Software

Open Access

RiboSubstrates: a web application addressing the cleavage specificities of ribozymes in designated genomes Jean-François Lucier^{1,2}, Lucien Junior Bergeron^{1,3}, Francis P Brière^{1,3}, Rodney Ouellette^{4,5}, Sherif Abou Elela^{1,2} and Jean-Pierre Perreault^{*1,3}

Address: ¹RNA Group/Groupe ARN, Sherbrooke, Canada, ²Département de microbiologie et infectiologie, Université de Sherbrooke, Sherbrooke, Canada, ³Département de biochimie, Université de Sherbrooke, Sherbrooke, Canada, ⁴Institut Atlantique de Recherche sur le Cancer, Moncton, Canada and ⁵Département de Chimie et Biochimie, Université de Moncton, Moncton, Canada

Email: Jean-François Lucier - jean-francois.lucier@usherbrooke.ca; Lucien Junior Bergeron - lucien.junior.bergeron@usherbrooke.ca; Francis P Brière - francis.briere@usherbrooke.ca; Rodney Ouellette - rodneyo@health.nb.ca;

Sherif Abou Elela - Sherif.abou.elela@usherbrooke.ca; Jean-Pierre Perreault* - jean-pierre.perreault@usherbrooke.ca * Corresponding author

Published: 31 October 2006

BMC Bioinformatics 2006, 7:480 doi:10.1186/1471-2105-7-480

This article is available from: http://www.biomedcentral.com/1471-2105/7/480

© 2006 Lucier et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<u>http://creativecommons.org/licenses/by/2.0</u>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: 22 August 2006 Accepted: 31 October 2006

Abstract

Background: RNA-dependent gene silencing is becoming a routine tool used in laboratories worldwide. One of the important remaining hurdles in the selection of the target sequence, if not the most important one, is the designing of tools that have minimal off-target effects (i.e. cleaves only the desired sequence). Increasingly, in the current dawn of the post-genomic era, there is a heavy reliance on tools that are suitable for high-throughput functional genomics, consequently more and more bioinformatic software is becoming available. However, to date none have been designed to satisfy the ever-increasing need for the accurate selection of targets for a specific silencing reagent.

Results: In order to overcome this hurdle we have developed RiboSubstrates <u>http://</u><u>www.riboclub.org/ribosubstrates</u>. This integrated bioinformatic software permits the searching of a cDNA database for all potential substrates for a given ribozyme. This includes the mRNAs that perfectly match the specific requirements of a given ribozyme, as well those including Wobble base pairs and mismatches. The results generated allow rapid selection of sequences suitable as targets for RNA degradation. The current web-based RiboSubstrates version permits the identification of potential gene targets for both SOFA-HDV ribozymes and for hammerhead ribozymes. Moreover, a minimal template for the search of siRNAs is also available. This flexible and reliable tool is easily adaptable for use with any RNA tool (i.e. other ribozymes, deoxyribozymes and antisense), and may use the information present in any cDNA bank.

Conclusion: RiboSubstrates should become an essential step for all, even including "non-RNA biologists", who endeavor to develop a gene-inactivation system.

Background

The development of gene-inactivation systems is an active and important field for both functional genomics and gene therapy (reviewed in references [1-3]). In fact, RNA dependent gene silencing is becoming a routine tool used by laboratories worldwide. However, the issue of sub-

strate specificity remains an important hurdle, regardless of the approach used (e.g. see references [4-7]). A key question is how to design target-specific RNA molecules that do not produce any side effects. There are several potential sources for side effects, including triggerring immunological responses, as well as degradation and translational suppression, to name only these ones. Moreover, in many cases side effects may result from the cleavage of non-desired mRNA species. There are several means for verifying the specificity of a given RNA, including profiling its impact on global gene expression using microarray analysis [4,6,7]. However, all of these approaches are performed following the silencing experiments and, consequently, have no predictive value. Only in the case of siRNA, have algorithms been developed to identify potential targeting of non-desired mRNA species before to attempting any manipulation [e.g. [8,9]]. However, since it is uncertain how siRNAs cause off-target effects, the accuracy of these algorithms remains to be demonstrated. Moreover, no such algorithm has been reported for ribozyme design. Thus, the selection of a unique target sequence remains a key process that, in most cases, is determined empirically.

Ideally, we would like to possess either a bioinformatic or a biochemical approach that permits verification of whether or not an RNA silencing tool offers the specificity required in order to target a unique mRNA species. This should permit the identification of all potential substrates, including both those that are perfectly matching as well those for which binding involves either Wobble base pairs or the presence of some mismatches. The commonly used Blast based target search is of limited utility in this regard because it is not suitable for evaluating short sequences like that found in the Hepatitis Delta Virus ribozyme (HDV ribozyme) binding domain (i.e. a stretch of 7 nucleotides) [7]. More specifically, the Blastn algorithm uses an heuristic approach to search for a minimal sequence size of 7 consecutive nucleotides which cannot contain any mismatches [10]. Alternatively, there is the WU-BLASTalgorithm that supports word-length search as small as only 1 nucleotide [11]. However, this algorithm is not suitable for RNA target search since it can miss potential hits. For example, if the searched sequence is composed of only G and U nucleotides, because of wobble base pairing, the searched signature will be composed of Y and R nucleotides. With such a signature, WU-BLAST cannot find any hits because no perfectly matched nucleotide can be retreived. Finally, there is the Smith and Waterman algorithm for searching minimal sequence; however it is known to be slow and memory consuming [12]. Most importantly there is no software that offers predefined templates for specific RNA silencing tools. At the dawn of post-genomic era, where there is a heavy reliance on tools that are suitable for high-throughput functional genomics, more and more bioinformatic software is becoming available, but, to date, none have been designed to satisfy the ever increasing need for the fast and accurate selection of targets for a specific silencing reagent.

Here we describe "RiboSubstrates", a new solution for the selection of RNA targets. RiboSubstrates is a portable software that can be adapted to any kind of RNA silencing requirements, and may be used to scan any predefined cDNA database. The main goal of the program is to evaluate whether or not a given ribozyme sequence has a suitable and unique target sequence within the transcriptome of a particular genome. The results obtained by the Ribo-Substrates scan provides an overall evaluation of the potential for non-specific cleavages, and thus helps to identify the ribozyme that should have the least amount of non-specific cleavage for a given application in vivo. To our knowledge, to date no other software permitting the identification of non-specific targets has been reported for ribozyme development. The current version of RiboSubstrates is a web-based application that permits the identification of potential gene targets for SOFA-HDV ribozymes, hammerhead ribozymes and siRNAs.

Implementation

RiboSubstrates has been written in Perl using the freely available CPAN modules CGI::Application, HTML::Template and Class::DBI [13,14]. In order to facilitate the use of the software, a help document was developed and is accessible at any moment at the top of the search pages.

