BMC Genomics

BioMed Central

Software

Open Access

FragIdent – Automatic identification and characterisation of cDNA-fragments

Dominik Seelow[1,2], Heike Goehler[3,4] and Katrin Hoffmann*[1,5]

Address: [1]Institut für Medizinische Genetik, Charité – Universitätsmedizin Berlin, Augustenburger Platz, 13353 Berlin, Germany, [2]Department of Neuropaediatrics, Charité – Universitätsmedizin Berlin, Augustenburger Platz, 13353 Berlin, Germany, [3]Max Delbrück Center für Molekulare Medizin, 13125 Berlin, Germany, [4]Ruhr-Universität Bochum, Medizinisches Proteom-Center, 44801 Bochum, Germany and [5]Max-Planck-Institut für Molekulare Genetik, Ihnestr. 73, 14169 Berlin, Germany

E-mail: Dominik Seelow - dominik.seelow@charite.de; Heike Goehler - Heike.Goehler@rub.de;
Katrin Hoffmann* - katrin.hoffmann.genetik@charite.de
*Corresponding author

Published: 02 March 2009

BMC Genomics 2009, **10**:95 doi: 10.1186/1471-2164-10-95

Received: 23 July 2008

Accepted: 2 March 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/95

Abstract

Background: Many genetic studies and functional assays are based on cDNA fragments. After the generation of cDNA fragments from an mRNA sample, their content is at first unknown and must be assigned by sequencing reactions or hybridisation experiments.

Even in characterised libraries, a considerable number of clones are wrongly annotated. Furthermore, mix-ups can happen in the laboratory. It is therefore essential to the relevance of experimental results to confirm or determine the identity of the employed cDNA fragments. However, the manual approach for the characterisation of these fragments using BLAST web interfaces is not suited for larger number of sequences and so far, no user-friendly software is publicly available.

Results: Here we present the development of FragIdent, an application for the automatic identification of open reading frames (ORFs) within cDNA-fragments. The software performs BLAST analyses to identify the genes represented by the sequences and suggests primers to complete the sequencing of the whole insert. Gene-specific information as well as the protein domains encoded by the cDNA fragment are retrieved from Internet-based databases and included in the output. The application features an intuitive graphical interface and is designed for researchers without any bioinformatics skills. It is suited for projects comprising up to several hundred different clones.

Conclusion: We used FragIdent to identify 84 cDNA clones from a yeast two-hybrid experiment. Furthermore, we identified 131 protein domains within our analysed clones. The source code is freely available from our homepage at http://compbio.charite.de/genetik/FragIdent/.

Background

cDNA clones or the cDNA contained in them are frequently used in yeast two-hybrid assays [1] and hybridisation studies [2]. Although whole clones are employed in some hybridisation studies [3], usually only the insert or a fragment of it is used as a probe. The DNA can be obtained either by amplification of the corresponding plasmid, by insert or vector specific PCR reactions.

While the contents of clones experimentally derived from complex mRNA samples are necessarily unknown, even clones from characterised libraries are in many cases wrongly annotated [4] or might have been mixed up in the laboratory. To draw conclusions from experiments involving such clones, it is inevitable to sequence at least a part of the insert to confirm its identity. A conventional approach is to determine the identities of the clones by sequencing from one or both ends using primers specific to the vector sequence [5-7]. Subsequently, the obtained sequence is matched to annotated sequences in public databases to identify the corresponding mRNA or protein. Although such an initial sequencing is in most cases enough to ascertain the gene encoded by the ORF, it may not cover the coding region completely. In such cases, it is impossible to determine the transcript variant, to detect new transcripts hitherto unknown [8] or to spot 'contaminated' clones that contain sequences not present in the original gene. Especially in yeast two-hybrid screens, it may be crucial to gain knowledge not only of the protein encoded but also of the functional domains covered by the actual clone since interactions are often mediated by protein domains [9].

