BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 18 2009, pages 2302–2308
doi:10.1093/bioinformatics/btp410

Sequence analysis

k-link EST clustering: evaluating error introduced by chimeric sequences under different degrees of linkage

Lauren M. Bragg[1,2,*] and Glenn Stone[1]

[1]CSIRO Mathematical and Information Sciences, North Ryde, NSW 2113 and [2]Preventative Health National Research Flagship, Locked Bag 17, North Ryde, NSW 1670, Australia

Received on April 6, 2009; revised on June 4, 2009; accepted on June 26, 2009

Advance Access publication July 1, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: The clustering of expressed sequence tags (ESTs) is a crucial step in many sequence analysis studies that require a high level of redundancy. Chimeric sequences, while uncommon, can make achieving the optimal EST clustering a challenge. Single-linkage algorithms are particularly vulnerable to the effects of chimeras. To avoid chimera-facilitated erroneous merges, researchers using single-linkage algorithms are forced to use stringent sequence–similarity thresholds. Such thresholds reduce the sensitivity of the clustering algorithm.

Results: We introduce the concept of `k-link` clustering for EST data. We evaluate how clustering error rates vary over a range of linkage thresholds. Using `k-link`, we show that Type II error decreases in response to increasing the number of shared ESTs (ie. links) required. We observe a base level of Type II error likely caused by the presence of unmasked low-complexity or repetitive sequence. We find that Type I error increases gradually with increased linkage. To minimize the Type I error introduced by increased linkage requirements, we propose an extension to `k-link` which modifies the required number of links with respect to the size of clusters being compared.

Availability: The implementation of `k-link` is available under the terms of the GPL from http://www.bioinformatics.csiro.au/products.shtml. `k-link` is licensed under the GNU General Public License, and can be downloaded from http://www.bioinformatics.csiro.au/products.shtml. `k-link` is written in C++.

Contact: lauren.bragg@csiro.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

While sequencing may be becoming more economical, many species are unlikely to be fully sequenced in the near future. Genome size and complexity can often decide whether an organism will become a candidate for genome sequencing. Even when a genome has been sequenced, there is no guarantee that sequencing reads can be assembled successfully. High-repeat content and extensive gene families (due to rounds of genome duplication) can often confound sequence assemblers, resulting in fragmented or ambiguous assemblies. *Sorghum*, perhaps one of the genetically simpler grasses, with a haploid genome of 10 chromosomes, a genome size ≈730 Mb, 61% of which is repetitive content, has been sequenced with at least 8× redundancy, yet still has over 3000 supercontigs (constituting roughly 10% of the genomic sequence) that cannot be assigned to a chromosome (Patterson *et al*., 2009).

Expressed sequence tags (ESTs) have often provided an inexpensive, transcriptionally biased, genomic substitute for many types of analyses. Studies that require high levels of sequence redundancy, such as polymorphism discovery, often use EST data. A crucial step within most sequence analysis pipelines is the process of clustering.

Clustering acts as a coarse filter in these analysis pipelines, reducing the input size to the, often computationally expensive, alignment process. The algorithm used often depends on whether the user wants to construct clusters of ESTs representing the same transcript (*transcript assemblies*), or alternatively, grouping all transcriptional products of the same gene (*gene index)*.

Many clustering algorithms are designed to cluster ESTs given a set of reference sequences, be it a genome or a set of complete mRNA sequences [TGICL (Pertea *et al*., 2003); Unigene (Boguski and Schuler, 1995); ECgene (Kim *et al*., 2005); to name a few). Many benefits can be gained from the use of genomic scaffolds: they reduce the number of pair wise sequence comparisons, they are often less error-prone than ESTs and they ensure that phylogenetically divergent transcripts with shared domains (such as paralogues, chimeras and repeat-containing transcripts) are not placed in the same cluster.

When species-specific reference sequences are unavailable, it is not uncommon to use the genome of a related species to guide the clustering process. However, use of a reference derived from a related organism should be approached with caution. To account for genomic distance, sequence identity thresholds may need to be dropped significantly. This can easily confound efforts to separate paralogous sequences. This will be particularly deleterious for those species which have experienced several rounds of genome duplication.

The presence of chimeric sequences in a EST dataset is considered a significant barrier to achieving accurate clusterings in the absence of a reference genome (Sorek and Safer, 2003). Chimerism is typically a PCR-induced artefact whereby two or more ESTs, typically from different genes, are artificially ligated. The detrimental influence of chimeric sequences may be somewhat restrained by increasing the align-length threshold (for example,

*To whom correspondence should be addressed.

requiring that 75% of the larger EST aligns to the smaller EST). However, such an approach would reduce the overall sensitivity of the clustering algorithm. Clustering algorithms which implement the single-linkage merging method will be particularly vulnerable to chimera-facilitated erroneous merges.