Essentially, RiboSubstrates includes four modules and two configuration files (see Figure 1). The first module was developed for new cDNA database integration. To date, the databases from Homo sapiens, Mus musculus, Saccharomyces cerevisiae, Escherichia coli, Lactococcus lactis and Leishmania major are available. Additional information on databases is available in the online help (see Database notes). These databases were recovered from public libraries and integrated into RiboSubstrates using the described module. In the case of Lactococcus lactis and Leishmania major, the mRNAs were extracted directly from the annotation of these genomes. Additional cDNA databases can easily be added on request; all that is required is a file in fasta format containing all of the entries. For a full Ribo-Substrates integration, the header must be in the following format: >ID|Name|Description. The ID field must be the ID provided by the external source for a direct link to the external source description. The description can be anything; we used the gene function and location in genome for the default description. If the header is not in the above format, the search result will display the header without any interpretation.



Figure I

Process flow for RiboSubstrates. The cDNA database management module, the RNA tool interface module, the substrate search and characterization module and the substrate interface module are illustrated.

The second module interprets a configuration file provided to the application (RiboParticularities.xml). This file describes the signature format to be searched for each ribozyme. For example, it will include information on the length of the recognition domain of a given ribozyme; the specific nucleotide requirement within, before, or after the binding domain, etc. Position constraints, including Watson-Crick and Wobble pairing, can be also considered. An illustration of the ribozymes, including the characteristics of the descriptors, is posted on the web page. A domain for a given ribozyme which a researcher is currently entering a sequence is simultaneously boxed in red in order to prevent any error. Researchers can enter sequences that include either Us or dTs without any difference.

The third module is the substrate search and score calculation. Initially, the cDNA database is scanned for the substrate sequence signature generated from the user inputs combined to the ribozyme constraints defined in the configuration file. Possible constraints include a maximum number of either Wobble base pairs or mismatches. In order to minimize the calculation time a maximum of 3 mismatches has been defined by default. This module produces a preliminary result file of potential substrates. A score is then calculated for each potential substrate. Each Wobble base pair encountered increases the score by 10, and each mismatch by 100. These arbitrarily values have been fixed based on kinetic data from different sources (e.g. see references [15,16] for the HDV ribozyme). For example, in the case of a SOFA-HDV ribozyme, the presence of a mismatch has been shown to reduce the overall specificity substrate from 5 to 18 times, while the presence of a Wobble base pair has been shown to have a limited effect. If the information for a spacer (see SOFA-HDV ribozyme example below) is specified in RiboParticularities.xml, then the tool score can be increased by the factors specified in the configuration file.

RiboSubstrates does not have a score cut off that indicates that a ribozyme should be re-designed because each RNA tool has its own specific requirements in order to be functional and the presence of both a Wobble base pair or a mismatch might affect different ribozymes differently. Moreover, the relationship between substrate specificity and RNA silencing efficacy is not defined by a simple linear equation based on the binding. The score serves mainly to help in identifying potential target sites including Wobble base pairs versus those possessing mismatches. Moreover, it is not realistic for a unique scoring system to be relevant for all potential RNA silencing tools since they all possess different requirements in terms of substrate specificity. However, regardless of the specific case, the presence of a mismatch will be more detrimental than the presence of a Wobble base pair.

Finally, the fourth module is the substrates list and detailed display. The list display is generated using a second configuration file (RiboSubstParticularities.xml) that describes the substrate list display. Each substrate sequence has different constraints for color mapping based on nucleotide binding to ribozyme. The color mapping is based on position, mismatches and wobble. Result tables are sorted as a function of the score, the lowest scores being the best matches. The first result table contains all substrates that are perfectly matched for the RNA tool in question (Figures 2 and 3). The second table contains all hits that have no mismatches, but that contain Wobble base-pairings. The last table contains hits with both mismatches and Wobble base pairs. The information displayed in each table includes the cDNA ID, the cDNA's name, the recognition sequence match and other information specific for each RNA tool. The ID for each hit is a link to a more detailed view of the potential substrate. This view presents essential information about the targeted substrates, including a description of the targeted mRNA. Depending on the interpretation of the header provided with the cDNA database, this may also include a link to the external source information (e.g. the NCBI detailed view of the mRNA). Other information displayed shows the position within the mRNA of the sequence targeted by the designed ribozyme. This position is mapped in color on the sequence. This module also permits the user to store the substrates in an Excel file.

The user can either wait online in order to receive the result, or it may be delivered via an e-mail message sent by the system once the computation is completed. For relatively small cDNA databases, the search is almost instantaneous, while for larger ones it takes few minutes. New RNA tools (i.e. ribozymes, deoxyribozymes, antisense DNAs and RNAs, etc.) can easily be integrated since only two configurations files must be edited for full RNA tool integration into the software. All web pages related to this newly integrated RNA tool are automatically generated from these files.

Results & Discussion

Since the current state of knowledge of the features involved the substrate specificities of siRNAs is limited, ribozymes represent a suitable RNA silencing tool with which to illustrate the potential of this software. More specifically, we have developed examples for both SOFA-HDV ribozymes and hammerhead ribozymes. Moreover, a signature file is also available for siRNA search.

The SOFA-HDV ribozymes (SOFA stands for <u>Specific Qn</u>/ o<u>Ff A</u>daptor) are an improved generation of ribozymes that possess significant potential in both functional genomics and gene therapy [15,17]. These ribozymes recognize their substrates through the formation of two