In these cases, it is therefore indispensable to sequence the whole insert. This can be achieved by successive sequencing reactions with primers aligning at the end of the prior sequence until the vector sequence or a stop codon is reached ("primer walking"). To construct the sequence of the clone, the obtained sequences have to be merged by aligning the overlapping part of the sequences. Alternatively, after the initial sequencing, primers can be designed in advance when the encoded gene is known. If the size of the insert is experimentally determined, the primer design can be restricted to the estimated region.

Whatever method is chosen, the steps performed by the researcher turn out to be a tremendous work when carried out manually for a large number of clones. In addition, the manual alignment of clone sequences to DNA, mRNA or protein databases bears the risk of copy and paste errors as well as the accidental use of different BLAST settings. Also, the generation of suitable insert-specific sequencing primers can require a huge effort when large numbers of clones have to be sequenced, each with multiple primers.

For the systematic analysis of DNA fragments, several bioinformatics tools are freely available. However, some of these tools are either addressed at dedicated bioinformaticians (e.g. EMBOSS [8]) or are specific for other purposes, such as the SABIA system for bacterial genomes (SABIA [9]). Other applications focus on EST sequences, and although they are useful to identify and characterise genes contained in a cDNA clone (AutoFACT [10], EST-PAC [11], OREST [12]), they cannot easily be employed to judge the length of the insert in a cDNA clone nor design new primers needed to sequence the whole insert. EST Express [13], on the other hand, can discriminate between full length and partial sequences and even provides filters for vector sequences but is a comprehensive clone management database which requires a complex installation and might therefore be oversized for projects which comprise less than a few thousand clones and do only require sequence analysis. However, since the use of any such tool is only rarely been mentioned in manuscripts, apparently BLAST analyses are currently mainly carried out either by copying and pasting sequences into one of the BLAST web interfaces or by proprietary software that is not specified in the respective publications and not publicly available. To meet the challenge to analyse more than 80 clones from a yeast two-hybrid experiment, we developed FragIdent, a software that combines the single steps into a single application. Our approach provides a user-friendly interface that guides the researcher through the single steps necessary to identify and to further characterise the cDNA fragments, hence making larger analyses feasible for researchers without any bioinformatics skills.