A previous study evaluated the effect of the sequence–similarity threshold on EST clustering error (Wang *et al*., 2004). Analogous to their statistical interpretation, errors in clustering can be designated as Type I or Type II errors. A Type I clustering error indicates that ESTs from the same gene have been placed in separate clusters. A Type II error describes the case where sequences from different genes have been clustered together. Wang *et al*. (2004) identified that Type I and Type II errors were minimized while using a sequence–similarity threshold of 25–40 bases aligned at 90% identity. Despite these findings, many clustering algorithms default to significantly more stringent parameters [CLOBB (Parkinson *et al*., 2002), with 30 bases at 95% identity, d2_cluster (Burke *et al*., 1999) with 100 bases at 90% identity, just to name a few]. Curiously, no one has investigated how the degree of linkage is used when clustering influences the magnitude and the type of clustering error. Here, we investigate how, when using this optimal sequence–similarity threshold, the required number of links (k) between clusters influences clustering error. To facilitate this investigation, we have developed a clustering algorithm, *k*-link, to produce a k-link EST clustering.

1.1 Existing clustering methods

EST clustering has been an extensively explored research area. Quite a number of algorithms and computational tools have been described in the literature. Perhaps the most commonly used reference clusterings are the 'gene indices' (where each cluster contains transcripts from the same gene) produced by the TGICL algorithm, developed by TIGR (Pertea *et al*., 2003).

TIGR gene index construction involves comparing EST and gene sequences using FLAST (Quackenbush *et al*., 2001). Sequences sharing a minimum identity of 95% over a $\geq$40 nt region, with each end of the alignment having less than 40 mismatches, are grouped into a cluster. This means one EST can belong to multiple clusters, and relies on anchor sequences (genomic or cDNA) to ensure that ESTs generated from different regions of the same gene are clustered together. Strictly speaking, TIGR does not need reference sequences to produce clusters; however, we expect the quality of the clustering will suffer as merges are not conducted between clearly overlapping EST clusters.

CAP3 (Huang and Madan, 1999), a 'DNA sequence assembly program', has been the mainstay (with over 700 citations) in sequence clustering and assembly since its inception. While it is intended as genomic sequence-assembly program, its versatility has lead to its use in numerous EST analysis pipelines [for example, autoSNP (Barker *et al*., 2003); QualitySNP (Tang *et al*., 2006)]. In our personal experience, CAP3 is a highly conservative algorithm, which while being very precise about which sequences are placed together, produced significantly large numbers ($\approx$80% of input dataset) of 'singlet' (or singleton) clusters when using 90% identity and 40 base overlap as parameters (Bragg, unpublished data). The CAP3 algorithm uses numerous heuristics which will not be summarized here.

There are several algorithms which have been specifically developed to cluster ESTs in the absence of genomic or full-length mRNA scaffolds. Most prominently, a succession of single-linkage EST clustering algorithms which use the d^2 distance metric (Hide *et al*., 1994) [see d2_cluster (Burke *et al*., 1999); CLU (Ptitsyn and Hide, 2005); and more recently, wcd (Hazelhurst *et al*., 2008)]. While these algorithms are fast and highly scalable, the single-linkage clustering method implemented therein is highly susceptible to merges between unrelated clusters due to chimeric sequences. While chimeras are uncommon [frequency of 0.01 (Hillier *et al*., 1996)], it is intuitive, especially in large datasets, that random chimeric sequences can cause an inordinate amount of damage to clustering results. In the absence of a genomic reference, the influence of chimeric sequences on the clustering outcome becomes a real problem. It is for this reason that many clustering algorithms are run with very stringent similarity thresholds. Unfortunately, the end result is often a trade-off of errors introduced by chimeric sequences for errors introduced by high sequence–similarity thresholds.

The study conducted by Wang *et al*. (2004) evaluated how the sequence-identity and align-length clustering parameters contributed to clustering error. Here, we investigate how the number of links required to merge clusters influences the error of the clustering.

