d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences

John Burke,[1,4] Dan Davison,[2] and Winston Hide[3]

[1]*Pangea Systems, Oakland, California 94612 USA;* [2]*Bioinformatics Department, Bristol-Myers Squibb Pharmaceutical Research Institute, Wallingford, Connecticut 06492-7660 USA;* [3]*South African National Bioinformatics Institute, Bellville 7535, University of the Western Cape, South Africa*

Several efforts are under way to condense single-read expressed sequence tags (ESTs) and full-length transcript data on a large scale by means of clustering or assembly. One goal of these projects is the construction of gene indices where transcripts are partitioned into index classes (or clusters) such that they are put into the same index class if and only if they represent the same gene. Accurate gene indexing facilitates gene expression studies and inexpensive and early partial gene sequence discovery through the assembly of ESTs that are derived from genes that have yet to be positionally cloned or obtained directly through genomic sequencing. We describe d2_cluster, an agglomerative algorithm for rapidly and accurately partitioning transcript databases into index classes by clustering sequences according to minimal linkage or "transitive closure" rules. We then evaluate the relative efficiency of d2_cluster with respect to other clustering tools. UniGene is chosen for comparison because of its high quality and wide acceptance. It is shown that although d2_cluster and UniGene produce results that are between 83% and 90% identical, the joining rate of d2_cluster is between 8% and 20% greater than UniGene. Finally, we present the first published rigorous evaluation of under and over clustering (in other words, of type I and type II errors) of a sequence clustering algorithm, although the existence of highly identical gene paralogs means that care must be taken in the interpretation of the type II error. Upper bounds for these d2_cluster error rates are estimated at 0.4% and 0.8%, respectively. In other words, the sensitivity and selectivity of d2_cluster are estimated to be >99.6% and 99.2%.

[Supplementary material to this paper may be found online at www.genome.org and at www.pangeasystems.com.]

The rapid generation of single-read sequence from the 3′ ends and 5′ portions of sufficiently expressed mRNAs (popularly referred to as ESTs) (Adams et al. 1991; Okubo et al. 1991; Wilcox et al. 1991) has resulted in the discovery of many genes well before the projected completion of the human genome project and before the completion of sequencing efforts in other organisms (Adams et al. 1992; Matsubara and Okubo 1993; Venter 1993). Because the source information is available for every EST, the intracluster representation of libraries from discrete disease and developmental states can be contrasted and hence large-scale expression studies can be performed (Okubo et al. 1992, 1994; Adams et al. 1995; Vasmatzis et al. 1998). EST sequence has enabled the construction of a physical map of the human genome (Hudson et al. 1995) as well a gene map that localizes many genes with respect to the markers of the physical map (Schuler et al. 1996). The utility of EST data has also been increased greatly by the establishment of centralized databases (Boguski et al. 1993).

The fragmented nature and vast quantity of EST data pose an obstacle to harvesting the full potential from this data source. Hence, several projects are in progress to construct information frameworks, called gene indices, where the fragmented, error-prone EST data and the known gene sequence data can be consolidated and placed in a correct pathologic and mapping context indexed by gene such that all data concerning a single gene is in a single index class and each index class contains the information for only one gene. Algorithmically, these projects all comprise some type of cluster analysis in which sequence similarity and possibly other criteria are used to form the clusters or index classes. Below, we detail some of the clustering methods used in several gene index projects.

The Institute for Genome Research (TIGR) Gene Index (TGI) (http://www.tigr.org/tdb/hgi/hgi.html; Adams et al. 1995; Sutton et al. 1995; White and Kerlavage 1996) is constructed by assembling full-length sequence [from the Expressed Gene and Anatomy Database (EGAD); White and Kervalage 1996)] and ESTs to form tentative human consensi (THCs). The **THC_BUILD** program (G. Sutton, pers. comm.) constructs index groups according to the following schedule: (1) BLAST and FASTA (Pearson 1990) are used to identify all sequence overlaps, (2) all detected overlaps

[4]**Corresponding author. Present address: Pangea Systems, Oakland, California 94612 USA.**
E-MAIL jburke@pangeasystems.com; FAX (510) 628-0108.

9:1135–1142 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

are stored in a relational database, (3) the CLUB program forms transitive closure groups from the overlap database, and (4) groups are subjected to assembly using TIGR assembler (Sutton et al. 1995; http://www.tigr.org/hgi/hgi_info.html). The assembler also imposes matching constraints on the ends of sequences and a minimum sequence identity within an index group. Most sequence assembly programs share similar properties with TIGR assembler; however, it is worth noting that some assemblers, such as the PHRAP package, incorporate sequence quality data derived from sequence traces into the assembly process (P. Green, pers. comm.) allowing for the incorporation of higher error data.