		See search	example Notes (on datak	ases						
		<u>see search</u>		Ji uatal							
Search N	lame:										
cDNA Ba	ank:	nk: Human NCBI build 35.1 mRNAs + HCV type 1b (+ / -) complete genom									
P1 seque	ence (5'->3') AUGGCL U										
BS sequ	ence (5'->3	ence (5'->3') CGGUUCCGCAGA									
BS mism	natch	2									
Minimum	n spacer ler	ngth 1									
Maximun	n spacer le	ngth 10									
Email:		Jean-Franc	ois.Lucier@USherbrool	ke.ca							
		C Send	email when resul	ts are re	eady						
		Wait	for results here								
			Submit Bosot								
		a BS	SPACER P1	4.00							
		3NNNNNNNN	NNNN NNNNN	IG							
		5' NNNNNNN		" "	-						
) ~)						
			\sim	\sim							
			3'								
В											
			Search parameters	5							
BS seque	ence (5'->3')) CGGTTCCGC	AGA								
BS misma	atch	2									
P1 seque	ence (5'->3')) ATGGCTT									
Minimum	spacer leng	th 1									
Maximum	n spacer leng	gth 10									
Leg	end: Orange	= -2,-1 substrat	Save in Excel te position Blue = we Perfect match	obble bas	e pair Red =	= missmato	h				
Lege Id	end: Orange	= -2,-1 substra	Save in Excel te position Blue = we Perfect match BS sequence	obble bas	e pair Red =	BS5	h Score				
Legi Id G	Gene P name n	= -2,-1 substrat	Save in Excel te position Blue = we Perfect match BS sequence match TCTGCGGAACCG	Spacer	e pair Red = 3855 mismatch	BS5 wobble	h Score				
Leg Id G 5420376 A	Gene P name n 0238799.1 C	= -2,-1 substrat	Save in Excel te position Blue = we Perfect match BS sequence match TCTGCGGAACCG	Spacer	e pair Red = 3855 mismatch	BS5 wobble	h Score				
Lega Id G 5420376 A	Gene P hame n 1/238799.1	= -2,-1 substrat	Save in Excel te position Blue = wo Perfect match BS sequence match TCTGCGGAACCG	Spacer 5	e pair Red = 3S5 nismatch	BS5 wobble	h Score				
Legi Id G 5420376 A	Gene P name n D238799.1 C	= -2,-1 substrat	Save in Excel te position Blue = wr Perfect match BS sequence match TCTGCGGAACCG	Spacer 5 (e pair Red = 3S5 mismatch	BS5 wobble 0	h Score				
Leg Id G 5420376 A	Gene name	= -2,-1 substrat	save in Excel te position Blue = wir Perfect match BS sequence match TCTGCGGAACCG	Spacer 5	BS5 mismatch	BS5 wobble 0 BS5 wobble	h Scon 0				
Leg Id G 5420376 A	Gene name XM_498542.:	e -2,-1 substrat	save in Excel te position Blue = w Perfect match BS sequence match TCTGCGGAACCG	Spacer 5 0	BS5 mismatch BS5 mismatch 1	BS5 wobble 0 BS5 wobble	h Scorr 0 Scorr 116				
Leg Id G 5420376 A Id 51471040 51477462	Gene name XM_498542.: XM_172801.:	= -2,-1 substrai	save in Excel te position Blue = w Perfect match BS sequence match TCTGCGGAACCG	Spacer 5 0 atch 10 1	BS5 mismatch b BS5 mismatch 1	BS5 wobble 0 BS5 wobble	h Score 0 Score 116 120				
Leg Id Gn 5420376 A Id 51471040 51477462 8922602	Gene name XM_498542. XM_172801.3	 -2,-1 substrait P1 sequence match P1 sequence match AGGAGCCAT GAGAGTCAT TAGAGCCAT 	Save in Excel te position Blue = wr Perfect match BS sequence match TCTGCGGAACCG	spacer 5 0 atch 10 1 7	BS5 mismatch mismatch 1 1	BS5 wobble 0 1 2 2	h Score 0 Score 116 120 124				
Leg Id Gn 5420376 A 51471040 51477462 8922602 21389558	Gene name XM_498542.: XM_172801.: NM_018185. NM_144669.	 = -2,-1 substrait P1 sequence match P1 sequence match AGAGCCAT AGGAGCCAT GAGAGTCAT TAGAGCCAT TAGAGCCAT TGGAGTCAT 	Save in Excel te position Blue = wr Perfect match BS sequence match TCTGCGGAACCG	Spacer 5 (Spacer 5 (Spacer 10 1 1 7 2	BSS mismatch 1 1 1	BS5 wobble 0 BS5 wobble 1 2 2 3	h Scorr 0 116 120 124 130				
Leg Id G 5420376 A Id 51471040 51477462 8922602 21389558 3134140	Gene P XM_498542. XM_12881.1 XM_12881.1 NM_0141669. NM_144669. NM_144669.	 = -2,-1 substrai P1 sequence match MGAGCCAT P1 sequence match AGGAGCCAT GAGAGCCAT GAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT 	Save in Excel te position Blue = wr Perfect match BS sequence match TCTGCGGAACCG BS sequence match TCTGCGGAGCCC TCTGTGGGACCCG TTGGTGGAACTG TCTATGGGACCTG	Spacer 5 (Spacer 10 1 7 2 1	BSS nismatch 1 1 1	BS5 wobble 0 BS5 wobble 1 2 2 3 3 3	h Scorn 0 116 120 124 130 130				
Leg Id G 5420376 A Id 51471040 51477462 8922602 21389558 3134140 4639736	end: Orange 3ene p name n JJ238799.1 c Gene name XM_498542. XM_172801.1 MM_174961.1 NM_144669. NM_174941.1 MM_207117.1	 = -2,-1 substrai P1 sequence match AGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT TAGAGCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TGGAGCCAT TGGAGCCAT TGGAGCCAT 	save in Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG	Spacer Spacer 5 0 ■ 10 1 7 2 1 8 0	BS5 mismatch 1 1 1 1 1	BS5 wobble 0 1 2 3 3 3 3 5	h Score 0 Score 116 120 124 130 130 130				
