences.fasta`*. Extract this file to a path of your choosing.*


RNAcentral
Search by gene, species, accession, or any keyword
Search
Examples: human RN7SL, C. elegans, miRNA, miRBase, GO:0043410
How to search
v17
Databases
Tools
API
Downloads
Browse
About
Help
Survey
Feedback
Fill out this survey to help us build a better RNAcentral
Sequence search
Local alignment using nhmmer
Search
Clear
Upload file
Up to 50 queries (beta)
Examples: miRNA hsa-let-7a-1 SNORD3A lysine riboswitch 5S rRNA 16S rRNA NKILA lncRNA
Secondary structure
R2DT
Visualise RNA secondary structure in standard orientations using RNA 2D Templates.
View
Rfam classification
Similar sequences 999 of 1357796
Text search within results
Filter
Clear
Sort by E-value (min to max) - default
Hide alignments
See details
Download
RNA types
ncRNA (987)
tRNA (9)
lncRNA (3)
Organisms
tRNA Phe (unmodified bases) from Thermus thermophilus (PDB 4V4P, chain BC)
URS000011107D_274 PDBe

```
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACC
A
```

Family	Accession	Start	End	Bit score	E-value	Strand	Alignment
tRNA	RF00005	1	73	65.6	3.2e-15	+	▶ Show

```
Query  1 GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 76
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1 GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 76
```

Figure 8 The sequence search page of RNACentral, with the top of the results for FMN riboswitch (PDB ID 6BFB, chain y). The search box with a query sequence is highlighted in green (appears the same before and after search). The download button for the results is highlighted in orange.

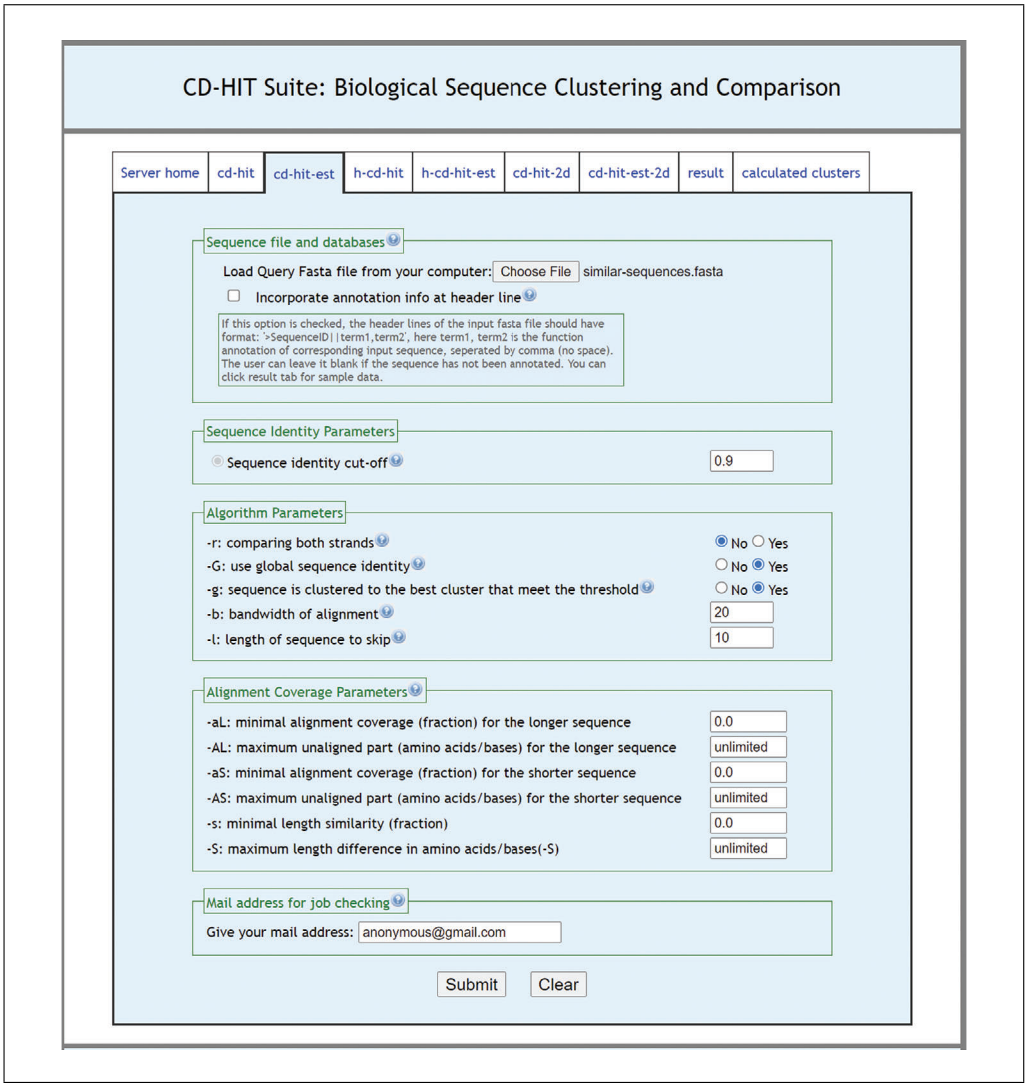
3. Adjust the sequences for compatibility with ConSurf.

   ConSurf can only work with DNA nucleobases, i.e., A, C, G, and T. When constructing an RNA MSA, ConSurf replaces all U characters in the sequences with the character T, but when constructing an MSA externally, this adjustment needs to be done manually. To do so, open the FASTA file you downloaded from RNACentral using Notepad++ (or a similar text editor), replace all U's with T's, and save the changes.

4. Cluster the sequences using cd-hit-est (*http://weizhong-lab.ucsd.edu/cdhit-web-server/cgi-bin/index.cgi?cmd=cd-hit-est*); see Fig. 9.

   cd-hit-est is a tool that clusters nucleotide sequences based on a sequence identity threshold set by the user (Fu, Niu, Zhu, Wu, & Li, 2012; Huang, Niu, Gao, Fu, & Li, 2010; Li & Godzik, 2006). The idea is to reduce the number of homologs in the dataset while keeping the integrity of the sequence diversity in the dataset, i.e., to produce a list of "effective" homologs.

   Increasing the threshold increases the number of representatives. We recommend setting the sequence identity threshold between 0.95 and 0.8. For the best results, aim for the highest threshold that generates at most 300 effective homologs. ConSurf needs at least



CD-HIT Suite: Biological Sequence Clustering and Comparison

Sequence file and databases

Load Query Fasta file from your computer: similar-sequences.fasta

☐ Incorporate annotation info at header line

If this option is checked, the header lines of the input fasta file should have format: '>SequenceID||term1,term2', here term1, term2 is the function annotation of corresponding input sequence, seperated by comma (no space). The user can leave it blank if the sequence has not been annotated. You can click result tab for sample data.

Sequence Identity Parameters

○ Sequence identity cut-off: 0.9

Algorithm Parameters

-r: comparing both strands: ● No ○ Yes

-G: use global sequence identity: ○ No ● Yes

-g: sequence is clustered to the best cluster that meet the threshold: ○ No ● Yes

-b: bandwidth of alignment: 20

-l: length of sequence to skip: 10

Alignment Coverage Parameters

-aL: minimal alignment coverage (fraction) for the longer sequence: 0.0

-AL: maximum unaligned part (amino acids/bases) for the longer sequence: unlimited

-aS: minimal alignment coverage (fraction) for the shorter sequence: 0.0

-AS: maximum unaligned part (amino acids/bases) for the shorter sequence: unlimited

-s: minimal length similarity (fraction): 0.0

-S: maximum length difference in amino acids/bases(-S): unlimited

Mail address for job checking

Give your mail address: anonymous@gmail.com

Figure 9 CD-HIT run page with sequence identity cut-off set at 0.9 (90%) and default settings left in all other fields.

5 homologs, and when the number of homologs exceeds 300, the run time significantly increases.

