In silico prediction of UTR repeats using clustered EST data

Stefan A. Rensing*, Daniel Lang and Ralf Reski

University of Freiburg, Plant Biotechnology, Sonnenstr. 5, D-79104 Freiburg, Germany

* stefan.rensing@biologie.uni-freiburg.de, fon +49 761 203-6974, fax -6990

Citation: Rensing S.A., Lang D. and Reski R. (2003): *In silico* prediction of UTR repeats using clustered EST data. In: Proceedings of the German Conference on Bioinformatics 2003, Mewes H.-W., Heun V., Frishman D., Kramer S. (eds.), pp 117-122, Belleville Verlag Michael Farin, Munich, Germany

Abstract

Clustering of EST data is a method for the non-redundant representation of an organisms transcriptome. During clustering of large amounts of EST data, usually some large clusters (>500 sequences) are created. Those can lead to iterative contig builds, consumation of lots of computing time and improbable exon alignments, which is unfavourable. In addition, these clusters sometimes contain transcripts for more than one gene, which is not desired. Such large clusters come into existence due to: (1) large numbers of identical ESTs / high transcript levels; (2) large gene families with highly similar members; (3) false clustering due to a) unremoved vector or rRNA sequences, b) undetected cloning artifacts or c) repetitive elements in UTRs.
During pre-processing (filtering and masking) of the sequence raw data, contaminations such as vector or linker sequences as well as bacterial genes are being removed (clipping). In the same process, it is essential to mask repetitive elements in order to avoid wrong clustering due to these sequence stretches. Therefore, determination of UTR repeats (to use in masking) is a method to avoid false clustering.
When dealing with organisms where repetitive elements are unknown, it is crucial to extract those sequences from the data prior to clustering. Here we present three *in silico* approaches to detect UTR repeats using clustered EST data. All three approaches yielded several putative repeats, of which the majority could be proven to be of repetitive nature in the genome. Usage of the predicted repeats enabled us to save computing time while increasing the quality of the clustered data.

Abbreviations: CDS = CoDing Sequence; EST = Expressed Sequence Tag; FCS = False Clustering Stretch; PCR = Polymerase Chain Reaction; UTR = UnTranslated Region

Definitions:
A contig is a non-redundant representation of a multiple sequence alignment by means of a consensus sequence.
A cluster is a set of sequences that share pairwise homologous stretches; during assembly contigs are built out of clusters.
A singlet is a sequence that either had no pairwise homology partner within the total pool of sequences or within a given cluster (= clustered singlet).

Introduction

Large scale EST projects usually aim at the non-redundant representation of an organisms transcriptome. Due to the nature of EST data, which represents a shotgun approach and is prone to sequencing errors as well as cloning artifacts, correct clustering tends to be difficult. In order to represent the transcriptome, great care is necessary to establish species- and dataset-specific parameters.
Our test datasets contain a large amount of EST data from the moss, *Physcomitrella patens*. For this organism, no information about repetitive elements is known, a situation that most groups face

who do not work with well characterised model organisms. When using such a dataset, it is imperative to figure out repetitive sequence stretches to use in the masking prior to clustering. Otherwise, clusters come into existence which contain transcripts of unrelated genes. This is due to the fact that during pairwise alignment the similar sequence stretches of repetitive elements lead to false clustering.
Using EST data that had been clustered without the aid of known repetitive elements, we developed three approaches to predict repetitive elements in the untranslated regions of the transcript, i.e. UTR repeats. Afterwards we applied these sequence stretches during filtering of the sequence raw data to avoid false clustering. The putative repeats have been analysed both by sequence analysis and in the wet lab.

Methods

Datasets

Because part of our data is proprietary, we carried out the analyses using two different data sets. The public data set consists of all the publicly available ESTs, totalling around 70,000 sequences. The complete data set additionally contains about 110,000 proprietary sequences (Rensing et al. 2002a). The public ESTs mainly comprise the 5' ends of the transcripts, whereas the proprietary data mainly comprises the 3' ends. The cDNA libraries used for the EST sequencing represent the whole life cycle of the organism. In addition, due to normalization and subtraction procedures, the collection has a low level of redundancy and is thought to cover the transcriptome nearly completely (Rensing et al. 2002a+b, Nishiyama et al., 2003).