Leg Id G 5420376 A 51471040 514714040 514714040 8922602 8922602 81341400 46397376 7662073	end: Orange 3ene P J238799.1 C Gene name XM_498542. XM_172801.1 XM_172801.1 XM_14465.1 NM_144641.1 NM_207117.1 NM_014872.1	 = -2,-1 substrait P1 sequence match P1 sequence match AGGAGCCAT P1 sequence GAGAGTCAT AGGAGCCAT AGGAGCCAT AGGAGTCAT TGGAGTCAT TGGAGTCAT TGGAGTCAT 	Save in Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG h wobble and mism BS sequence match TCTGCGGAGCCA TCTGTGGGAGCCA TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTAGGGACTG TCTAGGGGTTTG TTTGTGGGGGTCT	Spacer Spacer 5 0 10 1 7 2 1 8 2 2	BS5 mismatch 1 1 1 1 1 1	BSS wobble 0 BSS wobble 1 2 2 2 3 3 3 3 5	h Scorr 0 Scorr 116 120 124 130 130 134 150				
Leg Id C 5420376 A 5420376 A 51471040 51477462 8922602 21389258 83341400 46397376 7662073 C	Gene name Pn W1238799.1 C Gene name XM_498542. XM_172801.1 XM_018185. NM_018185.1 NM_174941. NM_207117 NM_014872.	 = -2,-1 substrait P1 sequence match P1 sequence match P1 sequence match AGGAGCCAT GAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TGGAGCCAT TGGAGCCAT 	Save in Excel te position Blue = wi Perfect match BS sequence match CTCGCGGAACCG h wobble and mism BS sequence match TCTGCGGAGCCC TCTGTGGAGCCA CCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGCGGACCG TCTGCGGACCG TCTGCGGGACCG	Spacer Spacer 5 0 atch Spacer 10 1 7 2 1 8 2 1 8 2	BS5 mismatch 1 1 1 1 1 1 1	BSS wobble 0 BSS wobble 1 2 2 3 3 3 3 5	h 0 3 3 3 3 3 3 3 3 3 3 4 13 4 13 0 13 4 150				
Leg Id C 5420376 A 5420376 A 51471040 51477452 5922602 1384140 44397375 7662073 C Id	end: Orange Gene name XM_498542. XM_172801 NM_018185. NM_144669. NM_014872. NM_014872. Status of the status of the sta	 = -2,-1 substrai P1 sequence match P1 sequence match AGGAGCCAT GAGAGTCAT TAGAGCCAT TAGAGTCAT TGGAGTCAT TGGAGTCAT TGGAGTCAT 	Save in Excel te position Blue = wi Perfect match BS sequence match CTCTGCGGAACCG	Spacer I 5 0 atch Spacer 10 1 7 2 1 8 2 1	BS5 mismatch 1 1 1 1 1 1	BS5 wobble 0 8 2 2 3 3 3 5	A Score 0 116 120 134 130 134				
Legr Id G 5420376 A 5420376 A 51471040 5147142 8922602 1389558 31341400 46397376 762073 762075 762	end: Orange Bene P hame P J238799.1 (C Gene name XM_498542. XM_12801.1 NM_018185. NM_124465. NM_124941. NM_014472. S420376 A3238799.1	 = -2,-1 substral P1 sequence natch AGAGCCAT P1 sequence match 1 AGGAGCCAT 3 GAGAGTCAT 1 TAGAGCAT 1 TAGAGTCAT 2 TIGAAGTCAT 2 TIGAAGTCAT 2 TIGAAGTCAT 1 TIGAAGTCAT 	Save in Excel te position Blue = wr Perfect match BS sequence match TCTGCGGAACCG M wobble and mism BS sequence match TCTGCGGAGCCC TCTGTGGAGCCA CCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGCGGGTGTTG	spacer 5 spacer 10 1 7 2 1 8 8 2	BS5 mismatch 1 1 1 1 1 1 1	BS5 wobble 0 1 2 2 3 3 3 5	Score 0 116 120 124 130 134				
Leg Id C 5420376 A 5420376 A 51471040 51477462 8922602 21389558 3134140 46397376 7662073 C Id Name Descriptio	end: Orange Bane P name P J238799.1 (c C Gene name XM_498542.1 XM_172801.3 NM_018185. NM_144669. NM_1474941.1 NM_207117.7 NM_014872. S420376 A3238799.1 P A3238799.1 NM_014872. S420376 A3238799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_01288799.1 NM_0128879.1 NM_01288799.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_0128879.1 NM_012887.1 NM_01287.	 = -2,-1 substrai P1 sequence match AGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT AGGAGCCAT TAGAGCCAT T	save in Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG h wobble and mism BS sequence match TCTGCGGAGCCC TCTGTGGGACCCG TCTGTGGGACCCG TCGTGGGACCG TCGTGGGACCG TCTGTGGGACCG TCTGCGGGTGTG TCTGCGGGTGTG TCTGCGGGTGTG TCTGCGGGGTCT TTGTGGGGCTCT	spacer 5 0 10 1 7 2 1 8 2 2	BS5 mismatch 1 1 1 1 1 1 1 1 1 1 1 5 5 5 5 5 5 5 5	BSS wobble 0 BSS wobble 1 2 2 3 3 3 5 5	h Score 116 120 124 130 134 150				
Legi Id G 5420376 A 5420376 A 51471040 51471040 51477462 3922602 2012 1389558 31341400 46397376 7662073 C C Id Name Descriptio	end: Orange 3ene n P name n JJ238799.1 c Mu_498542 XM_172801.1 XM_12801.1 XM_018185. NM_014872. S420376 AJ238799.1 n HCJ238799.3 GCCAGCAA	 = -2,-1 substral P1 sequence match AGGAGCCAT Witt P1 sequence match AGGAGCCAT AGGAGCCAT GAGAGTCAT TGGAGCCAT TGGAGCCAT	save in Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG h wobble and mism BS sequence match TCTGCGGACCCA TCTGTGGGACCA TCTGTGGGACCA TCTGTGGGACCG TTGGTGGAACTG TCTATGGGCTGT TCTGCGGTGTTG TCTGCGGGTGTTG TCTGCGGGTGTTG TCTGCGGGTGTTG TCTGCGGGTGTTG TCTGCGCGTGTTG	Spacer 5 C Spacer 10 1 1 2 1 8 2 2	BS5 mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BS5 wobble 0 1 2 2 3 3 3 5 5	h Score 116 120 130 134 150				
Leg Id G 5420376 A 51471040 51477462 8922602 21389558 31341400 46397376 7662073 C Id Name Descriptio	end: Orange Gene name XM_498542. XM_172801. XM_172801. XM_172801. XM_172801. XM_018185. NM_144691. XM_014872. XM_01472. XM_01472. XM_01472. XM_01472.	 = -2,-1 substrai P1 sequence match AGAGCCAT P1 sequence match AGGAGCCAT GAGAGCCAT GAGAGCCAT TAGGAGCCAT TAGGAGCAT TAGGAGCAT<td>stype 1b complete g stype 1b complete g caccactractactactactactactactactactactactactacta</td><td>spacer 5 (Spacer 10 1 1 8 2 2 9 enome, B ATCACTCC GTCCGTGG GTACACTCC</td><td>BSS mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1</td><td>BSS wobble 0 BSS wobble 1 2 2 2 3 3 3 3 5 5 SAACTACTG GGACCCCC</td><td>h Score 0 116 120 124 130 134 150 TCTTC CCCCC CGGGT</td>	stype 1b complete g stype 1b complete g caccactractactactactactactactactactactactactacta	spacer 5 (Spacer 10 1 1 8 2 2 9 enome, B ATCACTCC GTCCGTGG GTACACTCC	BSS mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BSS wobble 0 BSS wobble 1 2 2 2 3 3 3 3 5 5 SAACTACTG GGACCCCC	h Score 0 116 120 124 130 134 150 TCTTC CCCCC CGGGT				