Upload the homologs file produced in the previous step and leave "Incorporate annotations info at header line" unchecked. Change the "Sequence identity cut-off" if needed (default is 0.9), enter an e-mail address (optional but highly recommended), and submit the job, leaving the rest of the parameters on the default settings.

5. Check the resulting representative sequences for your query.

In the results page, open the FASTA file for representative sequences at X% identity (X being the percentage you entered in the search) and check for your query sequence. If included, the query should appear at the top of the list, but it might be excluded from the list of effective homologs. Select all the sequences and copy them.

6. Align the sequence dataset using MAFFT *(https://mafft.cbrc.jp/alignment/server/)*.

Open the MAFFT alignment server (Katoh, Rozewicki, & Yamada, 2017) and paste the sequences copied from the CD-HIT results. The headers of the sequences will be edited by MAFFT, leaving only the number code at the beginning of the header. It is therefore recommended to change the header of the query sequence to a short name that you would easily recognize, e.g., `myquery`. *If the query was excluded from the list, add it manually* *to the top of the list (again, making sure to replace all Us with Ts) with a header you will recognize.*

Enter a job title and an e-mail address (optional but highly recommended) and submit the job, leaving the rest of the parameters on the default settings.

7. Download MSA in Clustal format.

At the top of the results page of the MAFFT alignment, there is a "Clustal format" link. Press it to download a file of the alignment in Clustal format.

SUPPORT PROTOCOL 2

CONSTRUCTION OF AN MSA FOR AN RNA QUERY USING NHMMER LOCALLY

The RNACentral search engine for detecting RNA homologs follows a single rigid pipeline. Alternatively, for more flexible and comprehensive homology search it is possible to install and use nHMMER locally. This advanced option is meant for experts.

Necessary Resources

Hardware

Computer with Internet connection, under Windows, Mac, or Linux

Software

HMMER, can be downloaded from the HMMER site (*http://hmmer.org/*) and run on a Linux-based platform
Notepad++, or any other similar text editor

Files

RNA database in FASTA format downloaded from RNAcentral (*http://ftp.ebi.ac.uk/pub/databases/RNAcentral/current_release/sequences/rnacentral_active.fasta.gz*)
A file of the query sequence in FASTA format

1. HMMER installation.

If you do not have HMMER on your computer, download it from the HMMER web page (http://hmmer.org/). Mac and Linux users can install it following the instructions in the "User Guide," which is included in the download folder. For Windows users, a Linux subsystem for windows (LSW) is required. Cygwin is one such LSW that you can install by following the instructional video at https://www.youtube.com/watch?v=MBtbgZ7OmNM&ab_channel=ebioinformaticsYoutubeChannel (this guide is applicable for windows 10 as well as Windows 7). Once installed, follow the instructions in the "User Guide" to install HMMER on the LSW.

2. Search for homologs using nHMMER locally.

Run the query sequence against an RNA database using the following Synopsis:

nhmmer `--rna -E <x> --incE <x'> -A <f> queryfile seqdb`

`--rna` *asserts that the sequences are RNA.*
`-E <x>` *Target sequences with an E-value of <= <x> will be reported; the default is 10.0.*
`--incE <x'>` *Use an E-value of <= <x'> as the inclusion threshold, the default being 0.01.*
`-A <f>` *Save a multiple alignment of all hits that satisfy the inclusion thresholds to the file <f>.*
`Queryfile` *is the file with the query sequence, including the path if needed.*

`Seqdb` *is the database of sequences from RNACentral, including the path if needed.*

3. Convert the alignment file from Stockholm to FASTA format.

   *Using the HMMER miniapp **esl-reformat**, the aligned output can be converted to a dataset in FASTA format, which is required for the next steps.*

   Synopsis:

   esl-reformat `-o <f> fasta seqfile`
   `-o <f>` *saves the output to file* `<f>`.
   `Seqfile` *is the aligned output file from the previous step.*

4. Adjust the sequences for compatibility with ConSurf.

   ConSurf can only work with DNA nucleobases, i.e., A, C, G, and T. When constructing an RNA MSA, ConSurf replaces all U characters in the sequences with the character T, but when constructing an MSA externally, this adjustment needs to be done manually. To do so, open the FASTA file `<f>` *using Notepad++ (or a similar text editor), replace all U's with T's, and save the changes.*

5. Cluster the sequences using cd-hit-est (*http://weizhong-lab.ucsd.edu/cdhit-web-server/cgi-bin/index.cgi?cmd=h-cd-hit*; see Fig. 9).

   cd-hit-est is a tool that clusters nucleotide sequences based on a sequence identity threshold set by the user. The idea is to reduce the number of homologs in the dataset while keeping the integrity of the sequence diversity in the dataset, i.e., to produce a list of "effective" homologs.

   Increasing the threshold increases the number of representatives. We recommend setting the sequence identity threshold between 0.95 and 0.8. For the best results, aim for the highest threshold that generates at most 300 effective homologs. ConSurf needs at least 5 homologs, and when the number of homologs exceeds 300, the run time significantly increases.

   Upload the homologs file produced in the previous step and leave "Incorporate annotations info at header line" unchecked. Change the "Sequence identity cut-off" if needed (default is 0.9), enter an e-mail address (optional but highly recommended), and submit the job, leaving the rest of the parameters on the default settings.

6. Check the resulting representative sequences for your query.

   In the results page, open the FASTA file for representative sequences at X% identity (X being the percentage you entered in the search) and check for your query sequence. If included, the query should appear at the top of the list, but it might be excluded from the list of effective homologs. Select all the sequences and copy them.

7. Align the sequence dataset using MAFFT *(https://mafft.cbrc.jp/alignment/server/)*.

   Open the MAFFT alignment server and paste the sequences copied from the CD-HIT results. The headers of the sequences will be edited by MAFFT, leaving only the number code at the beginning of the header. It is therefore recommended to change the header of the query sequence to a short name that you would easily recognize. If the query was excluded from the list, add it manually to the top of the list (again, making sure to replace all U with T) with a header you will recognize.

   Enter a job title and an e-mail address (optional but highly recommended) and submit the job, leaving the rest of the parameters on the default settings.