Clustering

The filtering / masking of the raw data as well as clustering and assembly were carried out using the Paracel Transcript Assembler (PTA, www.paracel.com) with a parameter set optimized for the dataset / organism in question. PTA uses HASTE = Hash Accelerated Search Tool, a Smith-Waterman (Smith and Waterman, 1981) adaptation, for the pairwise comparison during clustering. The clustering and assembly is divided into two steps. In the first step, the seed-clustering, all known Physcomitrella CDS are being used to pull homologues from the input data set. These are then clustered and assembled independently from the rest of the sequences in order to save computing time.

BLAST searches

Standalone BLAST 2 (www.ncbi.nlm.nih.gov, Altschul et al., 1997) as well as GCG BLAST (see below) and the parallelized Paracel BLAST (www.paracel.com) have been utilized to carry out homology searches. The so-called "HASTE-BLAST" searches (simulating the HASTE algorithm) were performed using the BLAST 2 parameters NOGAP, NOFILTER, MATCH=3, MISMATCH=-6, WORDSIZE=12 and EXPECT=$1x10^{-3}$. For other BLAST searches, an E-value cutoff of $1x10^{-4}$ for peptide alignments and $1x10^{-2}$ for BLASTN have been used. When using the predicted repeats as query, the parameters NOGAP and NOFILTER have been used as well.

Additional software and databases

The GCG suite (10.3 UNIX, www.accelrys.com) and REPuter 3.0 (www.genomes.de; Kurtz et al., 2001) have been used. Textual clustering / filtering was to some extent carried out using Microsoft Excel. The following database releases were used: GENPEPT 133.0 (www.ncbi.nlm.nih.gov, and the plant subset termed PLANTPEP), a NCBI plant EST subset lacking *Physcomitrella patens* sequences (state of 22.8.03; containing 3,286,421 sequences), UTR DB release 14 (January, 2002; Pesole et al., 1996), Repbase Update Volume 8 (Issue 7; August 13, 2003, www.girinst.org, Jurka, 2000), SWISSPROT release 41.19 (4.8.2003, www.expasy.ch/swissprot). Several perl scripts have been written for the purpose of automating BLAST searches, parsing/filtering of BLAST output and pipelining of process input/output as well as parsing of the Paracel CAML (XML) files.

We developed and tested the following three approaches for the *in silico* detection of putative repetitive elements in untranslated regions of protein encoding genes (UTR repeats).

The REPuter approach (I)
The REPuter (Kurtz et al., 2001) approach looks for direct repeats in singlets and afterwards checks for the occurrence of those within contigs. In order to do this, all the singlets that are left after clustering and assembly, i.e. that are not part of the assembly, are concatenated. Using REPuter, they are then scanned for forward repeats of length >= 100bp with an allowed error rate of 3%. The resulting hits are afterwards being used as queries for BLAST searches against PLANTPEP. Those queries that yield significant hits (and therefore are not within the UTR but within the CDS) are discarded. The remainder is now being used as query for a BLAST search against the assembled sequence database, this time discarding all queries that produce hits against (non-clustered) singlets only. The remaining sequences are putative UTR-repeats and can be used for masking.