Leg Id C 5420376 A 5420376 A 51471040 51477462 8922602 21389558 31341400 46397376 7662073 C Id Name Descriptio	Bene P Gene name P Multiple Multiple V238799.1 C KM_498542. XM_172801.1 XM_172801.1 XM_17481.1 NM_018185. NM_14469.1 NM_124911.1 NM_207117 NM_014872. S420326 A1238799.3 GCCAGCCC ACGGGAGAG GCCAGCCC ACGCAGAA CCTITCTIT	 = -2,-1 substrai P1 sequence match P1 sequence match P1 sequence match AGGAGCCAT GAGAGCCAT GAGAGCCAT TAGAGCCAT GAGAGTCAT TAGAGCCAT GAGAGTCAT TAGAGCCAT TAGAGCAT TAG	stype 1b complete g Stype 1b	Spacer 5 (Spacer 10 11 7 2 1 1 8 2 2 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	BS5 mismatch 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BSS wobble 0 BSS wobble 1 2 2 3 3 3 3 5 SAACTACTG GGACCGCC	A Score 0 0 116 120 124 130 130 134 150 150				
Leg Id G 5420376 A 5420376 A 51471040 5147142 8922602 21389558 31341400 46397376 7662073 C Id Name Descriptio	end: Orange Bene p hame p J238799.1 (C Gene name XM_498542. XM_12801.1 NM_124465. NM_124461. NM_018185. NM_124461. NM_0144669. S420376 A1238799.1 GC2A3C79 GC2A3C	 -2,-1 substral P1 sequence match AGGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT AGGAGCCAT TAGAGCCAT TAGAGCCAT TGGAGTCAT TGGAGTCAT TGGAGTCAT TGGAGTCAT TGGAGTCAT TGGAGTCAT TGGAGTCAT GGAGTCACCGGCT GGATCAACCGGCT GGACCGCCCGAAC 	stype 1b complete g CCTGCGGAACCG b wobble and mism BS sequence match TCTGCGGAACCG tCTGTGGGAGCCC TCTGTGGGAGCCA TCTGTGGGAGCCA TCTGTGGGAGCCG TTGGTGGAGCCG TTGGTGGGACTG TCTGCGGGGTGTG TTTGTGGGGGTGTG TTTGTGGGGGCTGTG TGGCGGAACCGGGAACTC	spacer 5 (10) 10 1 7 2 1 1 8 2 2 9 enome, li 3 1 8 2 2	BS5 mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BS5 wobble 0 1 2 2 3 3 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	h Correction Cor				
Leg Id C 5420376 A 5420376 A 51471040 51477462 8922602 21389558 3134140 46397376 7662073 C Id Name Descriptio	end: Orange Bane P hame P J238799.1 (c C C C C C C C C C C C C C	 = -2,-1 substral P1 sequence match AGGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT AGGAGCCAT TAGAGCCAT GAGAGCCAT GAGAGCCAT GAGAGCCAT GAGACCAT GAGACCATCAT GGGCTACACCGCCAAA AACCGTCCACCAC 	swe n Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG TCTGCGGAACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGTGGGACCG TCTGCGGTGTTG TCTGCGGTGTTG TCTGCGGTGTTG TCTGCGGTGTTG TCTGCGGTGTTG TCTGCGGTGTTG TCTGCGGACCGTGA CCCGTGGGACCGGTGA TGGCGTTGTGTATGAGACTG TGGCCTTGTGGTACCG TGGGCTTGTGGTACCG TGGGCTTGTGGTACCG TGGGCTTGTGGTACCG	Spacer 5 0 10 1 1 7 2 1 8 8 2 2	BSS mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BSS wobble 0 BSS mobble 1 2 2 3 3 3 3 3 5 S BAACTACTG GGACCCCCC AGGACGAC AGGACGAC AGGACGAC CAAACGTA	A Correction of the second sec				
Legi Id G 5420376 A 5420376 A 51471040 51471040 51471040 51471040 51471040 51471040 51471040 51471040 51471040 51471040 7662073 C C Id Name	end: Orange Gene name P Mul 498542. XM_498542. XM_172801.1 XM_12801.1 XM_12801.1 XM_12801.1 XM_12801.1 XM_14460. XM_174941.1 MUL 207117.1 MUL 20	 = -2,-1 substral P1 sequence match AGGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT AGGAGCCAT GAGAGTCAT TGGAGCCAT TGGAGCAGCAT TGGAGCAGCAA TGGAGCAGCAA TGGAGCAGCAA TGGAGCAA TGGAGCAA TGGAGCAA TGGAGCAA TGGAGCAA TGGAGCAA TGGAGCAAA TGGAGCAAAA TGGAGCAAAA TGGAGCAAAAA TGGAGCAAAAAA TGGAGCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	Save in Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG TCTGTGGGACCG TCTGTGGGACCG TCGTGGGACCG TTGGTGGAACCG TCTATGGGCTTGTGGGGTCT is type 1b complete of CGACACTCCACCATAGG TGGCGTTAGTATGAGTACGGTACTGAATCCAATAGCCTGAGATCCAATAGCACGAATCCTA	Spacer 5 0 Spacer 10 10 1 7 2 1 8 2 2 Spacer 10 1 8 2 2 Spacer 10 1 1 7 2 1 8 2 2 Spacer 10 1 1 1 1 1 1 1 1 1 1 1 1 1	BS5 mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BS5 wobble 0 1 2 2 3 3 3 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5	h Scorr 116 120 130 134 130 134 150 TCTTC CCGGGA ACCCC				
Legi Id G 5420376 A 51471040 51477462 8922602 21389558 31341400 46397376 7662073 C Id Name Descriptio	end: Orange Gene name XM_498542. XM_172801. XM_172801. XM_172801. XM_172801. XM_172801. XM_174815. NM_144669. XM_144669. XM_14469. XM_14469. XM_14461. XM_207117. XM_014872. XM_14981. XM_498542 XM_	 = -2,-1 substral P1 sequence match AGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT GAGAGTCAT TGGAGCCAT TGGAGTCAT TGGGTCGCGAA AGCCGTCACCCGCA AGACCGTGCACC 2.1 	Save in Excel te position Blue = w Perfect match BS sequence match CTGCGGAACCG tTGGGAACCG TCTGTGGGACCCG TCGTGGGACCCG TCGTGGGACCG TCGTGGGACCG TCGTGGGACCG TCGTGGGACCGG TCGGCGTGTGGGGTCT GCGCATCGCCGGACCGGGCA tGGCCTTAGTATGAGC	Spacer 5 (Spacer 10 1 1 7 2 1 8 2 2 9 enome, B ATCACTCC GTCCTGG GTCCACC GTCCACC GTCCACC GTCCACC GTCCACC GTCCACC GTCCACC GTCCACC GTCCACCACC GTCCACCACCACCACCACCACCACCACCACCACCACCACCA	BSS mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BSS wobble 0 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8	h Contraction Co				