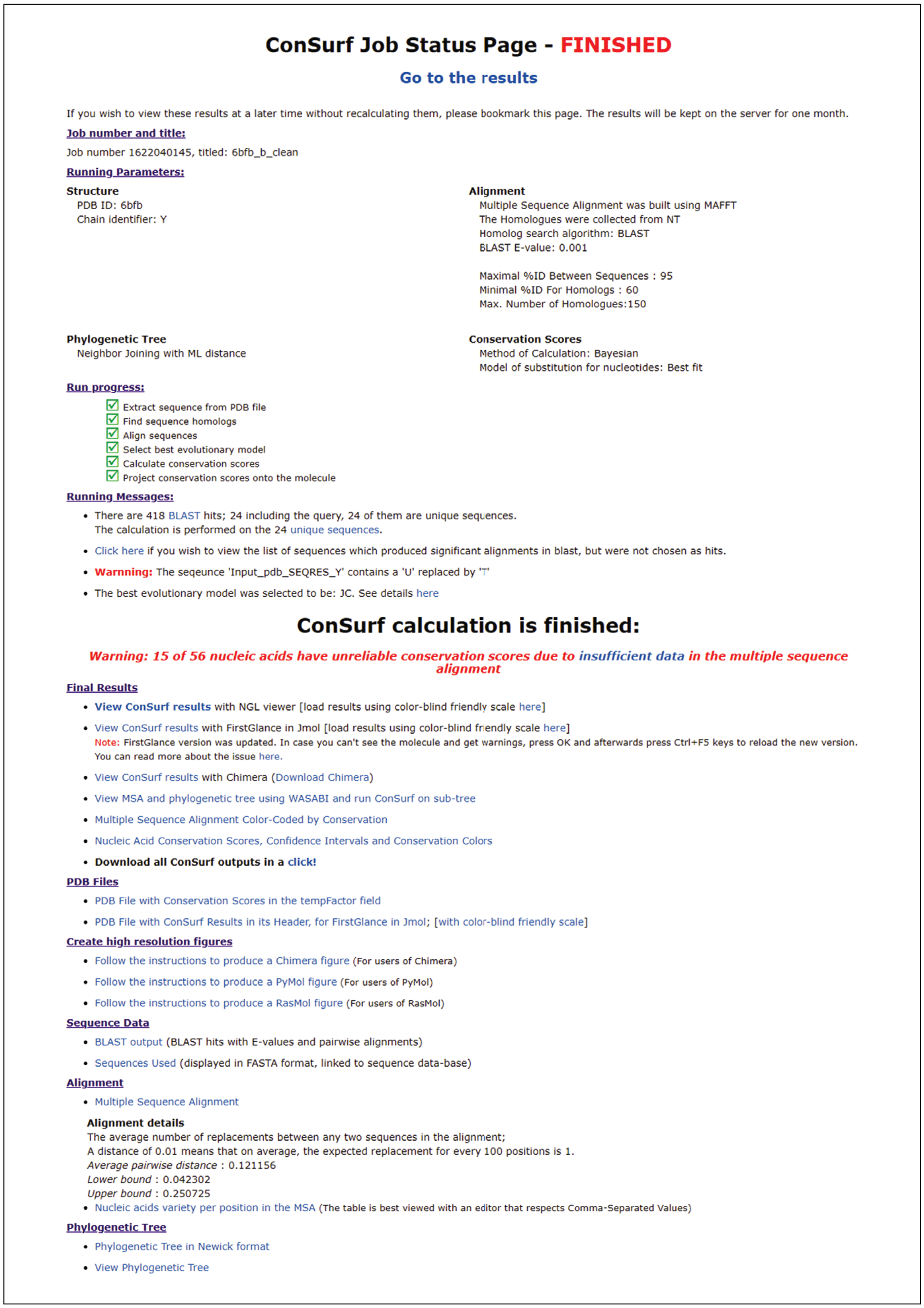
8. Download MSA in Clustal format.

   At the top of the results page of the MAFFT alignment, there is a "Clustal format" link. Press it to download a file of the alignment in clustal format.

GUIDELINES FOR UNDERSTANDING RESULTS

The results page of your ConSurf run will indicate the current stage of the analysis. Once complete, the job status at the top of the page will indicate in red “FINISHED” if the run completed successfully (Fig. 10), or “FAILED” if it did not (Fig. 11).

The parameters of the run will be detailed bellow the Job status, followed by a “Run progress” checklist and “Running massages.” In the case of a failed run, an error massage explaining the issue will be included.



ConSurf Job Status Page - FINISHED

Go to the results

If you wish to view these results at a later time without recalculating them, please bookmark this page. The results will be kept on the server for one month.

Job number and title:

Job number 1622040145, titled: 6bfb_b_clean

Running Parameters:

Structure
PDB ID: 6bfb
Chain identifier: Y

Phylogenetic Tree
Neighbor Joining with ML distance

Alignment
Multiple Sequence Alignment was built using MAFFT
The Homologues were collected from NT
Homolog search algorithm: BLAST
BLAST E-value: 0.001

Maximal %ID Between Sequences : 95
Minimal %ID For Homologs : 60
Max. Number of Homologues:150

Conservation Scores
Method of Calculation: Bayesian
Model of substitution for nucleotides: Best fit

Run progress:

☑ Extract sequence from PDB file
☑ Find sequence homologs
☑ Align sequences
☑ Select best evolutionary model
☑ Calculate conservation scores
☑ Project conservation scores onto the molecule

Running Messages:

- There are 418 BLAST hits; 24 including the query, 24 of them are unique seqυences.
  The calculation is performed on the 24 unique sequences.
- Click here if you wish to view the list of sequences which produced significant alignments in blast, but were not chosen as hits.
- **Warnning:** The seqeunce 'Input_pdb_SEQRES_Y' contains a 'U' replaced by 'T'
- The best evolutionary model was selected to be: JC. See details here

ConSurf calculation is finished:

Warning: 15 of 56 nucleic acids have unreliable conservation scores due to insufficient data in the multiple sequence alignment

Final Results

- **View ConSurf results** with NGL viewer [load results using color-blind friendly scale here]
- View ConSurf results with FirstGlance in Jmol [load results using color-blind friendly scale here]
  Note: FirstGlance version was updated. In case you can't see the molecule and get warnings, press OK and afterwards press Ctrl+F5 keys to reload the new version.
  You can read more about the issue here.
- View ConSurf results with Chimera (Download Chimera)
- View MSA and phylogenetic tree using WASABI and run ConSurf on sub-tree
- Multiple Sequence Alignment Color-Coded by Conservation
- Nucleic Acid Conservation Scores, Confidence Intervals and Conservation Colors
- **Download all ConSurf outputs in a click!**

PDB Files

- PDB File with Conservation Scores in the tempFactor field
- PDB File with ConSurf Results in its Header, for FirstGlance in Jmol; [with color-blind friendly scale]

Create high resolution figures

- Follow the instructions to produce a Chimera figure (For users of Chimera)
- Follow the instructions to produce a PyMol figure (For users of PyMol)
- Follow the instructions to produce a RasMol figure (For users of RasMol)

Sequence Data

- BLAST output (BLAST hits with E-values and pairwise alignments)
- Sequences Used (displayed in FASTA format, linked to sequence data-base)

Alignment

- Multiple Sequence Alignment

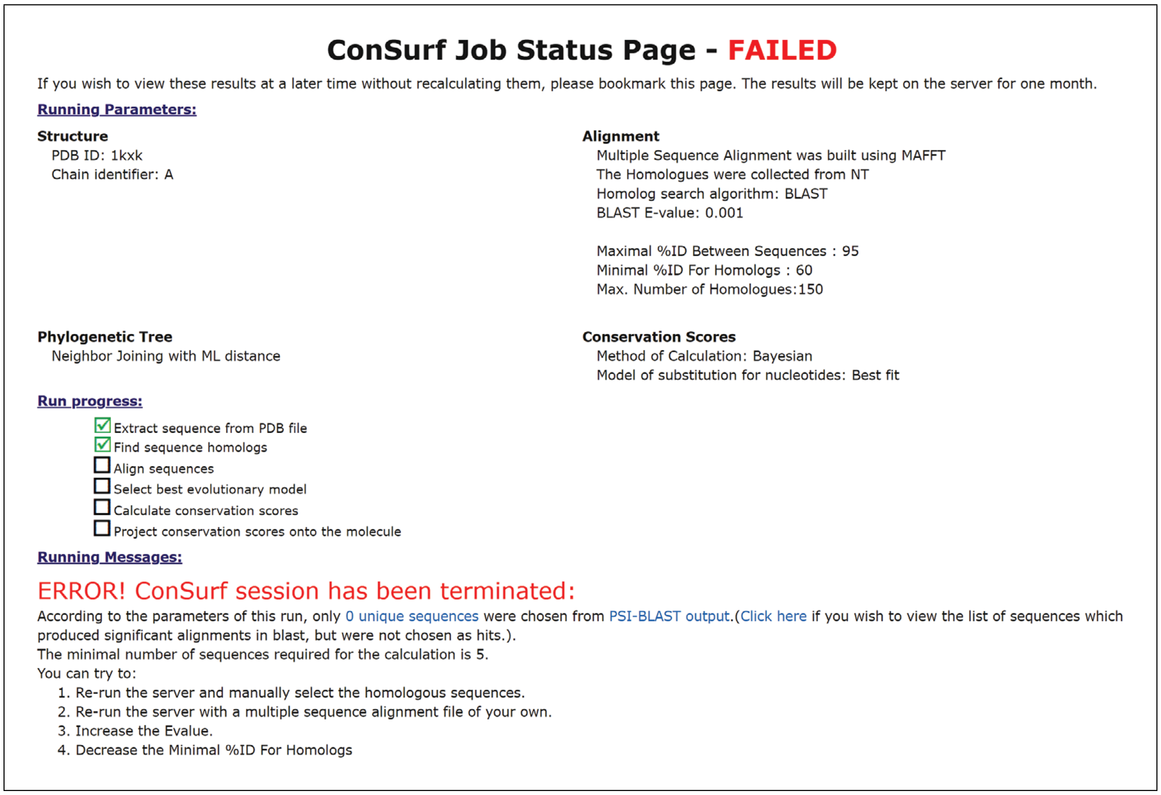
Alignment details
The average number of replacements between any two sequences in the alignment;
A distance of 0.01 means that on average, the expected replacement for every 100 positions is 1.
Average pairwise distance : 0.121156
Lower bound : 0.042302
Upper bound : 0.250725

- Nucleic acids variety per position in the MSA (The table is best viewed with an editor that respects Comma-Separated Values)

Phylogenetic Tree

- Phylogenetic Tree in Newick format
- View Phylogenetic Tree

Figure 10 A completed ConSurf run with results for an analysis starting with an RNA query of know structure following Basic Protocol 1.