2 METHODS

2.1 *k*-link clustering algorithm

The clustering algorithm begins with an all-against-all EST comparison. For every query EST, a cluster is created with this EST as the seed, and every EST that aligned at or above our similarity threshold is added to the cluster. This means for a set of n ESTs, there will be n initial clusters. This is seen as the initialization step of the clustering process, and is separate from the clustering algorithm itself. We have used megaBLAST for this comparison, and filtered the hits based on our earlier definition of a match. Non-default parameters ($N=0, E=0.1, W=12, T=21$) for megaBLAST were used as suggested in Korf *et al*. (2003). These initial clusters can be easily generated using other alignment methods, and for this reason, we have decoupled the cluster initialization process from the cluster comparison and merge steps. This flexibility enables users to decide which sequence alignment algorithm they wish to use to construct the initial cluster seeds.

The start of the iteration begins with the pair wise cluster comparison, where the number of ESTs shared by two clusters is calculated. If the number of shared ESTs is larger than the number of required links k, or the clusters are identical, this is recorded. Otherwise, if there are shared ESTs, but this is less than the number required, these ESTs are marked as potential chimeras. In the first iteration only, clusters seeded with an EST that is marked as chimeric will be removed. This is because a cluster seeded with a chimera (composed of genes A and B, say) will contain a mix of sequences from genes A to B. This chimera-seeded cluster, depending on the value of k, would be able to merge with both clusters containing A sequences, and clusters containing B sequences. To avoid this, we ensure that putative chimeras are removed prior to any merging operations. At the end of all the required cluster comparisons, the cluster pairs which had at least k links are now merged. This marks the end of the iteration.

As the parameter k must be estimated from cluster sizes, a rough clustering iteration is performed using a large number of links (k defaults to 6). The actual number of links required is then calculated (this calculation is described in Section 2.2). The rough clustering is then discarded. The clustering algorithm then begins its first iteration using the estimated k.

Consecutive iterations consist of the cluster comparison and merge steps. Clustering stops when no merges are recorded within an iteration, or when the maximum number of iterations has been reached.

2.2 Estimating the links parameter

The `k-link` algorithm outlined requires users to specify k, the number of links required. Clearly, to choose an appropriate linkage level, the chimeric frequency, dataset size and number of expected clusters (or 'gene indices' represented in the data) will need to be taken into account. Here, we describe the statistical model behind estimating the linkage parameter.

A chimeric sequence, or chimera, can suggest similarity between unrelated clusters of ESTs. As mentioned earlier, chimerism is a relatively rare event, and for this reason we have restricted our statistical model to consider only chimeras generated by a fusion of *two* unrelated ESTs.

In the single-link-based algorithms, non-exclusive clusters of ESTs will be merged when they have a common member. In practice, this will occur erroneously whenever there is a chimera that links two clusters. In the clustering algorithm that we propose, we will merge clusters based on k links. Thus, we need to consider the probability of obtaining multiple chimeras of the same chimeric type, that is, with one part of an EST derived from gene (cluster) A (say) and the second part derived from gene (cluster) B.

Clearly, if there are m genes in the organism of interest, there are $\binom{m}{2}$ different possible chimeric types. In order to control the erroneous merging of clusters, we need to understand the probability of observing a number of chimeras of the same chimeric type. When the number of links of interest is two, this is analogous to the so called 'Birthday Problem'.

The Birthday Problem considers a class of n individuals and then asks what is the probability that two (or more) individuals have the same birthday. In general, this probability can be written as

$$1-\frac{N!}{(N-n)!N^n}$$

where n is the number of individuals, and N is the number of possible birthdays (365). This formulation assumes for simplicity that the probability of an individual having a particular birthday is uniform across all possible birthdays.

For our example, we need to generalize this problem in two ways. First, we are interested in considering k or more common links between clusters. Second, it is unlikely that uniform probabilities of occurrence of chimeric types will even approximately hold. This then becomes a general problem in multinomial probabilities. Suppose we decide to merge all cluster with k or more links (ESTs in common). We will make an erroneous merge due to chimeras with probability q_k where

$$q_k = 1-\mathrm{P}(\max_i X_i < k)$$

and $(X_1,\ldots,X_N)$ is a random vector from a multinomial distribution with parameters n and $(p_1,\ldots,p_N)$. In this case, N is the number of possible chimeric types and is given by $\binom{m}{2}$, n is the total number of chimeras in the database of ESTs and p_i is the probability of observing a chimera of type i. In our example, N and n are likely very large and computation of q_k by enumeration is computationally tedious, however, Levin (1981) proposes an approximation approach based on Edgeworth expansions, and provides bounds for the true probabilities.