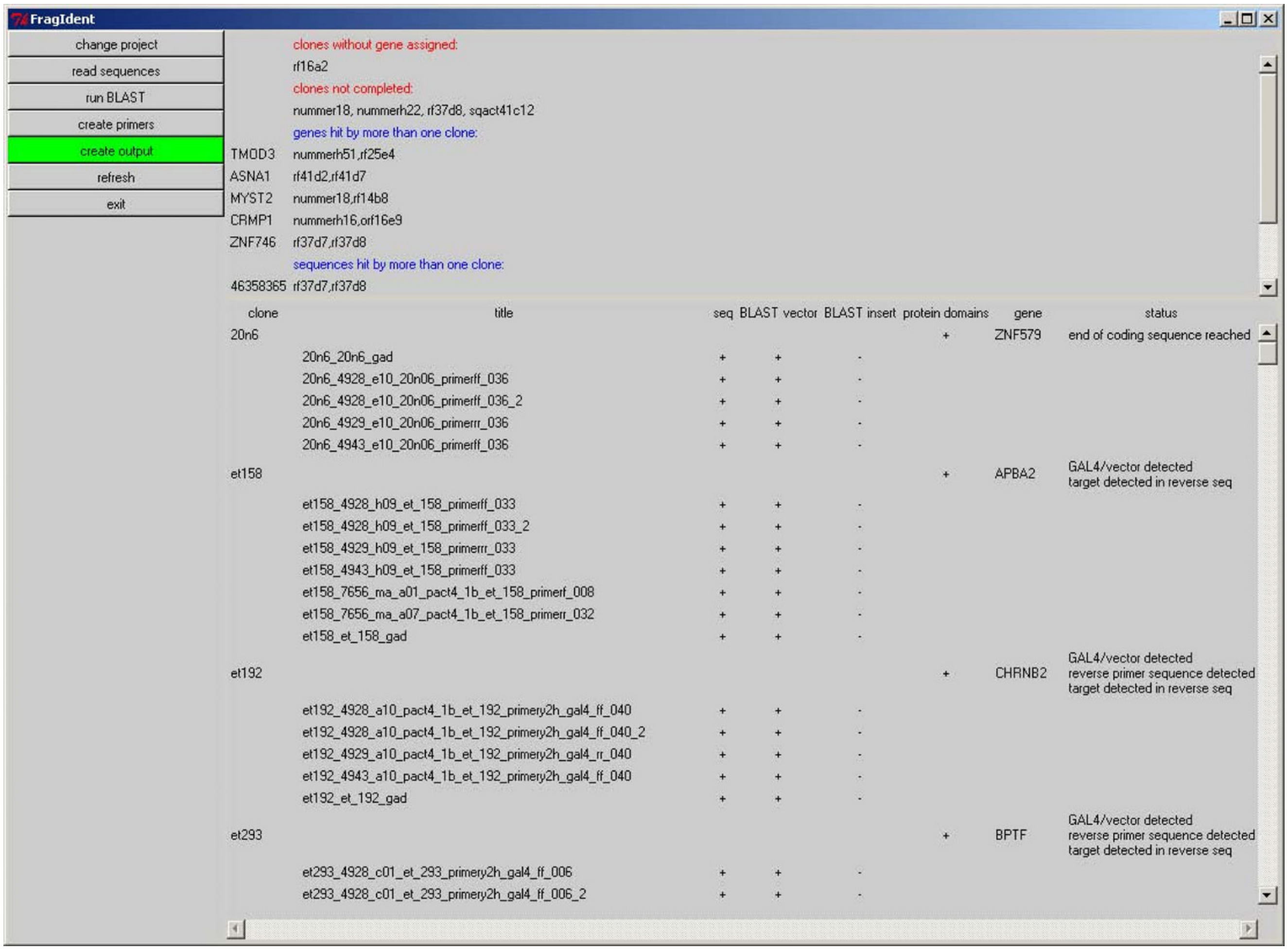
Implementation

FragIdent was programmed in Perl/Tk and is freely available. It uses Primer3 [14] for primer design. BLAST analyses are performed via the Internet at the NCBI [15,16]. Basic gene specific data is also obtained from the NCBI. The GeneDistiller database [17] is queried via the Internet to include more detailed information such as protein-protein interactions and reports from the OMIM disease database for the respective genes.

All results are stored in plain text files and can easily be transferred to other applications. Since the graphical output is written in HTML format with PNG images, it can be examined in any web browser without the need of installing our or any other additional software.

Results and discussion

The program flow resembles the steps in the manual approach. In the first step, FragIdent collects the sequences and blasts them against the target database(s) to identify the ORF represented by the inserts. In the next steps, the alignment of multiple sequences of one clone, the recognition and clipping of the vector sequences, the design of new sequencing primers and the identification of protein domains are carried out. In the final step, all information is integrated in concise output files and is combined with the gene specific information available in public databases. The user is guided through the steps of insert identification and characterisation by a very simple graphical user interface (figure 1).



clones without gene assigned:
rf16a2
clones not completed:
nummer18, nummerh22, rf37d8, sqact41c12
genes hit by more than one clone:
TMOD3 nummerh51,rf25e4
ASNA1 rf41d2,rf41d7
MYST2 nummer18,rf14b8
CRMP1 nummerh16,orf16e9
ZNF746 rf37d7,rf37d8
sequences hit by more than one clone:
46358365 rf37d7,rf37d8