ConSurf Job Status Page - FAILED

If you wish to view these results at a later time without recalculating them, please bookmark this page. The results will be kept on the server for one month.

Running Parameters:

Structure
PDB ID: 1kxk
Chain identifier: A

Alignment
Multiple Sequence Alignment was built using MAFFT
The Homologues were collected from NT
Homolog search algorithm: BLAST
BLAST E-value: 0.001

Maximal %ID Between Sequences : 95
Minimal %ID For Homologs : 60
Max. Number of Homologues:150

Phylogenetic Tree
Neighbor Joining with ML distance

Conservation Scores
Method of Calculation: Bayesian
Model of substitution for nucleotides: Best fit

Run progress:

☑ Extract sequence from PDB file
☑ Find sequence homologs
□ Align sequences
□ Select best evolutionary model
□ Calculate conservation scores
□ Project conservation scores onto the molecule

Running Messages:

ERROR! ConSurf session has been terminated:

According to the parameters of this run, only 0 unique sequences were chosen from PSI-BLAST output.(Click here if you wish to view the list of sequences which produced significant alignments in blast, but were not chosen as hits.).
The minimal number of sequences required for the calculation is 5.
You can try to:

1. Re-run the server and manually select the homologous sequences.
2. Re-run the server with a multiple sequence alignment file of your own.
3. Increase the Evalue.
4. Decrease the Minimal %ID For Homologs

Figure 11 A failed ConSurf run, due to lack of homologs to the RNA query.

For successful runs, the results page will include a secondary title, “ConSurf calculation is finished:” below the run details. Underneath it there may be a warning indicating the number of nucleic acid positions with insufficient data to reliably assign their conservation scores. If this number is too high, it is recommended to improve the search for homologs. The rest of the results will follow, and may vary slightly, depending on the input. A ConSurf analysis with a query structure will include the following in the results page (Fig. 10):

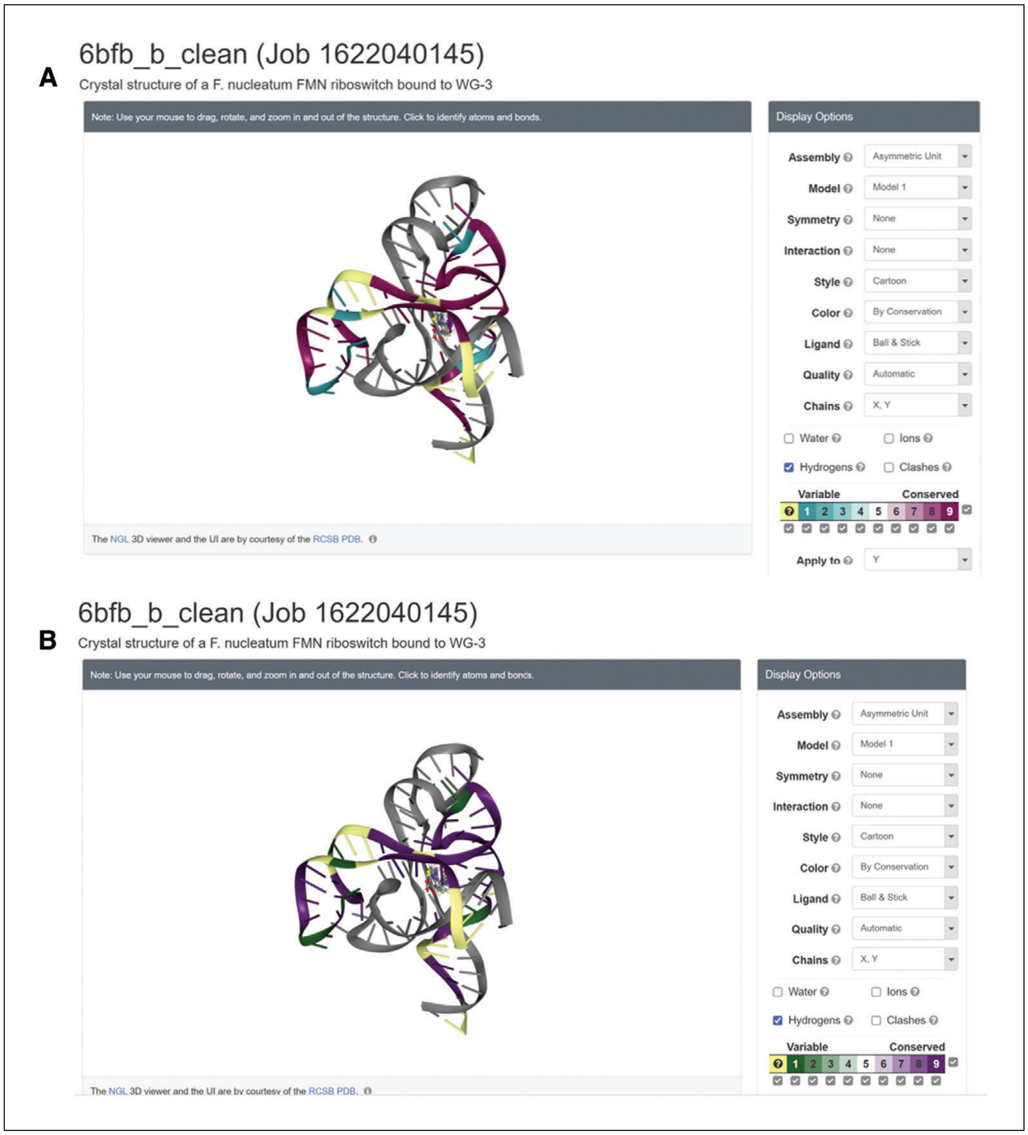
1. Final results

- The conservation profile can be viewed on the structure by selecting one of the “View ConSurf results” options.

  There are three viewing platforms: NGL viewer and FirstGlance in Jmol are online viewers, while Chimera is a platform that needs to be installed on your computer. Visualization is based on coding of the 1-through-9 conservation scores, 1 being the most variable and 9 the most highly conserved, into a cyan-through-magenta color palette. A tenth and separate color (yellow) indicates nucleobases for which the conservation score was not assigned a high enough confidence (mostly due to insufficient data, i.e., many insertions). On both online viewers, it is also possible to view the results in a color-blind-friendly scale of green-through-purple. (Fig. 12).

- The evolutionary tree and the MSA, which were either uploaded by the user or generated by the server, can be viewed on WASABI (Veidenberg, Medlar, & Löytynoja, 2016). The WASABI platform also allows the user to select a subtree and conduct a follow-up ConSurf analysis with the new section of homologs (Fig. 13).

  Open the WASABI viewer and click on the root of the subtree that you wish to conduct the follow-up analysis with. A list of actions will open, “Run ConSurf on subtree” will appear at the end. Select this option and wait (maybe 1 min). A message from ConSurf will pop up with the new search number and an “OK” button. To open the results page, click on the “OK” button, and a new tab with the job status will open.