Thus, given m, n and the vector of probabilities p, we can determine the probability of observing k or more chimeras of the same type. This can be used to set the number of links in the clustering algorithm, that is, choose k such that the probability of observing these many chimeras of the same type is small, which corresponds to the probability of erroneously merging two clusters because of chimeras being small. To estimate that the parameter n is relatively straightforward, we simply assume that 1% of ESTs are chimeric. Possible estimates for p can be derived from the distribution of cluster sizes expected (if they are known, or taken from a species with a similar number of genes and expression profile) or by simply assuming uniform probabilities (analogous to the Birthday Problem).

2.3 Datasets

Three EST datasets were selected for evaluating clustering error rates. All were downloaded from NCBI using eUtils on November 10, 2008. (i) *Caenorhabditis elegans* ESTs (352 043 sequences), (ii) *Oryza sativa* ssp. Indica (173 887 sequences) and (iii) *Sorghum Bicolor* (202 294 sequences).

For the construction of the reference clusters, genomic sequences for these species were also downloaded [*C.elegans* genome build Ce6, *S.bicolor* genome (Sbi1) and *O.sativa* ssp. Indica (Gramene build Jan 2005, 48)]. The *Sorghum* genome assembly Sbi1 consists of 10 chromosomes and over 3000 super-contigs. For ease-of-analysis with BLAT (Kent, 2002), the 3000 super-contigs were excluded from our study.

2.4 Masking

Masking is a crucial process that must be applied, especially in EST datasets, to prevent common EST features (such as linkers, poly-A tails and repeat sequences) from negatively influencing the clustering process. Repeat databases have been constructed for many species, and are commonly used in the masking process.

For novel or unsequenced species, a species-specific repeat database may not exist. Recent studies have shown that use of repeats from a related organism have little to no positive effect on clustering outcomes (Malde and Jonassen, 2008). The use of library-less repeat-masking, such as that provided by `RBR` (Malde *et al*., 2006), can overcome the lack of a species-specific repeat library.

All ESTs used in this experiment had been processed with `RBR` prior to clustering. The default configuration of $s=1.5$ and $d=6$ was used for the *C.elegans* and *O.sativa* datasets. *Sorghum*, with its high repeat content, was subjected to the more aggressive masking thresholds of $s=1$ and $d=4$. For all datasets, the default word size of 16 was used.

For each dataset, 100 000 ESTs were randomly selected and supplied to `RBR` for the generation of the oligomer repeat library. All ESTs in the dataset were then masked using this library.

As some low-complexity sequences are not statistically overrepresented in these datasets, any remaining low-complexity sequences were masked using `mdust` (R.L.Tatusov and D.J.Lipman, unpublished data) at default settings.

A preliminary run of `k-link` ($k=10$) showed a large cluster forming in each dataset. Upon manual inspection, the clustering appeared due to inadequate masking of repeats/transposed elements. Increasing `RBR` ($s=1$ and $d=3$) and `mdust` (word size = 12) masking failed to mute these uninformative regions within the large clusters. At the risk of losing a large proportion of the dataset by trying to aggressively mask these large clusters, we decided to omit these ESTs altogether.

After masking, sequences with less than 40 consecutively unmasked bases were removed from further analysis. For the creation of reference clusters, 314 784 *C.elegans* ESTs, 160 537 *O.sativa* ESTs and 187 894 *S.bicolor* sequences were available.

2.5 Reference clusters

Reference clusters were constructed in a similar manner to those in Malde and Jonassen (2008). ESTs were matched to their genomic location using `BLAT` (step size = 6) (Kent, 2002). To maintain consistency throughout the study, we define a match as a 40 base window in the alignment of query and target sequence where there is at least 90% sequence identity (i.e. at least 36 bases are identical). It has been shown that a sequence identity threshold of 90% minimizes the Type I and Type II clustering errors introduced by the sequence alignment component of the clustering process (Wang *et al*., 2004).

If multiple locations matched, the best hit was taken. Unmatched ESTs were removed from further analysis.

A sliding window was moved along the genomic sequence to identify clusters. Overlapping ESTs were aggregated into the same cluster if their genomic coordinates overlapped by at least 40 bases. Clustering results are shown in Table 1. For all three datasets, roughly half of the ESTs could

Table 1. Results of clustering ESTs using a genomic reference

Species	#Clusters	#ESTs	Percentage alignable[a] to genome
Caenorhabditis elegans	23 002	151 923	48.26
Oryza sativa	24 525	71 113	44.30
Sorghum bicolor	20 361	96 506	51.36

[a]With align threshold being 90% identity over a window of 40 bases.

be aligned to the genome. We believe this is due to both the high level of sequence masking and BLATs relatively poor performance on short (and/or divergent) query sequences (http://genome.ucsc.edu/FAQ/FAQblat).