clone	title	seq	BLAST vector	BLAST insert	protein domains	gene	status
20n6					+	ZNF579	end of coding sequence reached
	20n6_20n6_gad	+	+	-			
	20n6_4928_e10_20n06_primerff_036	+	+	-			
	20n6_4928_e10_20n06_primerff_036_2	+	+	-			
	20n6_4929_e10_20n06_primerrr_036	+	+	-			
	20n6_4943_e10_20n06_primerff_036	+	+	-			
et158					+	APBA2	GAL4/vector detected target detected in reverse seq
	et158_4928_h09_et_158_primerff_033	+	+	-			
	et158_4928_h09_et_158_primerff_033_2	+	+	-			
	et158_4929_h09_et_158_primerrr_033	+	+	-			
	et158_4943_h09_et_158_primerff_033	+	+	-			
	et158_7656_ma_a01_pact4_1b_et_158_primerf_008	+	+	-			
	et158_7656_ma_a07_pact4_1b_et_158_primerr_032	+	+	-			
	et158_et_158_gad	+	+	-			
et192					+	CHRNB2	GAL4/vector detected reverse primer sequence detected target detected in reverse seq
	et192_4928_a10_pact4_1b_et_192_primery2h_gal4_ff_040	+	+	-			
	et192_4928_a10_pact4_1b_et_192_primery2h_gal4_ff_040_2	+	+	-			
	et192_4929_a10_pact4_1b_et_192_primery2h_gal4_rr_040	+	+	-			
	et192_4943_a10_pact4_1b_et_192_primery2h_gal4_ff_040	+	+	-			
	et192_et_192_gad	+	+	-			
et293					+	BPTF	GAL4/vector detected reverse primer sequence detected target detected in reverse seq
	et293_4928_c01_et_293_primery2h_gal4_ff_006	+	+	-			
	et293_4928_c01_et_293_primery2h_gal4_ff_006_2	+	+	-			

Figure 1

User interface. This figure shows the user interface of our software. On the left side, buttons for the possible actions are located. On top of the main frame, possible problems that might require a closer examination (e.g. clones to which no gene could be assigned) are summarised. Below, all available clones with all sequences, BLAST results etc. are listed together with the genes assigned and the current sequencing status.

Sequence alignment and identification

The software reads sequences in FASTA format and assigns them to the respective clones. The sequences are blasted against the NCBI's databases (with BLASTN) to find the corresponding sequence represented by the insert as well as the vector-specific sequences, thereby assigning the start or end of the insert. Alternatively, a local installation of BLAST can be used.

The search for target ORFs encoded by the cDNA-fragment is performed directly against a (human) mRNA database (or another database as specified by the researcher) with stringent settings (i.e., a high degree of identity is required). For the identification of vector sequence, all available DNA databases are queried with less stringent settings. The default settings of the software (i.e. minimal identity and E values of alignments) and the sequence databases can be easily changed by the users in a configuration file if they desire to query different organisms or focus on genomic clones. Furthermore, it is possible to upload BLAST results obtained by other means or to align single sequences with different parameters. The software can also handle sequences complimentary to the reverse strand and will indicate when the reverse strand was sequenced. BLAST was given preference over BLAT [18], because it allows the explicit use of mRNA databases.

Out of all BLAST hits, the (mRNA) sequence with the best coverage is chosen as 'target sequence'. The software also allows the manual definition of a target gene of a clone. In this case, only BLAST hits against this gene are considered.

FragIdent lists all clones to which no homologous hit can be found in public databases and notifies the researchers of genes covered by more than one clone. If the assigned sequence is not completely covered by the experimentally derived sequences, gaps are indicated graphically and listed.

Our software can use an unlimited number of sequences per clone. Sequences can be added at any time during the analysis. Re-analysis will be restricted to the data that has actually changed.

Primer design

Additional primers based on the sequence of the cDNA-fragment can be designed automatically using Primer3 with the default settings. Primer design starts around 100 bases before the end of the experimentally confirmed sequence. New primers are created in regular intervals (with a user-specified length) and listed in a file together with their physical positions within the target sequence and the corresponding region (coding sequence or 3' UTR).

Finding protein domains

After the alignment, the software uses the GI number of the best hit to query Genbank [16] for gene specific information including protein domains. The position of the domains within the protein is then mapped to the coding sequence. The domains are shown in the graphical output and included in the results file. Furthermore, the user is notified whether or not a domain is completely covered by the insert.