6bfb_b_clean (Job 1622040145)

Crystal structure of a F. nucleatum FMN riboswitch bound to WG-3

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.

The NGL 3D viewer and the UI are by courtesy of the RCSB PDB.

Display Options

Assembly: Asymmetric Unit
Model: Model 1
Symmetry: None
Interaction: None
Style: Cartoon
Color: By Conservation
Ligand: Ball & Stick
Quality: Automatic
Chains: X, Y
☐ Water ☐ Ions
☑ Hydrogens ☐ Clashes
Apply to: Y

6bfb_b_clean (Job 1622040145)

Crystal structure of a F. nucleatum FMN riboswitch bound to WG-3

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.

The NGL 3D viewer and the UI are by courtesy of the RCSB PDB.

Display Options

Assembly: Asymmetric Unit
Model: Model 1
Symmetry: None
Interaction: None
Style: Cartoon
Color: By Conservation
Ligand: Ball & Stick
Quality: Automatic
Chains: X, Y
☐ Water ☐ Ions
☑ Hydrogens ☐ Clashes

Figure 12 ConSurf results of FMN riboswitch (PDB ID 6BFB chain Y) analysis following Basic Protocol 1, presented on the NGL online viewer. (**A**) defaults color scale. (**B**) colorblind friendly scale.

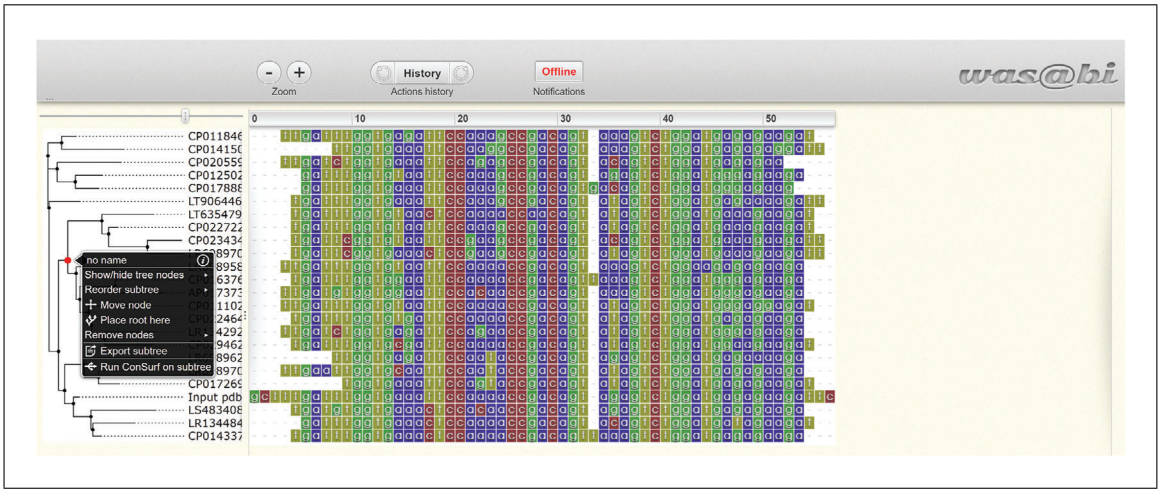


Figure 13 A phylogenetic tree viewed on WASABI with a subtree selected.

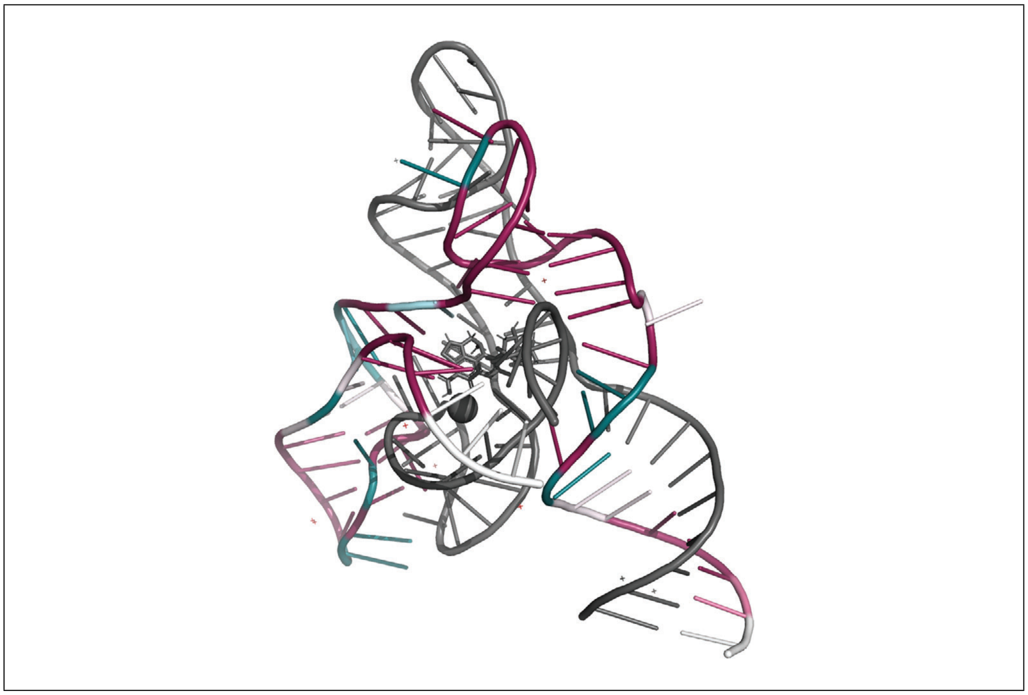


Figure 14 A high-resolution PyMOL image of FMN riboswitch (PDB ID 6BFB chain Y), hiding insufficient data.

- The MSA colored by ConSurf conservation scores.

  Each column is colored using the ConSurf color code. Yellow letters indicate columns for which the conservation score was not assigned a high enough confidence.

- A tabular text file summarizing the analysis for each base in the query sequence.

  For each position on the query, there are: a normalized score calculated, with the grade assigned on the 1-through-9 scale with 9 being the most highly conserved; the reliability estimation (for the Bayesian method); and the nucleotides observed in the respective MSA column for each position.

- A download button.

  When clicked, a compressed folder with all results will be downloaded.

2. PDB files

A file with the normalized conservation score in the Temperature Factor field and one with ConSurf results in the header can be downloaded directly from the results page.

3. Creating high-resolution figures

Figures can be produced in Chimera, PyMOL, or RasMol by following the instructions for each platform. On all platforms, there is an option to either show or hide low-confidence "insufficient" data. In the instructions for:

- Chimera–follow the instructions to produce the desired image
- PyMOL–the only option in the instructions is for figure hiding the insufficient data (Fig. 14). To create a PyMOL figure that does show insufficient data, download all the results and open the appropriate pdb file, following the instructions as they are found in the server.
- RasMol–the pdb file that can be found in the instructions cannot be used to create a figure in RasMol. Instead, download all results and follow the instructions using the desired pdb file from the results file.

4. Sequence data

If the homology search and MSA construction are conducted by ConSurf, it is possible to see the sequences found in the search as well as the final list of sequences selected to be used for the analysis.

5. Alignment

The MSA can be viewed in FASTA format, and a spreadsheet of the frequency of each nucleotide observed in each column of the MSA can be downloaded.

6. Phylogenetic tree

The tree can be viewed in Newick and Java format.

The results of a ConSurf analysis with query sequence (i.e., no structure) will be as follows (Fig. 15).

1. Final results

- The query sequence with a colored conservation profile.

  There are two viewing platforms, HTML and PDB. Visualization is based on coding of the 1-through-9 conservation scores, 1 being the most variable and 9 the most highly conserved, into a color palette. The HTML version uses the traditional cyan-through-magenta scale, and the pdf version offers a color-blind-friendly green-through-purple scale. A tenth and separate color (yellow) is used in both to indicate nucleobases for which the conservation score was not assigned a high enough confidence (mostly due to insufficient data, i.e., many insertions).