2.6 Clustering comparison

Common metrics for comparing the clusterings [Jaccard Index (Jaccard, 1901); Variation of Information metric (Meila, 2007)] assume that a clustering is a partitioning of a dataset into mutually disjoint sets. While this may be true of our reference clustering (where we took the 'best' hit for each EST), this one-to-one mapping is not guaranteed by the `k-link` algorithm (for $k>1$). This is because clusters with insufficient links will not be merged, thus the shared members will belong to both clusters. These sequences represent putative chimeras.

In the case of disjoint clusters, we can define

$R_{ij}=1$, whenever objects i and j are in the same cluster in the reference

$=0$, otherwise

$C_{ij}=1$, whenever objects i and j are in the same cluster in the comparison

$=0$, otherwise

Then the number of *true positives* is,

$$\sum_{\{i,j\}} R_{ij}C_{ij}$$

and the number of *true negatives* is,

$$\sum_{\{i,j\}} (1-R_{ij})(1-C_{ij})$$

where the sums are taken over all pairs of objects. For non-disjoint (comparison) clusterings, we need to adapt C_{ij}. Let A_i be the set of clusters containing object i. Then for disjoint clusters we can write,

$$C_{ij}=|A_i \cap A_j|$$

We propose for (possibly) non-disjoint clusters the measure,

$$C^{\star}_{ij}=\frac{|A_i \cap A_j|}{|A_i||A_j|}$$

which represents the probability that a cluster chosen randomly from the set A_i is the same as one chosen randomly from A_j. This has the effect of penalizing the amount of non-disjointness in the clustering.

The definitions of true negatives and positives then go through as before with $C^{\star}_{ij}$ replacing C_{ij}. The complete table of definitions is,

$$\text{True Positives} = \sum_{\{i,j\}} R_{ij}C^{\star}_{ij}$$

$$\text{False Positives} = \sum_{\{i,j\}} (1-R_{ij})C^{\star}_{ij}$$

$$\text{True Negatives} = \sum_{\{i,j\}} (1-R_{ij})(1-C^{\star}_{ij})$$

$$\text{False Negatives} = \sum_{\{i,j\}} R_{ij}(1-C^{\star}_{ij})$$

Table 2. Probability of seeing at least k occurrences of the same chimeric type

Species	Number of occurences	
	2	3
Caenorhabditis elegans	(0.3696, 0.4611)	(0.0049, 0.0049)
Oryza sativa	(0.0183, 0.0185)	(<0.0001, <0.0001)
Sorghum bicolor	(0.1888, 0.2092)	(0.0012, 0.0012)

Probabilities were calculated using a rough 6-link clustering.

For further reading, see Klastorin (1980). Using the definitions above, we proceed to define the type I error rate as the number of false negatives (FN) divided by the sum of true positives (TP) and false negatives. Correspondingly, the type II error rate is the number of false positives (FP) divided by the sum of the true positives and false positives.

To investigate how linkage degree effects clustering error, we ran `k-link` on the three datasets for $k \in \{1,2,3,4,5\}$. Error rates were calculated for the results of each run.

3 RESULTS

To validate whether use of a rough clustering ($k=6$) was a good enough approximation to the 'true' distribution of clusters, we estimated the probability bounds for co-occurrence using $k=2,3$. This bound estimation is provided in the `k-link` software. As a reference, we also estimated k using the reference cluster size distribution (see Table 1 of the supplementary figures). The probabilities estimated from the rough clusterings were similar in magnitude (but not identical) to the probabilities estimated from the reference cluster sizes (Table 2). Despite these differences, each rough clustering suggested the same number of required links ($k=3$, at the 0.01 significance level) as that estimated from the reference clustering for the dataset.

3.1 Evaluating clustering error

There appears to be a significant amount of Type II error in the datasets (Table 3). As expected, the Type I error FN/(TP+FN) is increasing as the number of required links is incremented. Conversely, the Type II error [FP/(TP+FP)] is decreasing as the number of required links is incremented.

The most significant result is observed in the *Sorghum* dataset, whereby increasing the required links from 2 to 3 decreased the Type II error from 30% to 20%. In contrast, the Type I error only increased from 0.02% to 1%. This disproportionately large decrease in Type II error is what one would expect from the removal of chimeric sequences joining two otherwise unrelated clusters. Examination of the cluster size statistics (Table 4) shows that of the three datasets, *Sorghum* had the largest decrease in maximum cluster size as required links were increased from 2 to 3 (1354 members down to 1086). This is consistent with the hypothesis that a chimeric sequence linked these otherwise unrelated clusters.