Graphical output and documentation

The alignments and the contained protein domains are displayed graphically in a printable HTML file, together with target gene specific information such as gene symbol and description (figure 2). Links provide fast access to the underlying data (sequences and BLAST results) for each clone. In a further step, information for the (human) genes represented by the inserts is retrieved from public databases and stored as an HTML file. In addition to the identification of the inserts, the software facilitates the documentation of the results, as the information is stored with a clear structure, in plain text files with a standardised format. Within the application only plain text, PNG images and HTML are used and all internal links are relative, hence a project can be shared by simply copying the files onto a web site. It is also possible to resume work on another computer or to join data from different projects by simply copying the files.

Conclusion

We have presented a data analysis suite to identify cDNA clones that were completely or partly sequenced, to find protein domains covered by the insert and to check the degree of coverage of the insert by experimentally derived sequences. The huge efforts on insert identification and eventually characterisation posed by large-scale studies using cDNA clones make manual approaches cumbersome and error-prone. We have designed our software to overcome the obstacles in an intuitive, user-friendly manner. This automatic approach is aimed at researchers who are not familiar with programming languages and want to analyse their data themselves. The graphical output and the possibility to interact with the analysis at various stages, give the users a high level of control over the identification process. We used this software successfully to analyse 84 cDNA clones from a yeast two-hybrid experiment. All cDNA fragments encoded a human gene. While most of the inserts could be completely sequenced with a single vector-specific primer in the first run, up to 6 further insert-specific primers were needed for longer cDNA fragments. In total, 131 protein domains were mapped onto the cDNA fragments.

Availability and requirements

Project name: cDNA-Alignment

Project home page: http://compbio.charite.de/genetik/FragIdent/

Operating systems: all

Programming languages: Perl, Perl/Tk

Other requirements: Primer3, Internet connection

License: free

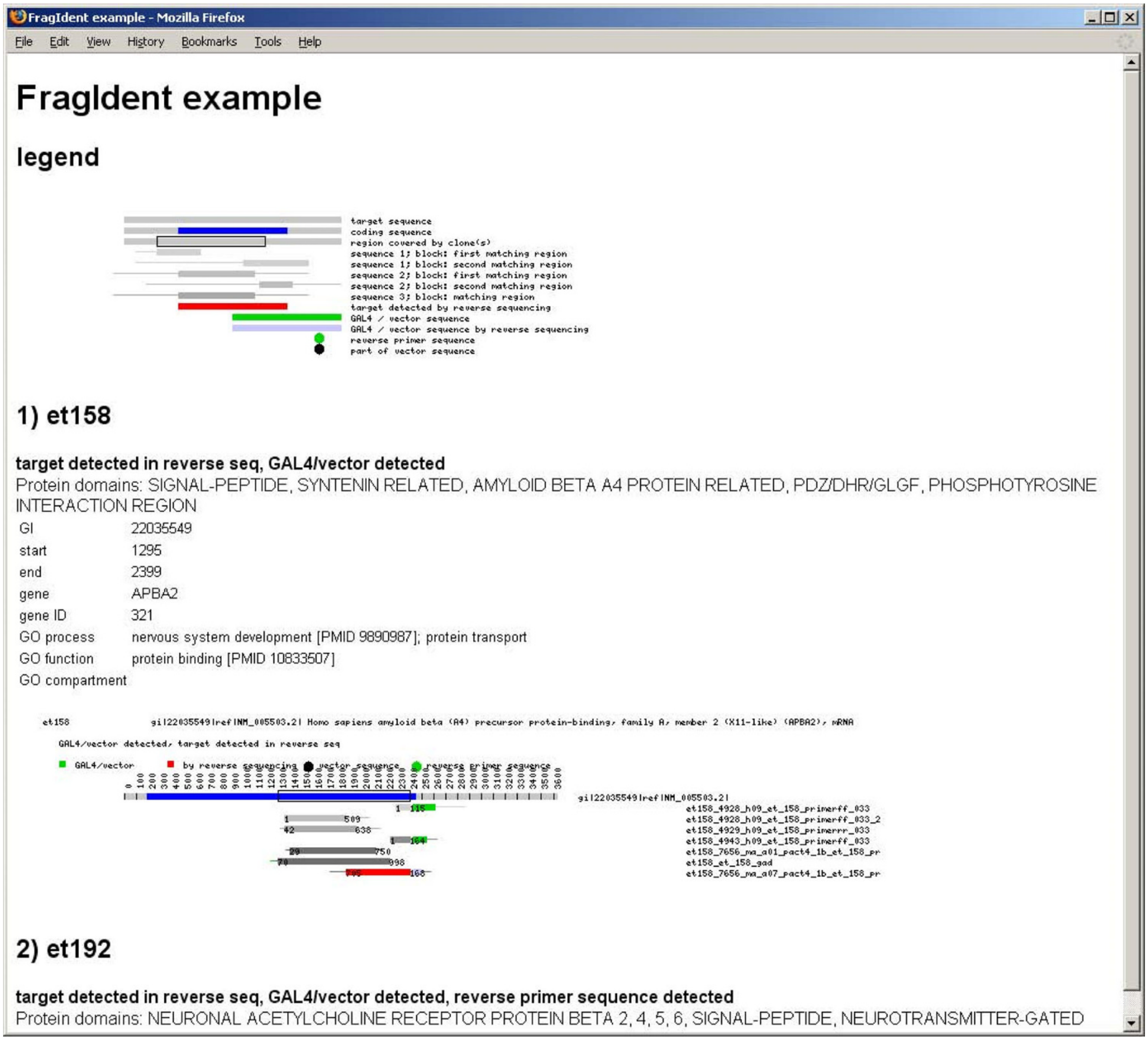
Any restrictions to use by non-academics: None