Leg Id G 5420376 A 5420376 A 51471452 8922602 21389558 31341400 46397376 7662073 C Id Name Descriptio	end: Orange Bane P hame P J238799.1 C Gene name XM_498542. XM_12801.1 M_127801.1 M	 = -2,-1 substrai P1 sequence match P1 sequence match P1 sequence match AGGAGCCAT GAGAGCCAT GAGAGCCAT TGGAGCCAT GGGAGCCAT GGGAGCCAT GGGAGCCAT GGGAGCCAT GGGAGCCAT TGGAGCCAT GGGAGCCAT GGGAGCCAT GGGAGCCAT GGGAGCCAT GGGAGCCAT GGGTCACTAACCGCC E Homo saplens GTCACTTACGGG 	Save in Excel te position Blue = wi Perfect match BS sequence match BS sequence match TCTGCGGAACCG CTGTGGGACCG CTGTGGGACCG CTGTGGGACCG TTGGTGGAGCGA CCTGTGGGACGG TTGGGGGACGG TTGGCGGACCCGGTGA CCTGTGGGGGCTGTGGTGTGG	Spacer 5 (10) 10 11 7 2 11 8 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	BS5 mismatch 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Missmatc BSS wobble 0 BSS wobble 1 2 2 2 3 3 3 3 3 3 3 5 SAACTACTG GGACGACCCCCC AGGACGACTACTGCT CGAGTGCCC CCAAACGTA	h Correction 116 120 130 134 130 134 130 134 150				
Legr Id G 5420376 A 5420376 A 51471040 51471040 51471040 8922602 21389558 31341400 46397376 7662073 C Id Name Descriptio	end: Orange	 -2,-1 substrai P1 sequence match AGAGCCAT P1 sequence match AGAGCCAT AGGAGCCAT GAGAGCCAT GAGAGCCAT TAGAGCCAT TAGAGCCATAGCC TAGAGCCATCAC Homo sapiens SIGCTACTACCGCCATCACC 	swe n Excel Perfect match Bs sequence match TCTGCGGAACCG twobble and mism Bs sequence match TCTGCGGAACCG TCTGTGGAACCA CCTGTGGAACCAGACA CCTGTGGGACCG TCGTGGGACCG TCGTGGGACCG TCGTGGGACCG CCTGTGGGACCG CCGTGGGACCGGTAG CCGTGGGACCGGTAG CCGCTTAGTATGAG CCGGTGGACCGGTAG CCGCTTAGTATGAGATT CGCCTTGTGTACAGATT CGCCTTGTGACCAGACA CCGGTGGACCGGTAGATCG TCGCGACCGGTAGATCG CGCACCCACACAGACA CCGGTGCACCCACCACAGA CCGGTGCACCCCACAGACA CCGGTGCACCCCCCACAGA CCGGTGCACCCCCACAGA CCGGTGCACCCCCCACAGA CCGGTGCACCCCCCCACAGA CCGGTGCACCCCCCACAGA CCGGTGCACCCCCCCATAG CGCCTTGTGTACGGACCG CCGCTCCCCCCCATAG CCGCTCCCCCCCATAG CCGCTCCCCCCCACAGA CCGGTGCACCCCCCCCACAGA CCGGCCCCCCCCCC	Spacer 5 0 Spacer 10 10 1 7 2 1 8 8 2 2 2 2 2 2 2 2 2 2 2 2 2	BSS mismatch 0 BSS mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BSS wobble 0 BSS wobble 1 2 2 3 3 3 3 5 SAACTACTG GGACCCCC AGGACGAC AGGACGCC CCAAACGAA AGAAGCTG	A Score 116 120 120 130 130 134 150				
Leg Id C 5420376 A 5420376 A 51471040 51477462 8922602 I389558 3134140 46397376 7662073 C C C C Descriptio Id Name Descriptio	end: Orange	 = -2,-1 substral P1 sequence match AGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT AGGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TAGAGCCAT TGGAGTCAT TGGAGCCAT TGGAGCCAT TGGAGCCATGACAACCGGCAAA CCATAGGGTC CCATAGGGTC CCATAGGGTC TGGAGCCATGACAACCGGCAAA CCACTTACGGG CATCACCGCAATGACAACGGCCAATGACAATCGGTGAACAA 	Save in Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG TCTGTGGGAACCG TCTGTGGGAGCCA TCGTGGGACCG TTGGTGGAACCG TCTATGGGACTG TCTATGGGATCTG TCTATGGGATCTG Stype 1b complete g GACACTCCACCATAGG TGGGAACCGGTGATGACTCGGAACCCGGTGATGGACCCGTGGAACTGA CCTGGGACCGGTGATGGAGCACGGTGATGGAGCACGGTGGAACTGGAACTCGACGATGGAACTGGAACTGGAACTCGTGGAACTCT	Spacer 5 Spacer 10 10 1 7 2 1 8 2 2 1 8 2 2 1 1 8 2 2 1 1 8 2 2 1 1 1 7 2 1 1 8 2 2 1 1 1 1 7 2 1 1 1 2 2 1 1 1 2 2 1 1 1 2 2 1 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2	BS5 mismatch 0 BS5 mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BS5 wobble 0 BS5 wobble 1 2 2 2 3 3 3 3 3 3 3 3 5 5 S ACTACTG GAACGAC AGAACGAC AGAACGAC AGAACGAC	A Score Control Control Contr				
Legi Id C 5420376 A 5420376 A 5420376 A 51471040 51471040 51471462 3022602 202 2022 2	end: Orange Gene name XM_498542. XM_172801.1 XM_498542. XM_172801.1 XM_498542. XM_172801.1 XM_498542. XM_172801.1 XM_498542. XM_498542. XM_498542. XM_498542. A1238799.1 GCCAGCCC A1238799.1 GCCAGCCC A1238799.1 S1421040 S14210	 = -2,-1 substral P1 sequence match AGGAGCCAT P1 sequence match AGGAGCCAT AGGAGCCAT AGGAGCCAT GAGAGTCAT TGGAGCCAT TGGAGCCATGGCGAA AGCCGTGACCACCGCC THOMO Saplens TGGACCATGACAAGCGCGCAACGCGC THOMO Saplens TGGACCACAGCG THOMO Saplens TGGACCACAGCGCGCAACGCGCGCACACGCG 	Save in Excel te position Blue = wi Perfect match BS sequence match TCTGCGGAACCG TCTGTGGGACCG TCTGTGGGACCG TCGTGGGACCG TGGTGGAACCG TCGTGGGACCG TGGTGGAACTG TCTATGGGACTG TCTATGGGACTG TGGCGTATGTGATGAGT CCAACGGCGCGCA	Spacer Spacer Spacer 10 10 1 7 2 1 8 2 2 2 2 2 2 2 2 2 2 2 2 2	BS5 mismatch 0 BS5 mismatch 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	BS5 wobble 0 1 2 2 2 3 3 3 3 3 5 5 5 5 5 5 5 5 5 5 5 5	h Control Control Cont				