- The MSA colored by ConSurf conservation scores (Fig. 16).

  Each column is colored using the traditional cyan-through-magenta ConSurf color-code. Yellow letters indicate columns for which the conservation score was not assigned a high enough confidence.

- The evolutionary tree and the MSA, which were either uploaded by the user or generated by the server, can be viewed on WASABI. The WASABI platform also allows the user to select a subtree and conduct a follow-up ConSurf analysis with the new section of select homologs.

  Open the WASABI viewer and click on the root of the subtree on you wish to conduct the follow-up analysis. A list of actions will open; "Run ConSurf on subtree" will appear at the end. Select this option and wait (around a minute); a message from ConSurf will pop up with the new search number and an "OK" button. To open the results page, click on the "OK" button and a new tab with the job status will open.

- Chimera view of the MSA.

  By following the instructions detailed under the question mark, the MSA, can be viewed with Chimera on your computer.

- A tabular text file summarizing the analysis for each base in the query sequence.

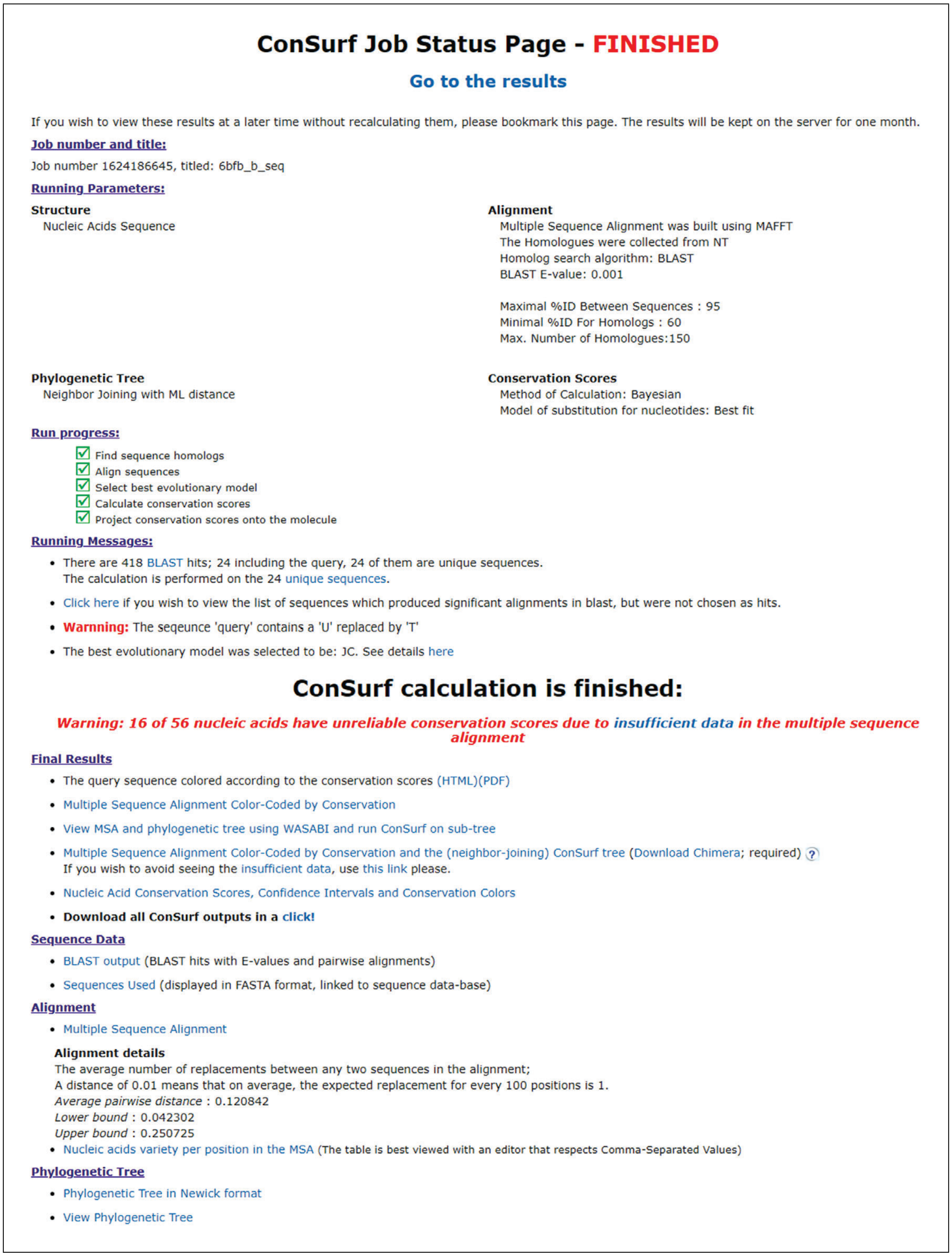
  For each position on the query there are: a normalized score calculated, with the grade assigned on the 1-through-9 scale, with 9 being the most highly conserved; the reliability estimation (for the Bayesian method); and the nucleotides observed in the respective MSA column for each position.

- A download button.

  When clicked, a compressed folder with all results will be downloaded.

2. Sequence data

If the homology search and MSA construction is conducted by ConSurf, it is possible to see the sequences found in the search as well as the final list of sequences selected to be used for the analysis.



ConSurf Job Status Page - FINISHED

Go to the results

If you wish to view these results at a later time without recalculating them, please bookmark this page. The results will be kept on the server for one month.

Job number and title:

Job number 1624186645, titled: 6bfb_b_seq

Running Parameters:

Structure
Nucleic Acids Sequence

Alignment
Multiple Sequence Alignment was built using MAFFT
The Homologues were collected from NT
Homolog search algorithm: BLAST
BLAST E-value: 0.001

Maximal %ID Between Sequences : 95
Minimal %ID For Homologs : 60
Max. Number of Homologues:150

Phylogenetic Tree
Neighbor Joining with ML distance

Conservation Scores
Method of Calculation: Bayesian
Model of substitution for nucleotides: Best fit

Run progress:

☑ Find sequence homologs
☑ Align sequences
☑ Select best evolutionary model
☑ Calculate conservation scores
☑ Project conservation scores onto the molecule

Running Messages:

- There are 418 BLAST hits; 24 including the query, 24 of them are unique sequences.
  The calculation is performed on the 24 unique sequences.
- Click here if you wish to view the list of sequences which produced significant alignments in blast, but were not chosen as hits.
- **Warnning:** The seqeunce 'query' contains a 'U' replaced by 'T'
- The best evolutionary model was selected to be: JC. See details here

ConSurf calculation is finished:

Warning: 16 of 56 nucleic acids have unreliable conservation scores due to insufficient data in the multiple sequence alignment

Final Results

- The query sequence colored according to the conservation scores (HTML)(PDF)
- Multiple Sequence Alignment Color-Coded by Conservation
- View MSA and phylogenetic tree using WASABI and run ConSurf on sub-tree
- Multiple Sequence Alignment Color-Coded by Conservation and the (neighbor-joining) ConSurf tree (Download Chimera; required) ?
  If you wish to avoid seeing the insufficient data, use this link please.
- Nucleic Acid Conservation Scores, Confidence Intervals and Conservation Colors
- **Download all ConSurf outputs in a click!**

Sequence Data

- BLAST output (BLAST hits with E-values and pairwise alignments)
- Sequences Used (displayed in FASTA format, linked to sequence data-base)

Alignment

- Multiple Sequence Alignment

Alignment details
The average number of replacements between any two sequences in the alignment;
A distance of 0.01 means that on average, the expected replacement for every 100 positions is 1.
Average pairwise distance : 0.120842
Lower bound : 0.042302
Upper bound : 0.250725

- Nucleic acids variety per position in the MSA (The table is best viewed with an editor that respects Comma-Separated Values)