All three datasets have a small amount of false negatives under the single-linkage requirement. Ideally, this would be zero, however, it is likely caused by differences between the BLAT and BLAST algorithms.

Interestingly, the difference between clusterings produced by single-linkage and 2-linkage is minimal, if any. This may suggest

that the data represents at least 2-fold coverage of the gene indicies in these species—excluding the singletons representing truly rare transcripts. However, this conclusion is confounded by the overwhelming effect of the unmasked repetitive sequences that still remain within the dataset.

Manual inspection of a subset of the putative chimeras revealed the ESTs to be a mixture of chimeras, splice variants, low expression transcripts and long ESTs which bridged the gap between sense and antisense clusters (Fig. 1). Images were obtained from WormBase (http://www.wormbase.org/, version WS201), and the UCSC Genome Browser (Kent *et al*., 2002).

For consistency with previous sequence clustering studies (Burke *et al*., 1999), we have also compared the size of clusters produced by the reference versus the `k-link` clusterings (Table 4). Excluding *O.sativa*, the 3-linkage clustering produced cluster sizes similar to those in the reference clustering. The largest cluster size in all 3-link clusterings was slightly larger than that in the reference. Boxplots of the cluster sizes for each clustering are supplied in the Supplementary Materials.

Table 3. Evaluation of error introduced by linkage degree

Species	TP	FP	FN	FN/(TP + FN)	FP/(TP + FP)
Caenorhabditis elegans					
Single-link	6 293 463	1 873 985	7256	0.0012	0.2294
2-link	6 293 463	1 873 985	7256	0.0012	0.2294
3-link	6 220 829.56	1 608 038.61	79 889.44	0.0127	0.2054
4-link	6 171 122.04	1 340 286.73	129 596.96	0.0206	0.1784
5-link	6 129 294.1	1 282 964.12	171 424.9	0.0272	0.1731
Oryza sativa					
Single-link	544 450	777 126	1723	0.0032	0.5880
2-link	544 409.5	771 706.75	1763.5	0.0032	0.5864
3-link	531 223.34	581 126.43	14 949.66	0.0274	0.5224
4-link	514 636.5	457 709.57	31 536.5	0.05772	0.4707
5-link	496 927.81	419 862.93	49 245.19	0.0902	0.4580
Sorghum bicolor					
Single-link	2 798 705	1 246 093	5950	0.0021	0.3081
2-link	2 798 705	1 246 093	5950	0.0021	0.3081
3-link	2 774 210	737 405.33	30445	0.0109	0.2100
4-link	2 754 055.07	680 983.11	50 599.92	0.01804	0.1982
5-link	2 728 152.31	628 988.71	76 502.69	0.0273	0.1874

TP, FP and FN values for `k-link` clustering (for various k) when compared with the reference clustering. The proportion of pairs in the reference clustering which were incorrectly separated (i.e. similar to Type I error) in the `k-link` clustering was calculated using FN/(TP + FN). The proportion of pairs in the non-reference clustering which were not together in the reference(similar to Type II error), was calculated using FP/(TP + FP).

4 DISCUSSION AND CONCLUSION

Understanding how error rates vary in response to parameter selection is crucial to obtaining the optimal (in this case, the

Table 4. Summary statistics for cluster sizes in the reference and `k-link` clusterings

Clustering	Min	Q1	Median	Mean	Q3	Max	No. of Clusters
Caenorhabditis elegans							
Reference	1	1	2	6.605	5	1321	23 002
Single-link	1	1	2	7.455	6	1407	20 378
2-link	1	1	2	7.455	6	1407	20 378
3-link	1	1	2	6.830	5	1407	23 019
4-link	1	1	2	6.471	5	1407	25 402
5-link	1	1	3	6.289	5	1407	27 420
Oryza sativa							
Reference	1	1	1	2.900	3	337	24 525
Single-link	1	1	1	3.668	3	713	19 387
2-link	1	1	1	3.668	3	713	19 390
3-link	1	1	2	3.413	3	712	22 887
4-link	1	1	2	3.367	3	653	25 786
5-link	1	1	2	3.420	4	653	28 136
Sorghum bicolor							
Reference	1	1	2	4.740	4	1060	20 361
Single-link	1	1	2	5.656	4	1354	17 062
2-link	1	1	2	5.656	4	1354	17 062
3-link	1	1	2	5.266	4	1086	19 102
4-link	1	1	2	5.075	4	1086	20 862
5-link	1	1	2	4.994	4	1080	22 572