Figure 2

Design of RiboSubstrates with the parameters for the SOFA-HDV ribozyme. (**A**) Secondary structure and nucleotide sequence of the SOFA-HDV ribozyme targeting an mRNA. The PI stem of the ribozyme and the biosensor (BS) of the SOFA module are identified. The letter N indicates the presence of an A, C, G or T residue. The arrow indicates the cleavage site. (**B**) The RiboSubstrates input with the signature to be searched for the SOFA-HDV-Rz targeting the HCV virus (SOFA-HDV-Rz-135) and the substrate display table. Mismatched nucleotides and Wobble base pairs are identified in red and blue, respectively. (**C**) The more detailed tables for both the mRNA encoding the HCV genome 1b and the predicted *Homo sapiens* LOC440096, respectively.

Α		Search f	or <u>Ha</u>	mn	nerhea	<u>d</u> substrate	s			
-	S	<u>iee search</u>	exam	ple	Note	s on databa	ses			
Search N	lame:	hammerhead	BCL2	targ	ets					
cDNA Ba	Bank: Human NCBI build 35.1 mRNAs							•		
Stem I s	equence	GGTCAGGTG								
stem I n	nismatch									
Stem III	sequence		GT							
Stem III	mismatch	0								
Email:		C Send e	maily	vho	n reguli	te are ready				
		 Wait for 	or resu	ults	here	to are ready				
			Sul	hmit	Roco	• 1				
			30	<u>.</u>	.5'					
				N	N N					
			Stem III	N- N-	N N					
				N- N-	N N					
				N- A-		NNNNN				
		C	11111	U		NNNNN 5'				
				\sim	Ste	em I				
В										
			Sea	rch	paramet	ers				
		Stem I sequ	atch	GG	TCAGGTG					
		Stem III see	quence	DA	CCACAGG	т				
		Stem III mi	smatch	0						
		cDNA bank		Hur	man NCBI	build 35.1 mR	NAs			
				Sav	e in Excel					
Le	egend: Orange	= UH substr	ate pos	erfe	ect match	wobbie base pa	air Red = mi	ssmatch		
		Stem I	Stem I		Stem I	Stem III	Stem III	Stem		
Id	Gene name	match	misma	tch	wobble	sequence	mismatch	III wobble	Score	
4557354	NM_000633.1	CACCTGACC	0	_	0	ACCTGTGGTC	0	0	0	
				With	wobble					
Id	Gene name	Stem I	Sten	١	Stem I	Stem III	Stem III	Stem	Score	
	oene name	match	misn	atc	wobble	sequence	mismatch	wobble	Con	
4557356	NM_000657.1	CACCTGGC	0		1	ACCTGTGGTC	0	0	10	
40538803	NM_170776.		το		3	GTTTGTGGT		3	80	
-	pan_010101.1				9	Ginorodi		5	00	
C										
Id	4557354 NM 000633	1								
Descriptio	Homo saple	ens B-cell CLL	/lymph	oma	2 (BCL2)	, nuclear gene	encoding mi	tochondri	al	
Descriptio	protein, tra	nscript variar	nt alpha	, mF	AVIS					
	GTTGGCCCC	GTGATGAAGT	ACATCO	GGG CATT	AAATATG ATAAGCT	GCGCACGCTGG GTCGCAGAGGG	GAGAACAGO	GTACGA TGGGATG	CGGG	
	GATGTGGG	GATGTGGGCGCCGCGCCCCGGGGGCCGCCCCCGGGCATCTTCTCCTCGCAGCCCGGGCA								
	CCCCCGGCC	GCCGCCGCGC	GGCCT	GCG	CTCAGCO	CGGTGCCACC	IGTGGTCC	ACCTGA	CCCT	
Id	4557356									
Name	NM_000657	NM_000657.1								
Descriptio	protein, tra	ns B-cell CLL nscript varian	/iympni it beta,	ma mRì	2 (BCL2), VA	, nuclear gene e	encoding mi	tochonari	ai	
	TGATTGAAG	TGATTGAAGACACCCCCTCGTCCAAGAATGCAAAGCACATCCAATAAAATAGCTGGATTATAACT								
	TTTTCCTCT	CCTCTTCTTCTCTGGGGGCCGTGGGGGGGGGGGGGGGG								
	AGTACATCO	AGTACATCCATTATAAGCTGTCGCAGAGGGGCTACGAGTGGGATGCGGGAGATGTGGGCGCCGCG								
	CGCATCCCC	CGCATCCCGCGACCCGGTCGCCAGGACCTCGCCGCTGCAGACCCCGGCGCCGCCGCGCGCCGCGCGCCGCGCGCCGC								
	CGGGGCCT	GCGCTCAGCC	CGGTG	CCA	сстбтб	GTCCACCTGG	CCCTCCGC	CAAGCCG	GCGA	
Id	40538803									
Name	NM_170776	.3		1.45	anter of	(CDD.07)				
Descriptio	GCAGGGTG	ns G protein- GGGGCAGGCC	COUPIEC	a reo	eptor 97 AGAGCCT	GGGGCCAGAGG	GCCAGACA	SCCACAG	AGCTO	
	TGGCGTGG	TGCGTGGGCAAGGCTGGCCAAGGATGGCGACGCCCAGGGGCCTGGGGGCCCTGCTCCTGCTCCT								
	CCTGCTCCC									
	GGCAGCGAG	GGCAGCGACTCCTGCAATGTGGAAAACTTGCAGAGATACTGGCTAAACTACGAGGCCCATCTGAT								
	CCAACACTG	UAAUUAAUUIIIIGAUGUAGAAGGIGAAUGUUTTTCCTGAAGGCTTTGGTCCAGAACCTCAGCA CCAACACTGCAGAAGACTTCTATTTCTCTCTGGAGCCCTCTCAGGTTCCGAGGCAGGTGATGAAG								
	GACGAGGAG	GACGAGGACAAGCCCCCTGACAGAGTGCGACTTCCCAAGAGCCTTTTTCGATCCCTGCCAGGCAA								
Id	33695094	Generati	1990	GI	CAUCATI	CIGGACATIGG	CCAGGGAC	i ci ci i ci	-04G	
Name	NM_016131	.2								
Descriptio	n Homo sapie	ns RAB10, m	ember f	RAS	oncogene	family (RAB10), mRNA			
Descriptio										

Figure 3

Design of RiboSubstrates with the parameters for the hammerhead ribozyme. (**A**) Secondary structure and nucleotide sequence of the hammerhead ribozyme targeting an mRNA. The two binding domains of the ribozyme are identified. The letter N indicates the presence of an A, C, G or T residue. The arrow indicates the cleavage site. (**B**) The RiboSubstrates input with the signature to be searched for the hammerhead ribozyme targeting the *Homo sapiens* Bcl-2 mRNA (hhRz-Bcl-2-1). Mismatched nucleotides and Wobble base pairs are identified in red and blue, respectively. (**C**) The substrate display table retrieved for this ribozyme. stems: i. the P1 stem that is composed of one GU Wobble base pair followed by six consecutive Watson-Crick base pairs; and, ii. the biosensor composed by ten to fifteen Watson-Crick base pairs (BS; see Figure 2A). The sequence located between these two domains within the substrate is referred to as "the spacer", and its optimal length has been determined to be 1 to 6 nucleotides [16]. The presence of a longer spacer sequence provides a substrate that should be the least cleaved.

As an example, we investigated the specificity of a Ribo-Substrate selected SOFA-HDV ribozyme targeting the internal ribosome entry site (IRES) of the hepatitis C virus (HCV) [16]. Once a configuration file (RiboParticularitites.xml) was created, the HDV-ribozyme was automatically integrated in RiboSubstrates. This configuration file includes specific information, such as the required presence of a guanosine residue at the 3'-end adjacent to the cleavage site, so as to permit formation of the essential GU Wobble base pair. The sequences of the P1 stem and biosensor (BS) were AUGGCUU and CGGUUCCGCAGA, respectively (SOFA-HDV-Rz-HCV-135). The maximal length of the spacer sequence was fixed at 10 nucleotides. These sequences were verified to be unique to the HCV genome. The number of mismatches permitted in the BS was 3, and the spacer (i.e. the distance between the P1 and BS stems) length varied between 1 and 10 nucleotides. As this ribozyme is to be used in human cells, an updated version of the Homo sapiens cDNA database, including an entry for a HCV type 1b complete genome, was selected and the search performed. The RiboSubstParticularities.xml files for the score calculation had some specific indications for SOFA-HDV ribozyme. For example, a score was included for the spacer length according to biochemical data published previously [15,16]. A spacer length between 1 and 6 nucleotides has a spacer score of 0; both the lack of a spacer (i.e. the P1 and BS sequences on the substrate are contiguous) and a spacer length between 7 and 9 has a spacer score of 4. An unspecified length is automatically attributed a spacer score of 10. The substrate display is illustrated in figure 2B (lower section). The designed ribozyme has only one perfect hit, specifically the HCV genome. After this entry, the lowest score was 116 for a predicted Homo sapiens mRNA (LOC440096) (Figure 2B and 2C). This hit possesses one mismatch and one Wobble base pair with SOFA-HDV-Rz-HCV. Biochemical experiments have shown that the presence of one mismatch causes a significant reduction of the cleavage activity of a SOFA-HDV ribozyme [15]. Therefore, this ribozyme should be specific for targeting the HCV IRES. In addition they were many other hits, with higher scores, that included several mismatches as well as Wobble base pairs. When the RiboSubstrates experiment was repeated using the original human database (i.e. without adding the HCV sequence) no perfect hit was retrieved. More importantly, several experiments have been performed both *in vitro* as well as in cultured cells (unpublished data, F Brière and JP Perreault). In each case SOFA-HDV-Rz-HCV-135 exhibited efficient cleavage of the HCV IRES and all data indicated that this ribozyme has a specific action. More importantly, RiboSubstrates had suggested that this SOFA-HDV ribozyme should be specific as was observed experimentally. In so doing Ribo-Substrates prevented the lose of time designing other HCV specific ribozymes that also cleave human mRNAs.