Phylogenetic Tree

- Phylogenetic Tree in Newick format
- View Phylogenetic Tree

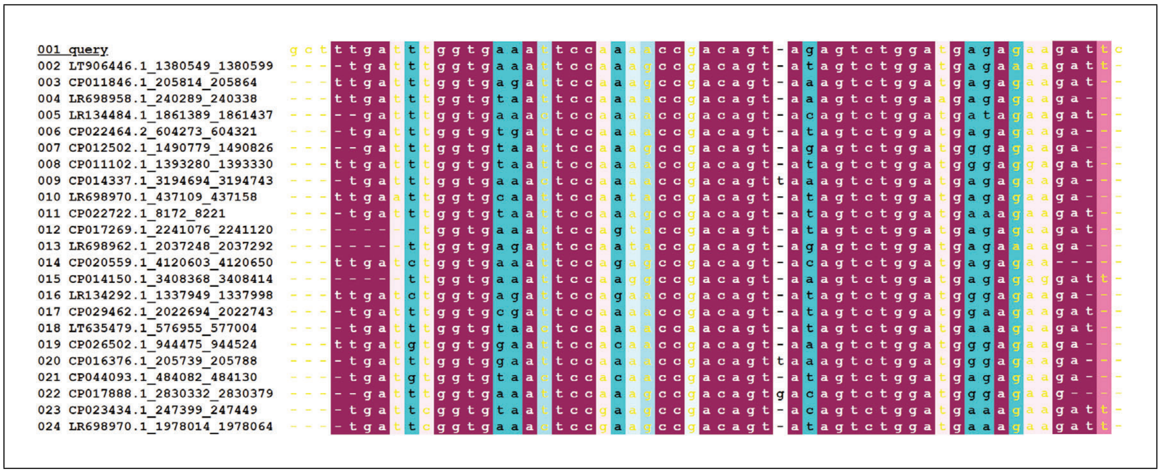
Figure 15 A completed ConSurf run with results for an analysis starting with an RNA sequence query following Basic Protocol 1.

3. Alignment

The MSA can be viewed in FASTA format, and an Excel file of the frequency of each nucleotide observed in each column of the MSA can be downloaded.

4. Phylogenetic tree

The tree can be viewed in Newick and Java format.



```
001 query                       g c t t t g a t t t g g t g a a a t t c c a a a a c c g a c a g t - a g a g t c t g g a t g a g a g a a g a t t c
002 LT906446.1_1380549_1380599  - - - - t g a t t t g g t g a a a t t c c a a a g c c g a c a g t - a t a g t c t g g a t g a g a a a a g a t t -
003 CP011846.1_205814_205864    - - - t t g a t t t g g t g a g a t t c c a a a g c c g a c a g t - a a a g t c t g g a t g a g a g a a g a t - -
004 LR698958.1_240289_240338    - - - t t g a t t t g g t g t a a t t c c a a a a c c g a c a g t - a a a g t c t g g a a g a g a g a a g a - - -
005 LR134484.1_1861389_1861437  - - - - - g a t t t g g t g a a a c t c c a a a a c c g a c a g t - a c a g t c t g g a t g a t a g a a g a t - -
006 CP022464.2_604273_604321    - - - - t g a t t t g g t g t g a t t c c a a a a c c g a c a g t - a t a g t c t g g a t g a g a g a a g a - - -
007 CP012502.1_1490779_1490826  - - - - - g a t t t g g t g t a a t t c c a a a g c c g a c a g t - a g a g t c t g g a t g g g a g a a g a - - -
008 CP011102.1_1393280_1393330  - - - t t g a t t t g g t g t a a t t c c a a a a c c g a c a g t - a t a g t c t g g a t g g g a g g a g a t - -
009 CP014337.1_3194694_3194743  - - - - t g a t t t g g t g a a a c t c c a a a a c c g a c a g t t a a a g t c t g g a t g a g a g a a g a - - -
010 LR698970.1_437109_437158    - - - t t g a a t t g g t g c a a t t c c a a t a c c g a c a g t - a t a g t c t g g a t g a g a g a a g a - - -
011 CP022722.1_8172_8221        - - - - t g a t t t g g t g t a a t t c c a a a g c c g a c a g t - a t a g t c t g g a t g a a a g a a g a t - -
012 CP017269.1_2241076_2241120  - - - - - - - - - t g g t g a a a t t c c a g t a c c g a c a g t - a t a g t c t g g a t g a g a g a a g a t - -
013 LR698962.1_2037248_2037292  - - - - - - - - t t g g t g a g a t t c c a a t a c c g a c a g t - a g a g t c t g g a t g a g a a a a g a - - -
014 CP020559.1_4120603_4120650  - - - t t g a t c t g g t g a a a t t c c a g a g c c g a c a g t - a c a g t c t g g a t g a g a g a a - - - - -
015 CP014150.1_3408368_3408414  - - - - - - - - t t g g t g a a a t t c c a a g g c c g a c a g t - a a a g t c t g g a t g a g a g a g g a t t -
016 LR134292.1_1337949_1337998  - - - t t g a t c t g g t g a g a t t c c a g a a c c g a c a g t - a t a g t c t g g a t g g g a g a a g a - - -
017 CP029462.1_2022694_2022743  - - - - t g a t t t g g t g c g a t t c c a a a a c c g a c a g t - a t a g t c t g g a t g g a a g a a g a t - -
018 LT635479.1_576955_577004    - - - - t g a t t t g g t g t a a c t c c a a a a c c a a c a g t - a t a g t c t g g a t g a a a g a a g a t - -
019 CP026502.1_944475_944524    - - - t t g a t g t g g t g g a a t t c c a c a a c c g a c a g t - a a a g t c t g g a t g g g a g a a g a - - -
020 CP016376.1_205739_205788    - - - - t g a t t t g g t g g a a t t c c a a a a c c g a c a g t t a a a g t c t g g a t g g g a g a a g a - - -
021 CP044093.1_484082_484130    - - - - t g a t g t g g t g t a a c t c c a c a a c c g a c a g t - a t a g t c t g g a t g a g a g a a g a - - -
022 CP017888.1_2830332_2830379  - - - - - g a t t t g g t g a a a t t c c a a a g c c g a c a g t g a c a g t c t g g a t g g g a g a a g - - - -
023 CP023434.1_247399_247449    - - - - t g a t t c g g t g t a a t t c c g a a g c c g a c a g t - a c a g t c t g g a t g a a a g a a g a t t -
024 LR698970.1_1978014_1978064  - - - - t g a t t c g g t g a a a c t c c g a a g c c g a c a g t - a t a g t c t g g a t g a a a g a a g a t t -
```

Figure 16 A color-coded MSA starting with a sequence query, analyzed by following Basic Protocol 1.

COMMENTARY

Background Information

ConSurf is a web-based tool which estimates the conservation of amino/nucleic acid positions in a protein/DNA/RNA molecule. Unfortunately, the pipeline for analyzing nucleic acids is mostly useless, in essence. The reason is that for the most part, the pipeline, which is based on dated databases and sequence search tools, fails to find a large enough set of effective homologs to allow a meaningful estimate of the evolutionary rates. Here we have described two complementary protocols (and variations thereof) based on state-of-the-art tools and databases for building a large enough multiple sequence alignment of effective homologs for your RNA molecule query. To examine the utility of these protocols, we utilized the pipelines to analyze a representative set of dozen RNA molecules of known structure (Table 1). By following Basic Protocol 1, four of the dozen queries we examined were successfully analyzed. When combining Basic Protocol 2 with Support Protocol 1, we were able to successfully analyze eleven of the twelve query sequences. All of the queries were successfully analyzed when combining Basic Protocol 2 with Support Protocol 2.