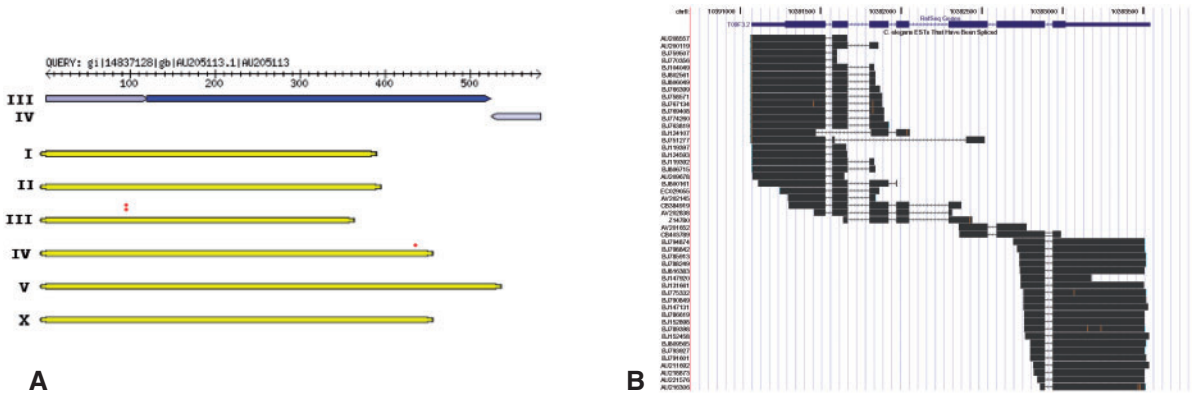

Fig. 1. Manual inspection of putative chimeras identified by `k-link` appear to be a mixture of true chimeric sequences and low-abundance transcripts. In (**A**), the EST shown is a *C.elegans* chimera of two genes, one fragment from *dod-6* on chromosome III (consisting of two exons–location is marked with two red asterisks), and the other fragment from *col-103* on chromosome IV (one exon–marked with one red asterisk). This chimera was the only sequence suggesting a link between the transcripts from these two separate genes. (**B**) ESTs derived from the T09F3.2 gene (http://genome.ucsc.edu, *C.elegans*, May 2008). A rare splice variant (BJ751277) spans the two otherwise non-overlapping EST clusters, and for this reason is marked as a putative chimera.

most reference-like) clustering. Past studies have indicated that a sequence identity threshold of 90%, with an alignment length between 25 and 40 bp, minimizes the Type I and Type II errors introduced by these thresholds. Despite these findings, the majority of clustering algorithms still maintain high sequence–similarity thresholds. Our results emphasize that single-linkage algorithms perform poorly (in terms of Type II error) when using the optimal sequence similarity threshold. Despite the high level of activity in this research area, no one has tried varying the cluster-merging algorithm to take into account the number of links between clusters. We feel that it is a natural progression from the study by Wang *et al*. (2004) to investigate how the linkage algorithm used in EST clustering influences Type I and Type II error rates.

To our knowledge, no algorithm has been developed which merges EST clusters based on a *k*-link threshold. Using our algorithm, `k-link`, we have shown that Type II error can be reduced by increasing the number of links required between clusters. It is difficult to say whether the reduction in error was due to appropriate handling of chimeric sequences, or the increased linkage requirements mitigating some of the effect of unmasked repetitive sequence. Given the very low frequency of chimeric sequences, intuition suggests that the significant initial decrease in Type II error as k is increased from 2 to 3 shows the algorithm overcoming the effect of the chimeric sequences. The substantial 'base' Type II error suggests that the remaining errors are due to uninformative sequences, be it repeats or stretches of low-complexity sequence.

The traditional definitions of Type I and Type II errors in statistics often imply an inverse relationship between the two measures. Many of the EST clustering approaches, described in the literature, have minimized Type I error at the cost of increased Type II error. This coincides with the prevailing view that Type I errors are more difficult to correct post-clustering than Type II errors (Hazelhurst *et al*., 2008). However, as the `k-link` algorithm permits an EST to belong to multiple clusters, this inverse relationship does not strictly hold. The negation of chimeric sequences, combined with the tolerance of non-disjoint clusters, causes the Type I error to grow disproportionately slower than the rate of loss in Type II errors.

The growth in Type I error as k is increased may be further impeded While there was a slight increase in type I error as k was increased, this is likely to be reduced by introducing an adaptive linkage algorithm. Such an algorithm would take into account the size of the compared clusters and adjust the required number of links accordingly.