In UniGene (Boguski and Schuler 1995; Schuler at al. 1996), genes are indexed by forming initial groups with full-length sequences (mRNA and genomic). Then groups are formed within the EST set and between the ESTs and the initial groups. EST matches are not allowed to join distinct initial groups and clusters without a polyadenylation signal or unless at least two 3′ ESTs are discarded. Finally, clone information was used to further join clusters and singleton clusters, and unmatched ESTs were added to nonsingleton clusters at lower stringency levels (http://ncbi.nlm.nih.gov/UniGene/TXT/build.html). To enhance the speed of the clustering, a two-phase searching process was used in which two sequences were compared with a constrained Smith–Waterman algorithm only if they shared two common words of length 13 separated by no more than 2 bases. The STACK gene indexing system that uses d2_cluster is covered in the discussion section of this paper and in a companion paper.

In this article we describe d2_cluster, an algorithm for clustering sequences into index classes. First, we describe the basic method (which we call D20 to distinguish it from future variants of the algorithm). We then demonstrate the utility of d2_cluster by performing a data analysis of the results of clustering a moderate sized data set (~43,000 sequences). We characterize the relative performance of d2_cluster and UniGene clustering showing that although the behavior of the two algorithms is very similar, there are measurable differences in the rate of merging sequences into clusters. Finally, we derive estimates of the absolute type I and type II error rates (the probability of under or over clustering) for d2_cluster.

Description of the d2_cluster Method (D20)

d2_cluster is an agglomerative clustering method [every sequence begins in its own cluster, and the final clustering is constructed through a series of mergers (Johnson and Witchern 1994)]. d2_cluster can be described in terms of minimal linkage clustering (sometimes called single linkage or transitive closure in the sequence analysis literature). The term transitive closure refers to the property that any two sequences with a given level of similarity will be in the same cluster; hence, **A** and **B** are in the same cluster even if they share no similarity but there exists a sequence **C** with enough similarity to both **A** and **B**. The criterion for joining clusters is the detection of two sequences that share a window of (**Window_Size**) bases that is (**Stringency**) percent or more identical. The only criterion for clustering is sequence overlap and source or annotation information is not used. To detect the overlap criterion, we use the d2 algorithm and set parameters and threshold values as described in previous work (Torney et al. 1990; Hide et al. 1994; Wu et al. 1997). The initial and final state of the algorithm is a partition of the input sequences in which each sequence is in a cluster and no sequence appears in more than one cluster.

For ease of notation, let the following conventions hold:

1. We signify the d2 distance between two sequences, say **A** and **B**, as d2(**A**,**B**).
2. Given two clusters, e.g., clusters i and j, the operation *MERGE*(cluster i, cluster j), also denoted *MERGE*(cluster i ← cluster j), means that all sequences in cluster j are assigned to cluster i.
3. The database to be clustered contains *N* sequences that are numbered 0 through (*N* − 1). Let sequence (i) be denoted Si or S(i).
4. The membership of sequence Si is denoted Ci.

The notation d2(**A**,**B**) is conveniently used, but, of course, d2(,) is not a function of only **A** and **B** but also of various parameters (specified in Torney et al. 1990; Hide et al. 1994; Wu et al. 1997). The *MERGE* operation can be expressed in terms of convention 4 above: For all sequences, Sr, such that Cr = j, Cr is reset to be Cr = i.

We describe the progression of d2_cluster inductively in that we first detail what happens in the first two iterations (I0 and I1) and then describe how one performs iteration (i + 1) given that iteration (i) has been completed. Technically speaking, it is sufficient to state only the first step and then to give the step (i) to step (i + 1) instructions, but we detail the first two steps for clarity.

The clustering is finished when *N* iterations are completed. Transitive closure is obtained because clusters are joined if they contain any sequences with sufficient identity. d2_cluster, as described above, can be mapped to the minimal linkage algorithm commonly seen in the statistics and engineering texts, and details of this are given in the online supplement to this paper (http://www.genome.org and www.pangeasystems.com).

Initial state: Each sequence is in its own cluster. (i.e., Si is in cluster i or Ci = i).

First iteration I1: The first sequence in the database, S0, is selected as a query. For each sequence in *Si* ($1 \le i < N$), *MERGE*(cluster C0) ← (cluster *Ci*) if d2(*S*0,*Si*) < THRESHHOLD.

Second iteration I2: The second sequence in the database (*S*1) is now selected as a query. Note that *C*1 = 1 unless sequence 1 was merged into cluster 0 during step I1. For all sequences, Si ($2 \le i < N$), *MERGE*(cluster C1 ← cluster *Ci*) if *d2*(*S*1,*Si*) < THRESHHOLD.