Authors' contributions

DS developed the software, HG provided the clones, and KH coordinated the underlying project and defined the software requirements. DS, HG and KH wrote the manuscript. KH is supported by Deutsche Forschungsgemeinschaft grant DFG, SFB 577, project A4, and is a recipient of a Rahel Hirsch Fellowship, provided by the Charité Medical Faculty.

Acknowledgements

We thank Mohsen Karbasiyan who helped us with the sequencing and Erich Wanker for his support within the Z2-project of the SFB 577.


FragIdent example - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help

FragIdent example

legend

target sequence
coding sequence
region covered by clone(s)
sequence 1; block: first matching region
sequence 1; block: second matching region
sequence 2; block: first matching region
sequence 2; block: second matching region
sequence 3; block: matching region
target detected by reverse sequencing
GAL4 / vector sequence
GAL4 / vector sequence by reverse sequencing
reverse primer sequence
part of vector sequence

1) et158

target detected in reverse seq, GAL4/vector detected

Protein domains: SIGNAL-PEPTIDE, SYNTENIN RELATED, AMYLOID BETA A4 PROTEIN RELATED, PDZ/DHR/GLGF, PHOSPHOTYROSINE INTERACTION REGION

GI	22035549
start	1295
end	2399
gene	APBA2
gene ID	321
GO process	nervous system development [PMID 9890987]; protein transport
GO function	protein binding [PMID 10833507]
GO compartment	

et158
gi|22035549|ref|NM_005503.2| Homo sapiens amyloid beta (A4) precursor protein-binding, family A, member 2 (X11-like) (APBA2), mRNA
GAL4/vector detected, target detected in reverse seq
GAL4/vector
by reverse sequencing
vector sequence
reverse primer sequence
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600
gi|22035549|ref|NM_005503.2|
1 115
1 509
42 638
1 164
29 750
70 998
705 168
et158_4928_h09_et_158_primerff_033
et158_4928_h09_et_158_primerff_033_2
et158_4929_h09_et_158_primerrr_033
et158_4943_h09_et_158_primerff_033
et158_7656_ma_a01_pact4_1b_et_158_pr
et158_et_158_gad
et158_7656_ma_a07_pact4_1b_et_158_pr