Our study has also shown that unmasked uninformative sequences (such as repeats, vector sequences or low-complexity stretches) can have a significant impact on the Type II clustering error. No clustering algorithm is impervious to the effects of unmasked repeat sequences. For a novel species, it is difficult to know how much masking is required. While we removed some super-clusters during the construction of the datasets, many repeat sequences still filtered through. This highlights the need to inspect cluster sizes produced by a clustering algorithm to identify whether sequence masking was adequate. We suggest that users of `k-link` perform a rough clustering (without parameter estimation, option –E) to identify any significantly large clusters. Future research will involve evaluating linkage-related error by running `k-link` on a EST dataset from a species with a well-characterized repeat database. We would like to emphasize that the RBR supplementary library option used in this study was implemented at our request—and its efficiency has not been benchmarked.

We have shown that single-linkage clustering produces a significant level of type II error when using less-stringent (but more sensitive) sequence—similarity parameters. By using the optimal number of required links (through probabilistic estimation), we mitigate the effect of chimeras, and potentially of some repeat sequences, in the EST dataset. The Type I errors introduced by using the smallest number of links required to overcome the effect of chimeric sequences, while minimized for the constant link-level case, may be further reduced by using an adaptive linkage algorithm.

The introduction of variable linkage to EST clustering should be seen as an additional parameterization of the clustering algorithm, allowing researchers to fine-tune the desired proportion of Type I and Type II errors. Understanding when either error type is preferable will be dependent on the analysis conducted post-clustering.

ACKNOWLEDGEMENTS

The authors are grateful to Andrew George, for encouraging research in this area, and for the highly valued advice and suggestions from Peter Humburg, Ketil Malde and Scott Hazelhurst. In addition, we would like to thank Ketil Malde for making an extension to the RBR software. Finally, we wish to thank the two anonymous referees for their helpful comments.

Funding: CSIRO's Transformational Biology Platform (in part).

Conflict of Interest: none declared.

REFERENCES

Barker,G. *et al*. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, **19**, 421–422.

Boguski,M.S. and Schuler,G.D. (1995) Establishing a human transcript map. *Nat. Genet.*, **10**, 369–371.

Burke,J. *et al*. (1999) d2_cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.*, **9**, 1135–1142.

Hazelhurst,S. *et al*. (2008) An overview of the wcd EST clustering tool. *Bioinformatics*, **24**, 1542–1546.

Hide,W. *et al*. (1994) Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J. Comput. Biol.*, **1**, 199–215.

Hillier,L.D. *et al*. (1996) Generation and analysis of 280 000 human expressed sequence tag. *Genome Res.*, **6**, 807–828.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Jaccard,P. (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat*, **37**, 547–579.

Kent,W.J. (2002) BLAT - the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kim,N. *et al*. (2005) ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res.*, **15**, 566–576.

Klastorin,T.D. (1980) Merging groups to maximize object partition comparison. *Psychometrika*, **45**, 425-433.

Korf,I. *et al*. (2003) *BLAST: An Essential Guide to the Basic Alignment Search Tool*. O'Reilly & Associates, Sebastopol.

Levin,B. (1981) A representation for multinomial cumulative distribution functions. *Ann. Stat.*, **9**, 1123–1126.

Malde,K. *et al*. (2006) RBR: library-less repeat detection for ESTs. *Bioinformatics*, **22**, 2232–2236.

Malde,K. and Jonassen,I. (2008) Repeats and EST analysis for new organisms. *BMC Genomics*, **9**, 1471–2164.

Meila,M. (2007) Comparing clusterings an information based distance. *J. Multivar. Anal.*, **98**, 873–895.

Patterson,A.H. *et al*. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Parkinson,J. *et al*. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*, **3**, 31.

Pertea,G. *et al*. (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Ptitsyn,A. and Hide,W. (2005) CLU: a new algorithm for EST clustering. *BMC Bioinformatics*, **6** (Suppl. 2), S3.

Quackenbush,J. *et al*. (2001) The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.

Sorek,R. and Safer,H. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.*, **31**, 1067–1074.

Tang,J.F. *et al*. (2006) QualitySNP: a pipeline for detecting Single nucleotide polymorphisms and insertions/deletions in EST data from diploids and polyploidy species. *BMC Bioinformatics*, **7**, Article 438.

Wang,J.Z. *et al*. (2004) EST clustering error evaluation and correction. *Bioinformatics*, **20**, 2973–2984.