(*k*)th iteration I(*k*): Suppose we have completed ($k - 1$) iterations. We select sequence *Sk* as a query. For all seqs, *Si* ($k + 1 \le i < N$), *MERGE*(cluster *Ck* ← cluster *Ci*) if d2(*Sk,Si*) < THRESHHOLD.

Data and Error Analysis

Relative Performance of d2_cluster and UniGene

The first step in evaluating d2_cluster performance was to compare it with other methods. The UniGene data set was chosen for comparison due to its high quality and wide acceptance. Because an implementation of the UniGene clustering suite was not available, it was not possible to actually run the UniGene clustering algorithm. Hence, the base sequences from a version of UniGene (Rat section, UniGene build 19, August 1998) were processed with d2_cluster, and the d2_cluster/UniGene groupings were compared. Unless explicitly stated otherwise, all UniGene clusters referenced in this paper come from build 19. Rat UniGene had 43,612 ESTs and full-length mRNA sequences. Because the screening information for UniGene was unavailable, it was necessary to rescreen the data set against repetitive elements and mitochondrial sequence. The cross_match (P. Green, unpubl.) program was used for this purpose, and an "x" screening marker was left over sequence positions matching repetitive elements. Sequences with <100 bases of nonscreened sequence remaining were dropped from the analysis, and the remaining 42,441 sequences were clustered with d2_cluster (at the parameter settings **Window_Size** = 100, **Stringency** = 0.9, **Min_Seq** = 100, and **Rev_Comp** = 1. At these settings two sequences are placed into the same cluster if they contained a window of size 100 bases with at least 90% identity. Sequences with <100 bases are disregarded, and the reverse strand is searched). Approximately 31 hr were required to complete the cluster analysis on a SUN machine E450 with a 400-MHz processor. As stated above, the clustering of sequence with d2_cluster was made strictly on the basis of sequence similarity, and annotation information was not used. After this processing, every sequence was a member of exactly one d2 cluster and one UniGene cluster.

Table 1 compares the cluster size distribution for the UniGene and d2_cluster groupings. d2_cluster produced ~20% fewer singleton sequences (6463 to 5198) and reduced the overall number of clusters by 10% (15,225–13,756). Generally, the numbers of smaller clusters are reduced, whereas larger clusters appear with slightly higher frequency.

To further quantify the higher join rate of d2_cluster and to assess relative quality, a subsetting analysis was executed and is summarized in Figure 1. Among all UniGene clusters, a negligible number, 60 (or <0.5% of UniGene clusters), were merges of d2 clusters. Conversely, 1078 d2_cluster groups (or ~8% of total) are merges of UniGene clusters, and this number can serve as another measure of the more aggressive joining of d2_cluster. The average of the three estimates (8%, 10%, 20%) of increased joining is 13%. A total of 12,389 clusters (or 83% of UniGene clusters and 90% of d2 clusters) are identical between UniGene and d2_cluster indicating that although there are measurable differences in the clustering rate, the answers produced by the two algorithms are consistent on a large scale.

There are several situations that could cause sequence members of a single (d2_cluster/UniGene) generated cluster to appear in several different clusters generated by the other method. Possibilities include (1) the failure of one method to join a valid cluster that was generated with the other method, (2) the introduction of a false join by one of the methods, or (3) the use of different clustering criterion by the methods. For example, UniGene uses clone information to augment EST sequence clustering, whereas d2_cluster uses no annotation or source information. An example of a cluster formed by d2_cluster that was fragmented into two UniGene clusters is given in Figure 2. The JAVA program CRAWview (Chow and Burke 1999) generates the color cluster representation (Fig. 2A). From the CRAW report shown in Figure 2A and the sequence alignment given in Figure 2B (online supplement to

Table 1. Cluster size distribution for UniGene and d2_cluster

Histogram of cluster sizes		
cluster size	UniGene (Rat-build 19)	d2_cluster
Singleton clusters	6463	5189
2	3496	3298
3–4	3002	2971
5–8	1635	1602
9–16	491	531
17–32	111	127
33–64	21	27
65–128	3	8
129–256	3	2
257–512	1	0
Total clusters	15,225	13,756