2) et192

target detected in reverse seq, GAL4/vector detected, reverse primer sequence detected

Protein domains: NEURONAL ACETYLCHOLINE RECEPTOR PROTEIN BETA 2, 4, 5, 6, SIGNAL-PEPTIDE, NEUROTRANSMITTER-GATED

Figure 2

Alignments – graphical overview. This screenshot shows the HTML page containing the graphical alignments. The page starts with a short legend followed by the different clones that were studied. This example shows the first clone ('et158'). The output lists the protein domains included in the insert and some gene specific information (more can be found in a separate list). The alignments are displayed in an image below: The target sequence is covered by three forward and one reverse sequences. Three further sequences could not be aligned to the target. The target gene is shown on top of this image as a grey bar with its coding sequence in blue. Below, the different sequences are plotted in different shades of grey with green parts representing the vector sequence. The sequence at the bottom is drawn in red to indicate that its orientation is reversed, i.e. that it was detected by sequencing from the 3' end and therefore determines the 3' end of the insert. The part of the gene that is actually covered by the sequences is emphasised with a box around it. Regions of the target gene not covered by sequences are marked with a thin red line on top of its bar and listed in the text. Protein domains are shown as horizontal purple bars reflecting their position in the cDNA. Below the figure, links to the raw data (sequences and BLAST results) are included.

References

1. Legrain P and Selig L: **Genome-wide protein interaction maps using two-hybrid systems.** *FEBS Lett* 2000, **480:**32–36.
2. Lennon GG and Lehrach H: **Hybridization analyses of arrayed cDNA libraries.** *Trends Genet* 1991, **7:**314–317.
3. Halgren RG, Fielden MR, Fong CJ and Zacharewski TR: **Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones.** *Nucleic Acids Res* 2001, **29:**582–588.
4. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T and Irie R, *et al*: **Complete sequencing and characterization of 21,243 full-length human cDNAs.** *Nat Genet* 2004, **36:**40–45.
5. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH and Goehler H, *et al*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122:**957–968.
6. VanBuren V, Piao Y, Dudekula DB, Qian Y, Carter MG and Martin PR, *et al*: **Assembly, verification, and initial annotation of the NIA mouse 7.4K cDNA clone set.** *Genome Res* 2002, **12:**1999–2003.
7. Porcel BM, Delfour O, Castelli V, De BV, Friedlander L and Cruaud C, *et al*: **Numerous novel annotations of the human genome sequence supported by a 5'-end-enriched cDNA collection.** *Genome Res* 2004, **14:**463–471.
8. Rice P, Longden I and Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276–277.
9. Almeida LG, Paixao R, Souza RC, Costa GC, Almeida DF and Vasconcelos AT: **A new set of bioinformatics tools for genome projects.** *Genet Mol Res* 2004, **3:**26–52.
10. Koski LB, Gray MW, Lang BF and Burger G: **AutoFACT: an automatic functional annotation and classification tool.** *BMC Bioinformatics* 2005, **6:**151.
11. Strahm Y, Powell D and Lefevre C: **EST-PAC a web package for EST annotation and protein sequence prediction.** *Source Code Biol Med* 2006, **1:**2.
12. Waegele B, Schmidt T, Mewes HW and Ruepp A: **OREST: the online resource for EST analysis.** *Nucleic Acids Res* 2008, **36:** W140–W144.
13. Smith RP, Buchser WJ, Lemmon MB, Pardinas JR, Bixby JL and Lemmon VP: **EST Express: PHP/MySQL based automated annotation of ESTs from expression libraries.** *BMC Bioinformatics* 2008, **9:**186.
14. Rozen S and Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132:**365–386.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z and Miller W, *et al*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389–3402.
16. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K and Chetvernin V, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008, **36:** D13–D21.
17. Seelow D, Schwarz JM and Schuelke M: **GeneDistiller – distilling candidate genes from linkage intervals.** *PLoS ONE* 2008, **3:** e3874.
18. Kent WJ: **BLAT – the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656–664.