The HASTE BLAST approach (II)
The HASTE BLAST approach uses BLAST with the same parameters than the HASTE algorithm - which is initially used for clustering - to determine regions of potentially erroneous clustering. For this approach, two sequence databases are created: one from the assembled clusters, termed *ass*, the other from the input sequences before assembly, yet after filtering, termed *raw*. For these databases only the sequences from the largest clusters are being used, because those most probably contain the problem sequences. The sequences from *ass* are now being used as query in a HASTE BLAST search against *raw*. This search yields as hits the false-positives that led to false clustering within the large pool of sequence stretches that led to correct clustering. The BLAST output is now being filtered according to bit score (110<s<300, corresponding to the clustering threshold used within PTA) and position (the match has to be in the first 40% of the subject length, in addition, this is applied just to those ESTs that are known to represent a 3'->5' sequence run). The output sequence stretches from this procedure are subjected to clustering and the respective longest sequence is written to file to avoid redundancy. We now have a much smaller pool of sequences with most of the stretches that yield correct clustering removed. Now a second HASTE-BLAST round uses these sequence stretches as query against *ass*. The output is again filtered, removing self-hits as well as hits that expand beyond the bordering 30% of the subject sequence length. At least 3 hits per query must remain to count it as significant; those sequences are now being used as putative UTR repeats.

The FCS approach (III)
This approach tries to find those repeats that were missed in the first two approaches by determining „false clustering stretches" from contigs that do not match the majority homology annotation of the cluster ("textual minority clusters") by utilizing BLAST and filtering. Intitially, an *ass* database, like described above, is BLASTed against PLANTPEP with a list size of 1, extracting the description lines for further processing. The hits are then filtered according to this crude textual annotation. This yields a majority annotation, that most of the sequences out of a cluster will share. However, we also find "textual minority clusters" (TMC), i.e. sequences that do not share the majority annotation. Pairwise comparison of the differentially annotated contigs and clustered singlets showed that they often do not have common subsequences, because those sequence stretches that initially led to wrong clustering ("false clustering stretches"; FCS) are part of the raw sequences, but not anymore of the contigs.
As potential FCS, such sequences are being extracted from the TMCs that are longer than 50bp and did not make it into the assembled contigs. The potential FCS are afterwards used as query in a HASTE-BLAST search against *raw*; then all those hits are being discarded that are part of the contig the potential FCS is derived from. The remaining BLAST hit stretches will now be uses to query *ass*.
Now we discard those hits that are part of the TMC the potential FCS was derived from. If there are other hits, those are checked out in terms of whether they are located in an UTR (by using BLAST against PLANTPEP and discarding queries that exclusively hit CDS). We then use all sequences that are derived from the same potential FCS for a multiple sequence alignment and produce a consensus sequence, which is a putative UTR repeat.

Southern analyses

In order to determine the respective number of genomic representations of the putative repeats, PCR primers were designed and tested on genomic DNA. Products of the expected length were non-radioactively labelled in PCR reactions using Digoxigenin (Roche, Germany) and used as probes for Southern hybridisation. Several restriction enzymes with recognition sites of 6 bases in length have been tested on genomic wildtype DNA and by Southern blotting. It was determined that *Hin*DIII is most suitable to produce an even restriction pattern of the genomic moss DNA without large molecular fragments remaining. Therefore, *Hin*DIII digested genomic wildtype DNA was used to prepare the blots on positvely charged nylon membranes. All probe sequences were checked not to contain *Hin*DIII restriction sites. The number of clearly distinguishable hybridisation bands after luminescence detection was counted as the number of genomic loci.

Results and Discussion

Predicted repeats

All three approaches together yielded 17 putative UTR repeats, of which 6 were derived from the REPuter approach, 5 from the HASTE BLAST and 6 from the FCS approach (table 1). The predicted repetitive regions were of 53 to 298 base pairs length (mean: 163). The sequences of the repetitive elements are available via www.plant-biotech.net.