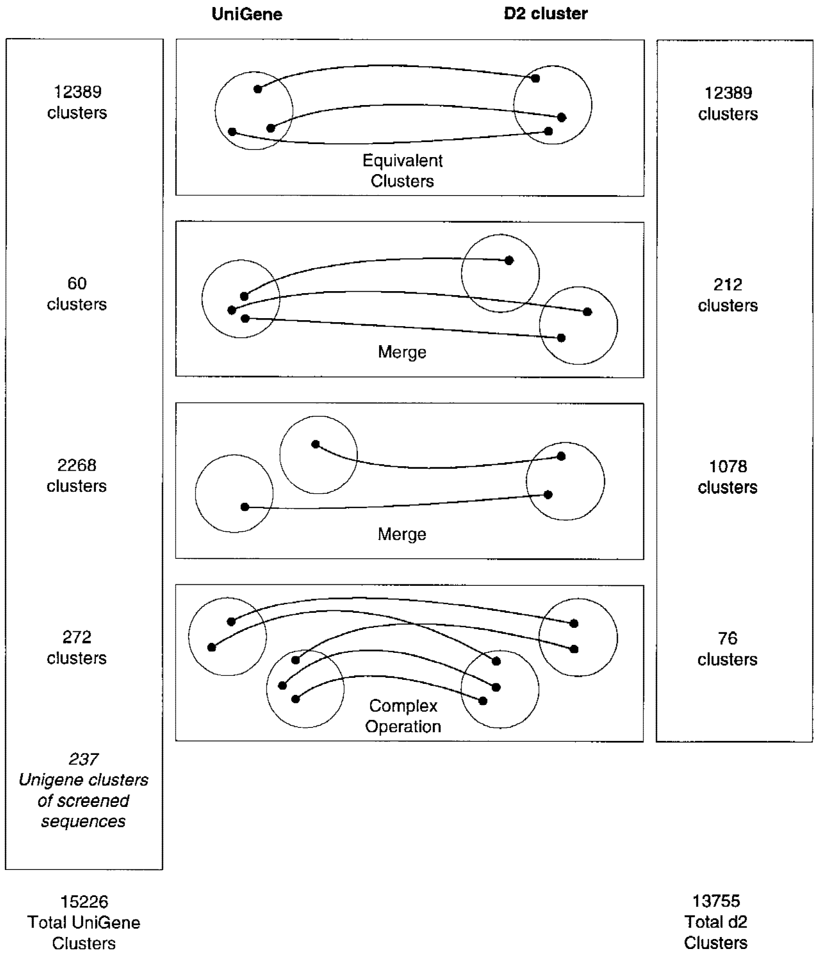


Figure 1 Subsetting comparison of UniGene and d2_cluster. Cluster equivalence means that all elements in one cluster are also present in the other cluster, and vice versa. Out of 14,989 (= 15,226 − 237) original UniGene clusters and 13,755 d2 clusters, 12,389 (or 83% of UniGene clusters and 90% of d2 clusters) are equivalent. Two hundred thirty-seven UniGene clusters were not considered in the analysis because they were composed of sequences that were screened out in our vector and repetitive elements screening stage.

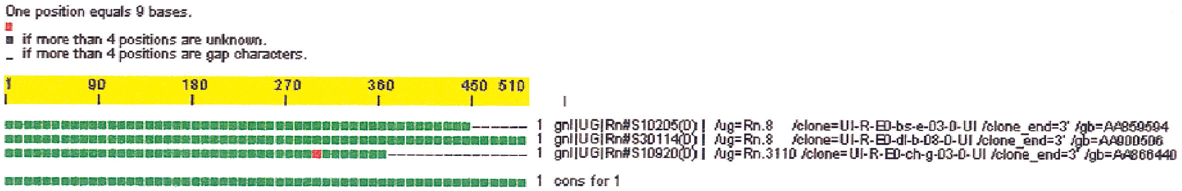
this article, www.genome.org and www.pangeasystems.com), it can clearly be seen that the elements of two distinct UniGene clusters (Rn.8 and Rn.3110) should probably be together as they are >98% identical over their entire lengths. Figure 3 gives another, more interesting case of a d2 cluster that contains isozymes of the rat cytochrome P-450 gene with membership corresponding to seven separate UniGene clusters. d2_cluster has put all of these sequences together because of regions of high identity (as seen in Fig. 3B included in the online supplement to this article). UniGene has separated isozymes into distinct clusters, although UniGene clusters Rn.18603, Rn.10842, and Rn.9104 should probably form a single cluster according to reasonable clustering rules due to their perfect assembly into subgroup 2 and high overlap (full alignment is not shown but is available from the authors upon request).

Estimating the Absolute Error Rate

Although a comparison of d2_clustering with UniGene can compare the relative efficiency of the two clustering methods and can demonstrate the tendency of d2_cluster to partition sets into fewer classes than UniGene, no information about the absolute correctness of either clustering method can be inferred. To gauge absolute upper bounds for the error rates of d2_cluster, we performed two rigorous analyses of the groups formed by d2_cluster to analyze sensitivity and selectivity (or, more precisely, to estimate upper bounds for the actual type I and type II errors).