Critical Parameters

Number of effective homologs

The number of effective homologs used to construct the MSA has great influence on the quality of the analysis. Too few homologs might not be sufficient to evenly sample the relevant sequence space. A minimum of five homologs (including the query) are required in ConSurf, but many more are usually needed for accurate estimate of the evolutionary rate per site. On the other hand, too many homologs may slow down the run or even prevent completion. We recommend including between 50-and-300 homologs.

Inclusion threshold E-value

Increasing the E-value of included homologs will increase the number of homologs. The E-value is set to determine the likelihood of a false positive homolog. If it is set to 1, then, on average, there will be one false positive in the list of homologs.

Sequence identity threshold

When clustering the homologs with CD-HIT, the sequence identity cut-off should be between 95% and 80%. Using a higher threshold may compromise the integrity of the final ConSurf analysis, as the relevant sequences may lack diversity. A lower threshold might compromise the clustering process and the CD-HIT run may not complete.

Troubleshooting

Table 2 lists common problems that may arise with the protocols in this article, along with their possible causes and solutions.

Acknowledgments

The authors thank Gal Masrati and Amit Kessel for their assistance on this protocol. The research was supported by Grant 450/16 of the Israeli Science Foundation (ISF). NB-T's research is supported in part by the Abraham E. Kazan Chair in Structural Biology, Tel Aviv University.

Author Contributions

Maya Rubin: Writing original draft; **Nir Ben-Tal:** supervision, writing review and editing.

Table 1 A Summary of the Results Obtained for the Sample of 12 RNA Queries[a]

Query details			Basic Protocol 1		Basic Protocol 2 + Support Protocol 1		Basic Protocol 2 + Support Protocol 2	
PDB id	Chain	Sequence length	Number of effective homologs	Job status	Number of effective homologs	Job status	Number of effective homologs	Job status
5fk4	A	93	0	Failed	156	Completed	226	Completed
4tra	A	76	27	Completed	34	Completed	174	Completed
1z43	A	101	2	Failed	166	Completed	59	Completed
1kxk	A	71	0	Failed	105	Completed	10[b]	Completed
2qbz	A	161	19	Completed	285	Completed	199	Completed
2gcs	B	125	1	Failed	234	Completed	48	Completed
2ydh	A	94	1	Failed	155	Completed	228	Completed
4wfm	A,B	103	11	Failed	102	Completed	141	Completed
6bfb	A	54	6	Completed	224	Completed	33	Completed
6bfb	B	56	23	Completed	231	Completed	81	Completed
6d8o	A,B	158	1	Failed	264	Completed	16[b]	Completed
1duh	A	45	0	Failed	1	Failed	15[b]	Completed

[a] The three leftmost columns list the PDB ID, chain, and number of RNA bases of the query. The next two columns list the number of effective homologs detected and whether the run using Basic Protocol 1 completed successfully. The next two columns list the number of effective homologs detected and whether the run using Basic Protocol 2 in combination with Support Protocol 1 completed successfully. The last two columns list the number of effective homologs detected and whether the run using Basic Protocol 2 in combination with Support Protocol 2 completed successfully. All analyses were carried out with RNA queries of known structure and default settings.

[b] When searching for homologs using nHMMER locally, the inclusion threshold E-value was set to 0.1 rather than the default value.

Table 2 Troubleshooting Guide

Problem	Possible cause	Solution
Many of the nucleic acids have unreliable conservation scores due to insufficient data	MSA was constructed with too few sequences	Increase number of effective homologs
MSA is too large, causing the run to fail	There are over 300 sequences included in the MSA	Decrease the number of effective homologs by decreasing the percentage "sequence ID cut-off" in CD-HIT.
	The query is a long sequence with a large MSA of over 200 effective homologs	
The character "U" is found in the MSA	The adjustment of replacing "U" with "T" when constructing an MSA externally was skipped	Modify the MSA to fit the requirement of the server and attempt to run the analysis again

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Literature Cited

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: An improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, *44*(W1), W344–50. doi: 10.1093/nar/gkw408.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., & Ben-Tal, N. (2010). ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, *38*(Web Server issue), W529–33. doi: 10.1093/nar/gkq399.

Capra, J. A., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, *23*(15), 1875–1882. doi: 10.1093/bioinformatics/btm270.

Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., … Ben-Tal, N. (2013). ConSurf: Using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry*, *53*(3-4), 199–206. doi: 10.1002/ijch.201200096.

Del Sol, A., Pazos, F., & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *Journal of Molecular Biology*, *326*(4), 1289–1302. doi: 10.1016/s0022-2836(02)01451-1.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, *23*(1), 205–211. doi: 10.1142/9781848165632_0019.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. doi: 10.1093/bioinformatics/bts565.

Gallet, X., Charloteaux, B., Thomas, A., & Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *Journal of Molecular Biology*, *302*(4), 917–926. doi: 10.1006/jmbi.2000.4092.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160–174. doi: 10.1007/BF02101694.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680–682. doi: 10.1093/bioinformatics/btq003.

Huang, Y.-F., & Golding, G. B. (2014). Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLoS Computational Biology*, *10*(1), e1003429. doi: 10.1371/journal.pcbi.1003429.

Huang, Y.-F., & Golding, G. B. (2015). FuncPatch: A web server for the fast Bayesian inference of conserved functional patches in protein 3D structures. *Bioinformatics*, *31*(4), 523–531. doi: 10.1093/bioinformatics/btu673.

Innis, C. A. (2007). siteFiNDER|3D: A web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Research*, *35*(Web Server issue), W489–94. doi: 10.1093/nar/gkm422.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian protein metabolism* (pp. 21–132). Elsevier. doi: 10.1016/B978-1-4832-3211-9.50009-7.

Katoh, K., Rozewicki, J., & Yamada, K. D. (2017). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, bbx108. doi: 10.1093/bib/bbx108.

Landgraf, R., Xenarios, I., & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *Journal of Molecular Biology*, *307*(5), 1487–1502. doi: 10.1006/jmbi.2001.4540.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. doi: 10.1093/bioinformatics/btl158.

Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996a). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, *257*(2), 342–358. doi: 10.1006/jmbi.1996.0167.

Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996b). Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(15), 7507–7511. doi: 10.1073/pnas.93.15.7507.

Lichtarge, O., Yamamoto, K. R., & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *Journal of Molecular Biology*, *274*(3), 325–337. doi: 10.1006/jmbi.1997.1395.

Mayrose, I., Graur, D., Ben-Tal, N., & Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Molecular Biology and Evolution*, *21*(9), 1781–1791. doi: 10.1093/molbev/msh194.

NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. doi: 10.1093/nar/gkx1095.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. doi: 10.1002/jcc.20084.

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., & Ben-Tal, N. (2002). Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, *18*(Suppl 1), S71–7. doi: 10.1093/bioinformatics/18.suppl_1.s71.

RNAcentralConsortium. (2021). RNAcentral 2021: Secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, *49*(D1), D212–D220. doi: 10.1093/nar/gkaa921.

Sayle, R. A., & Milner-White, E. J. (1995). RASMOL: Biomolecular graphics for all. *Trends in Biochemical Sciences*, *20*(9), 374. doi: 10.1016/s0968-0004(00)89080-5.

Schrödinger. (2021). PyMOL Molecular Graphics System (2.5.1). Computer software, New York: Schrödinger, LLC.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, *9*(4), 678–687. doi: 10.1093/oxfordjournals.molbev.a040752.

Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Some Mathematical Questions in Biology /DNA Sequence Analysis edited by RobertM.Miura*,.

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins*, *48*(2), 227–241. doi: 10.1002/prot.10146.

Veidenberg, A., Medlar, A., & Löytynoja, A. (2016). Wasabi: An integrated platform for evolutionary sequence analysis and data visualization. *Molecular Biology and Evolution*, *33*(4), 1126–1130. doi: 10.1093/molbev/msv333.

Wheeler, T. J., & Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, *29*(19), 2487–2489. doi: 10.1093/bioinformatics/btt403.

Internet Resources

https://www.ncbi.nlm.nih.gov/nucleotide/

Nucleotide Home Page, NCBI.