All predicted repeats have been checked in the wet lab for presence in the genome (data not shown) and could be detected with a copy number between 2 and 17, proofing their repetitive nature. The REPuter repeats R9 and R10 were shown to be present in the genome at only two locations. All other predicted repeats showed at least 4 and maximum 17 loci (mean: 8.9). No significant correlation (r=0.2) could be determined between the number of hits during masking and the number of Southern bands, i.e. genomic loci.

table 1

repeat				BLAST				large dataset		public data	
id	type of repeat	length of repeat	number of genomic loci	hits vs. EST plant	hits vs. SwissProt	hits vs. UTR DB plant	hits vs. Repbase plant	HASTE BLAST hits vs. raw	masked areas	HASTE BLAST hits vs. raw	masked areas
R2	HASTE-BLAST	80	5	0	-	0	0	42	156	5	37
R3	HASTE-BLAST	160	17	1161	CAB	0	0	434	2499	97	513
R4	HASTE-BLAST	155	13	434	CAB (1)	0	0	130	395	33	243
R5	HASTE-BLAST	169	8	1123	R-LSU	0	0	510	974	30	172
R6	HASTE-BLAST	88	6	1080	RA	0	0	296	432	40	114
R7	REPuter	194	6	236	-	60	1	57	1933	51	1388
R8	REPuter	145	10	18	-	8	0	38	910	2	117
R9	REPuter	171	2	0	-	3	0	4	960	0	112
R10	REPuter	257	2	5	-	7	8	8	1664	0	216
R11	REPuter	200	5	5	-	5	5	4	1297	0	236
R12	REPuter	152	11	131	-	21	21	26	1531	22	204
R13	FCS	53	15	0	-	0	0	22	608	0	28
R14	FCS	71	6	0	-	0	0	6	353	0	10
R15	FCS	117	4	0	-	0	0	10	144	4	79
R16	FCS	298	15	21	-	12	0	15	707	444	278
R17	FCS	242	15	1098	CAB (1)	12	0	251	1559	184	797
R18	FCS	215	12	0	-	0	0	8	549	4	295
	mean:	163	8,9	312		8	2	109	981	54	285
	total:			5312		128	35	1861	16671	916	4839

(1) repeat contains part of the C-terminal CDS and the 3' UTR

Sequence analysis

The repetitive elements were used as query in BLAST searches against SWISSPROT and the UTR DB plant subset (table 1). Whereas 12 of 17 repeats did not find a match in SWISSPROT, 5 sequences matched against chlorophyll a/b binding protein (CAB), RUBISCO large subunit (R-LSU) and RUBISCO activator (RA). Two of the three CAB matches (R4 and R17) match against the C-terminal end of the protein and extend into the 3'-UTR. The other three hits (R3, R5 and R6) lie within the CDS. Although they therefore do not represent UTR repeats, they seem to represent sequence stretches that lead to clustering of paralogues of these multigene families if not used for masking. We will hence call these three repeats CDS-repeats. Because of the fragmentary nature

of EST data, the HASTE-BLAST approach (by using hits at the EST edges as potential UTRs) seems to predict non-UTR repeats with a higher chance than the other two approaches.
Eight of 17 repeats found matching regions in UTR DB. This shows that these sequences are present with only slight variation in UTR regions of other species as well (the UTR DB plant subset used here contains only 42 of 25,531 = 0.16% *Physcomitrella patens* sequences, yet none of those were hit by our 17 repetitive sequences). The remaining 9 repeats (6, if we do not count the CDS-repeats R3, R5 and R6) do not have close homologues in UTR DB, which makes them candidates for being species-specific repeats.
Only 4 of the 17 repeats yielded hits in the Repbase plant subset, proving that most of the repetitive elements are novel. No hits from the HASTE-BLAST and FCS predictions were found against Repbase, so these two approaches seem to be especially suited to discover new and species-specific repeats.
In the comparison with the plant EST subset (which does not contain Physcomitrella sequences), a total of 5312 hits was found. Not surprisingly, most of those are due to the three CDS-repeats and the two repeats that cover the C-terminal end of the CDS. Besides those, another FCS repeat as well as 5 of the 6 REPuter repeats find matches in the plant EST collection. This demonstrates again, that some of our sequence stretches are well conserved across species while others seem to be species-specific.
The repeats R2, R13, R14, R15 and R18 did neither find similar sequences in the BLAST searches against plant ESTs nor against the UTR DB and Repbase plant subsets. Therefore these 5 repeats seem to be novel and specific for *Physcomitrella patens*. A further 8 repeats are novel in the sense that they are not present in Repbase.