In hypothesis testing the type II error rate is the probability of incorrectly rejecting the null hypothesis when it is true. We define the null hypothesis to be that any two sequences do not belong in the same cluster. In this case the type II error, in the context of sequence clustering, is the probability of placing two sequences into the same cluster by mistake, or, in other words, the probability of over clustering. We say that the presence of a type II error can be discounted in cases in which a single, high-quality assembly can represent the cluster in which overlap satisfies the matching criterion. At this point it must be noted that there are caveats to this error analysis. For example, there are biological reasons, such as gene paralogs, that ESTs that are actually from distinct genes might be perfectly alignable. Some situations, such as alternative splicing, require that more than a single consensus represent the entire cluster. In such cases, type II error can be excluded if the cluster consensi can be shown to contain sufficiently large domains of identity with each other. The Rat EST clusters formed by d2_cluster were aligned and analyzed with CRAW (Burke et al. 1998), a program that creates a minimal number of high-quality consensus representatives for a cluster and discriminates and models alternative gene forms. Because CRAW provides a control of the variability between a consensus and its member sequences and enforces a minimum overlap criterion, no type II errors should occur in the clusters that are represented by a single consensus. We set CRAW stringency such that the alignment of a sequence with the subcluster consensus contained no window of 50 bases with >10% mismatch.

Following the strategy outlined above, we empiri-



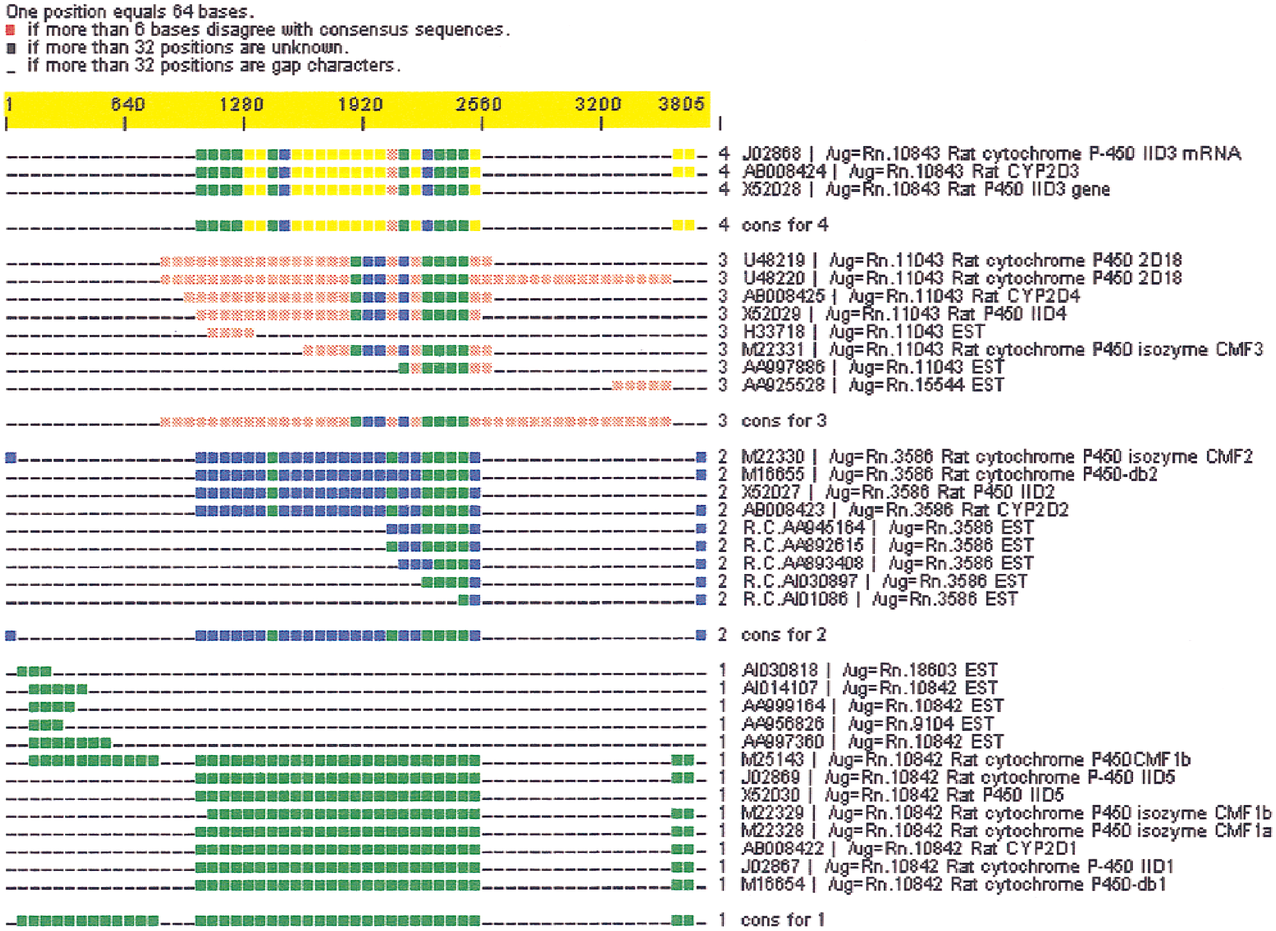
One position equals 9 bases.
▪ if more than 4 positions are unknown.
_ if more than 4 positions are gap characters.