A second example was developed for the hammerhead ribozyme since to date it has been the catalytic RNA motif receiving the greatest amount of attention in terms of the development of gene-inactivation systems [18]. This ribozyme possesses a substrate recognition domain that includes two stems located on either side of the cleavage site (i.e. stems I and III; Figure 3A). The sequence of stem I (i.e. the 5'end of the ribozyme) has no specific requirement. Conversely, the sequence of stem III, which corresponds to the 3' end of the ribozyme, starts with a required adenosine that base-pairs with the uridine adjacent to the residue located at the cleavage site. The latter residue can be adenosine, uridine or cytosine (symbolized by an H). Specific RiboParticularities.xml files describing the hammerhead signature to be searched for, and the substrate display, were appended to a previously written configuration file. An example of the search was performed for a hammerhead ribozyme targeting the human apoptosis-associated B-cell CLL/lymphoma 2 (Bcl2) mRNA (Figure 3B and 3C; [19]). The results display shows that, in addition to targeting the Bcl2 mRNA with a perfect score, this ribozyme would also have the potential to target both another variant of Bcl2 (score of 10) and the G protein-coupled receptor 97 (score of 50). Such a result provides the information required for the research decide whether or not to go further with this ribozyme, or to design another ribozyme that would be more specific for the Bcl2 mRNA. At a minimum, this result should raise a red flag, suggesting to the researcher that he should consider other information, such as if it is known that alternative targets are also expressed in the cell cultures used, before proceeding.

Finally, a template was developed for the siRNA search. Since the current knowledge does not provide an unambiguous description of the features that define the rules for siRNA specificity, only one parameter was considered. The minimal size of the siRNA should be at least 19 nucleotides. As rules become available, this signature file will be upgraded, which explains the indication "in development" appearing below the siRNA template available in the software.

Conclusion

The avoidance of any off-target effect (i.e. non-specific cleavage) is crucial to the development of a useful RNAbased tool for gene inactivation. Given this fact, RiboSubstrates is an essential tool for the rapid selection of sequences suitable as targets for RNA degradation. This flexible and reliable program is easily adaptable for any RNA's tool specifications, and may use information from any cDNA bank. In addition the same framework could be used for selecting targets for other ribozymes, deoxyribozymes and antisense-based RNA tools. Requests for the additional development of RiboParticularities.xml files for other RNA tools, and the addition of cDNA databases, will be performed primarily in collaboration with users and the authors (please contact Jean-François Lucier or Jean-Pierre Perreault). Alternatively, we are willing to share the code to all interested scientists.

Availability and requirements

The RiboSubstrates resource is platform that is independent and freely available for all non-commercial users at <u>http://www.riboclub.org/ribosubstrates</u> The software has been programmed in PERL and is licensed under GNU LGPL. This web site is the property of the Sherbrooke Ribo-Club group and includes many other interesting tools and information, as well as a forum dedicated to RNA scientists. The cDNA databases will be updated frequently as new versions become available.

List of Abbreviations

BCL2 Human apoptosis-associated B-cell CLL/lymphoma 2

HCV Hepatitis C Virus

HDV Hepatitis D Virus

SOFA Specific On/oFf Adaptor

Authors' contributions

JFL designed the software, participated in all phases of research and contributed to write the manuscript; LJB and FPB contributed to software design and testing; RO and SAB provided intellectual contributions and revised the manuscript; JPP initiated the idea, provided intellectual contributions, supervised all aspects of the work and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank G Robichaud, P Landry and MC April for their numerous discussions as well as J. Brassard and JC Houde for programming. This work was supported by grants from Genome Québec to SAE and JPP, an Atlantic Canada Opportunities Agency grant to RO and a Canadian Institute of Health Research (CIHR; EOP-38322) to JPP. The RNA group is supported by a grant from CIHR (PRG-80169) and Université de Sherbrooke. LJB was the recipient of a pre-doctoral fellowship from CIHR. SAE is a Chercheur Boursier Senior of the Fonds de la Recherche en Santé du Québec. JPP holds the Canada Research Chair in Genomics and Catalytic RNA.

References

- I. Breaker RR: Natural and engineered nucleic acids as tools to explorer biology. *Nature* 2004, **432**:838-845.
- Grunweller A, Hartmann RK: RNA interference as a gene-specific approach for molecular medicine. Curr Med Chem 2005, 12:3143-3161.
- 3. Akashi H, Matsumoto S, Taira K: Gene discovery by ribozyme and siRNA librairies. *Nat Rev Mol Cell Biol* 2005, 6:413-422.
- Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS: Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol 2003, 21:635-637.
- 5. Haley B, Zamore P: Kinetic analysis of the RNAi enzyme complex. Nat Struct Mol Biol 2004, 11:599-606.
- Birmingham A, Anderson EM, Reynolds A, Ilsley-Tyree D, Leake D, Fedorov Y, Baskerville S, Maksimova E, Robinson K, Karpilow J, Marshall WS, Khvorova A: 3' UTR seed matches, but not overall identity are associated with RNAi off-targets. Nat Methods 2006, 3:199-204.
- Bergeron LJ, Ouellet J, Perreault JP: Ribozyme-based gene-inactivation systems require a fine comprehension of their substrate specificities; the case of delta ribozyme. Curr Med Chem 2003, 10:2589-2597.
- Snove O Jr, Nedland M, Fjeldstad SH, Humberset H, Birkeland OR, Grunfeld T, Saetrom P: Designing effective siRNAs with off-target control. Biochem Biophys Res Commun 2004, 325:769-773.
- get control. Biochem Biophys Res Commun 2004, 325:769-773.
 9. Naito Y, Yamada T, Ui-Tei K, Morishita S, Saigo K: siDirect: highly effective, target-specific siRNA design software for mamma-lian RNA interference. Nucleic Acids Res 2004, 32:W124-W129.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25:3389-3402.
- Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W: WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res 2003, 31:3795-3798.
- 12. Smith TF, Waterman MS: Identification of common molecular subsequences. J Mol Biol 1981, 147:195-197.
- 13. The Perl directory [http://www.perl.org/]
- 14. CPAN: Comprehensive Perl archive network [http:// cpan.org/]
- Bergeron LJ, Perreault JP: Target-dependent on/off switch increases ribozyme fidelity. Nucleic Acids Res 2005, 33:1240-1248.
- Bergeron LJ, Reymond C, Perreault JP: Functional characterization of the SOFA delta ribozyme. RNA 2005, 11:1858-1868.
- 17. Fiola K, Perreault JP, Cousineau B: Gene targeting in the grampositive bacterium Lactococcus lactis using various delta ribozymes. Appl Env Microbiol 2006, 72:869-879.
- Blount KF, Uhlenbeck OC: The structure-function dilemma of the hammerhead ribozyme. Annu Rev Biophys Biomol Struct 2005, 34:415-440.
- Akgul C, Moulding DA, Edwards SW: Alternative splicing of Bcl-2 related genes: functional consequences and potential therapeutic applications. *Cell Mol Life Sci* 2004, 61:2189-2199.