Masked areas
When including the 17 repeats into the pre-processing (filtering/masking) of the EST data, a lot of repetitive regions could be masked (table 1). Using the 17 putative repeats on the large dataset, 16,671 regions in the input sequences were detected (corresponding to 9.7% of the input sequences). The number of masked areas per repetitive element was in the range of 156 to 2499 (mean: 981). For the public data, 4,839 regions were masked, corresponding to 7.1% of the input sequences. Here, the number of masked areas per repetitive element was in the range of 10 to 1388 (mean: 285). On average, a HASTE-BLAST repeat masked 4.9% of the input sequences, a REPuter repeat 8.1% and an FCS repeat 4.5%.
By performing HASTE-BLAST searches against the filtered raw data, the number of repetitive elements detectable by the HASTE algorithm during clustering was checked for. This yielded between 4 and 510 hits (mean: 109) for the large dataset and between 0 and 444 hits (mean: 54) for the public dataset.
The public dataset contains approximately 40% of the sequences present in the large dataset. The number of masked areas in the public set is 29%, the number of HASTE-BLAST hits 49% of those in the large dataset.
The inclusion of the putative repeats into pre-processing of the EST data removed 7.3% (3.8% for the public set) of input sequences from the seed clustering pool in exchange for including them as input sequences in the normal clustering as well as some problem sequences and singlets. The two largest clusters (majority annotation: chlorophyll a/b binding protein) could thus be reduced in size by roughly 25%. The assembly of large clusters is a time-consuming step, especially so if an iterative assembly is necessary. Therefore, computing when including the repeats for masking was around twofold quicker as compared to the clustering lacking this information.

Conclusions
Three approaches for the *in silico* prediction of UTR repeats have been used on two test datasets, resulting in the detection of sequence stretches in ~8% of the input sequences during filtering/masking and reduction in size of large clusters. Overall, this resulted in saving ~50% of computing time. 15 of 17 repeats have been proven to be repetitive in the genome by Southern blot analysis (4 to 17 representations in the genome). Thus, our methods are able to detect novel and species-specific UTR repeats using clustered EST data, which in turn can be utilized to enhance the results and increase the speed of the clustering process.

Acknowledgements
We would like to thank Dana Fritzowsky for skillfull technical assistance. The proprietary EST data has been produced in a joined project with BASF Plant Science GmbH.

References

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucl Acids Res*, **25** ,3389-3402

Jurka,J. (2000) Repbase Update: a database and an electronic journal of repetitive elements, *Trends Genet*, **16** ,418-420

Kurtz,S., Choudhuri,J., Ohlebusch,E., Schleiermacher,C., Stoyem,J. and Giegerich,R. (2001) REPuter: The manifold applications of repeat analysis on a genomic scale, *Nucl Acids Res*, **29** ,4633-4642

Nishiyama,T., Fujita,T., Shin-I,T., Seki,M., Nishide,H., Uchiyama,I., Kamiya,I., Carninci,P., Hayashizaki,Y., Shinozaki,K., Kohara,Y., and Hasebe,M. (2003) Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution, *Proc Natl Acad Sci USA*, **100** ,8007-8012

Pesole,G., Grillo,G. and Liuni,S. (1996) Databases of mRNA Untranslated Regions for Metazoa, *Computer Chem*, **20** ,141-144

Rensing,S.A., Rombauts,S., Hohe,A., Lang,D., Duwenig,E., Rouzé,P., Van de Peer,Y. and Reski,R. (2002a) The transcriptome of the moss *Physcomitrella patens*: comparative analysis reveals a rich source of new genes.
http://www.plant-biotech.net/Rensing_et_al_transcriptome2002.pdf

Rensing,S.A., Rombauts,S., Van de Peer,Y. and Reski,R. (2002b): Moss transcriptome and beyond, *Trends Plant Sci*, **7** ,535-538

Smith,T. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147** ,195-197