1 gnl|UG|Rn#S10205(D) | /ug=Rn.8 /clone=UI-R-E0-bs-e-03-0-UI /clone_end=3' /gb=AA859594
1 gnl|UG|Rn#S30114(D) | /ug=Rn.8 /clone=UI-R-E0-dl-b-08-0-UI /clone_end=3' /gb=AA900506
1 gnl|UG|Rn#S10920(D) | /ug=Rn.3110 /clone=UI-R-E0-ch-g-03-0-UI /clone_end=3' /gb=AA866440

1 cons for 1

Figure 2 (*A*) CRAW report (Burke et al. 1998) for a cluster formed by d2_cluster that contains two UniGene clusters: Rn.8 and Rn.3110. (*B*) (Available as on online supplement to this paper at www.genome.org and at the authors web site at www.pangeasystems.com) Interleaved sequence alignment shows a >300-bp region of near perfect match.

cally estimate an upper bound on this error by bounding by above the number of clusters with type II errors. After processing the 13,755 clusters formed by d2_cluster, all but 1617 can be represented by a single CRAW consensus sequence. Thus, an initial upper bound for the type II error rate is 11.8% (=100 * 1617/13,755). A more stringent upper bound was obtained by inspection of the multiple alignments of the 1617 clusters with multiple CRAW consensi to identify cases in which multiple alignments of the different consensi still contained a window of 100 bases with at least 90% identity. Fewer than 106 clusters could not easily be confirmed to satisfy this constraint. Hence, the type II error is bounded by above, for this data set, by 0.8% (=100 * 106/13,755). These 106 cases are most likely not all errors on the part of d2_cluster, and many are



One position equals 64 bases.
▪ if more than 6 bases disagree with consensus sequences.
▪ if more than 32 positions are unknown.
_ if more than 32 positions are gap characters.

4 J02868 | /ug=Rn.10843 Rat cytochrome P-450 IID3 mRNA
4 AB008424 | /ug=Rn.10843 Rat CYP2D3
4 X52028 | /ug=Rn.10843 Rat P450 IID3 gene

4 cons for 4

3 U48219 | /ug=Rn.11043 Rat cytochrome P450 2D18
3 U48220 | /ug=Rn.11043 Rat cytochrome P450 2D18
3 AB008425 | /ug=Rn.11043 Rat CYP2D4
3 X52029 | /ug=Rn.11043 Rat P450 IID4
3 H33718 | /ug=Rn.11043 EST
3 M22331 | /ug=Rn.11043 Rat cytochrome P450 isozyme CMF3
3 AA997886 | /ug=Rn.11043 EST
3 AA925528 | /ug=Rn.15544 EST

3 cons for 3

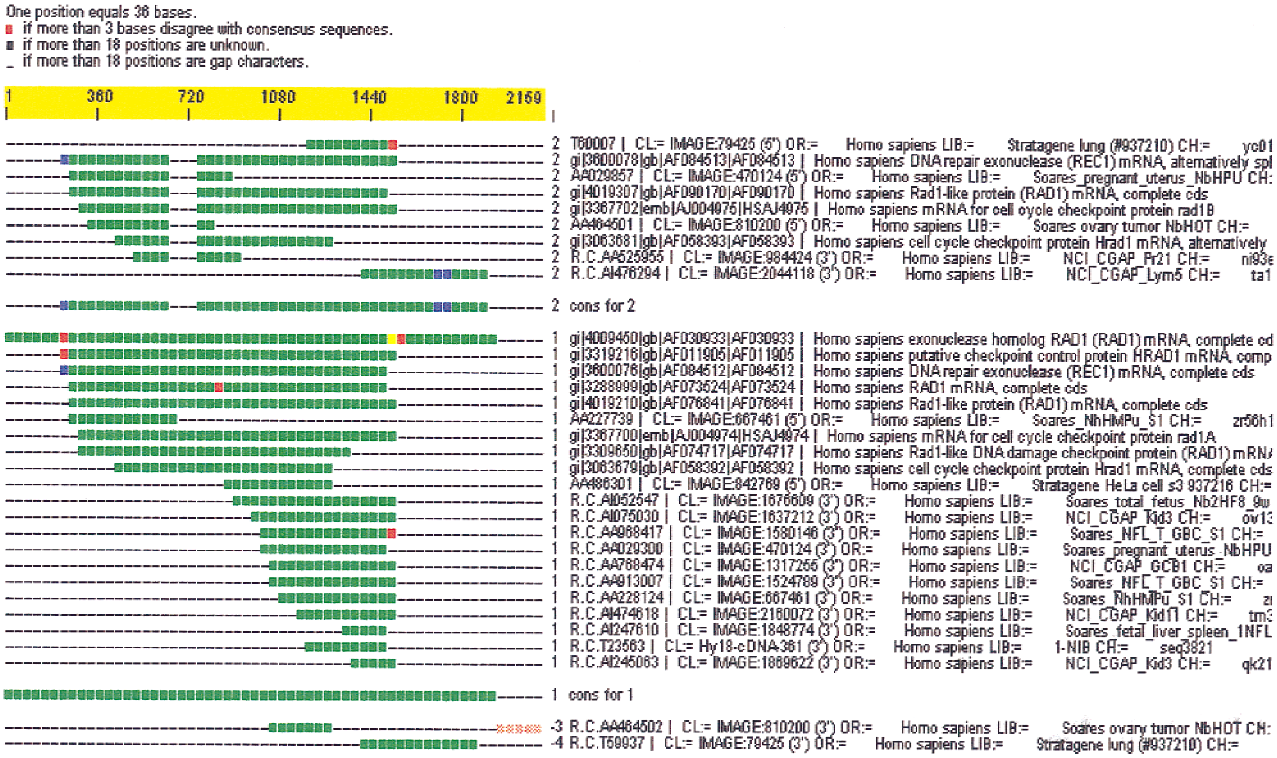
2 M22330 | /ug=Rn.3586 Rat cytochrome P450 isozyme CMF2
2 M16655 | /ug=Rn.3586 Rat cytochrome P450-db2
2 X52027 | /ug=Rn.3586 Rat P450 IID2
2 AB008423 | /ug=Rn.3586 Rat CYP2D2
2 R.C.AA945164 | /ug=Rn.3586 EST
2 R.C.AA892615 | /ug=Rn.3586 EST
2 R.C.AA893408 | /ug=Rn.3586 EST
2 R.C.AI030897 | /ug=Rn.3586 EST
2 R.C.AI01086 | /ug=Rn.3586 EST

2 cons for 2

1 AI030818 | /ug=Rn.18603 EST
1 AI014107 | /ug=Rn.10842 EST
1 AA999164 | /ug=Rn.10842 EST
1 AA956826 | /ug=Rn.9104 EST
1 AA997360 | /ug=Rn.10842 EST
1 M25143 | /ug=Rn.10842 Rat cytochrome P450CMF1b
1 J02869 | /ug=Rn.10842 Rat cytochrome P-450 IID5
1 X52030 | /ug=Rn.10842 Rat P450 IID5
1 M22329 | /ug=Rn.10842 Rat cytochrome P450 isozyme CMF1b
1 M22328 | /ug=Rn.10842 Rat cytochrome P450 isozyme CMF1a
1 AB008422 | /ug=Rn.10842 Rat CYP2D1
1 J02867 | /ug=Rn.10842 Rat cytochrome P-450 IID1
1 M16654 | /ug=Rn.10842 Rat cytochrome P450-db1

1 cons for 1

Figure 3 (*A*) CRAW report for a d2 cluster containing isozymes of mouse cytochrome P-450. Seven UniGene clusters (Rn.10843, Rn.3586, Rn.18603, Rn.10842, Rn.9104, Rn.11043, and Rn.15544) are merged. (*B*) (Online supplement available at www.genome.org and www.pangeasystems.com) Interleaved multiple alignment showing a region of 240 bases with high identity alignment between all four cluster assemblies. d2_cluster has put all of these sequences together because of regions of high identity (as seen in Fig. 3B). UniGene has separated isozymes into distinct clusters, although UniGene clusters Rn.18603, Rn.10842, and Rn.9104 should probably form a single cluster according to reasonable clustering rules due to their perfect assembly into subgroup 1 and high overlap.



```
One position equals 36 bases.
■ if more than 3 bases disagree with consensus sequences.
■ if more than 18 positions are unknown.
_ if more than 18 positions are gap characters.
```

```
2  T60007 |  CL:= IMAGE:79425 (5') OR:=      Homo sapiens LIB:=      Stratagene lung (#937210) CH:=      yc01
2  gi|3600078|gb|AF084513|AF084513 |  Homo sapiens DNA repair exonuclease (REC1) mRNA, alternatively spl
2  AA029857 |  CL:= IMAGE:470124 (5') OR:=      Homo sapiens LIB:=      Soares_pregnant_uterus_NbHPU CH:
2  gi|4019307|gb|AF090170|AF090170 |  Homo sapiens Rad1-like protein (RAD1) mRNA, complete cds
2  gi|3367702|emb|AJ004975|HSAJ4975 |  Homo sapiens mRNA for cell cycle checkpoint protein rad1B
2  AA464501 |  CL:= IMAGE:810200 (5') OR:=      Homo sapiens LIB:=      Soares ovary tumor NbHOT CH:=
2  gi|3063681|gb|AF058393|AF058393 |  Homo sapiens cell cycle checkpoint protein Hrad1 mRNA, alternatively
2  R.C.AA525955 |  CL:= IMAGE:984424 (3') OR:=      Homo sapiens LIB:=      NCI_CGAP_Pr21 CH:=      ni93e
2  R.C.AI476294 |  CL:= IMAGE:2044118 (3') OR:=      Homo sapiens LIB:=      NCI_CGAP_Lym5 CH:=      ta1

2  cons for 2

1  gi|4009450|gb|AF030933|AF030933 |  Homo sapiens exonuclease homolog RAD1 (RAD1) mRNA, complete cd
1  gi|3319216|gb|AF011905|AF011905 |  Homo sapiens putative checkpoint control protein HRAD1 mRNA, comp
1  gi|3600076|gb|AF084512|AF084512 |  Homo sapiens DNA repair exonuclease (REC1) mRNA, complete cds
1  gi|3288999|gb|AF073524|AF073524 |  Homo sapiens RAD1 mRNA, complete cds
1  gi|4019210|gb|AF076841|AF076841 |  Homo sapiens Rad1-like protein (RAD1) mRNA, complete cds
1  AA227739 |  CL:= IMAGE:667461 (5') OR:=      Homo sapiens LIB:=      Soares_NhHMPu_S1 CH:=      zr56h1
1  gi|3367700|emb|AJ004974|HSAJ4974 |  Homo sapiens mRNA for cell cycle checkpoint protein rad1A
1  gi|3309650|gb|AF074717|AF074717 |  Homo sapiens Rad1-like DNA damage checkpoint protein (RAD1) mRN
1  gi|3063679|gb|AF058392|AF058392 |  Homo sapiens cell cycle checkpoint protein Hrad1 mRNA, complete cds
1  AA486301 |  CL:= IMAGE:842769 (5') OR:=      Homo sapiens LIB:=      Stratagene HeLa cell s3 937216 CH:=
1  R.C.AI052547 |  CL:= IMAGE:1676609 (3') OR:=      Homo sapiens LIB:=      Soares_total_fetus_Nb2HF8_9w
1  R.C.AI075030 |  CL:= IMAGE:1637212 (3') OR:=      Homo sapiens LIB:=      NCI_CGAP_Kid3 CH:=      ov13
1  R.C.AA968417 |  CL:= IMAGE:1580146 (3') OR:=      Homo sapiens LIB:=      Soares_NFL_T_GBC_S1 CH:=
1  R.C.AA029300 |  CL:= IMAGE:470124 (3') OR:=      Homo sapiens LIB:=      Soares_pregnant_uterus_NbHPU
1  R.C.AA768474 |  CL:= IMAGE:1317255 (3') OR:=      Homo sapiens LIB:=      NCI_CGAP_GCB1 CH:=      oa
1  R.C.AA913007 |  CL:= IMAGE:1524789 (3') OR:=      Homo sapiens LIB:=      Soares_NFL_T_GBC_S1 CH:=
1  R.C.AA228124 |  CL:= IMAGE:667461 (3') OR:=      Homo sapiens LIB:=      Soares_NhHMPu_S1 CH:=      zi
1  R.C.AI474618 |  CL:= IMAGE:2160072 (3') OR:=      Homo sapiens LIB:=      NCI_CGAP_Kid11 CH:=      tm3
1  R.C.AI247610 |  CL:= IMAGE:1848774 (3') OR:=      Homo sapiens LIB:=      Soares_fetal_liver_spleen_1NFL
1  R.C.T23563 |  CL:= Hy18-cDNA-361 (3') OR:=      Homo sapiens LIB:=      1-NIB CH:=      seq3821
1  R.C.AI245063 |  CL:= IMAGE:1869622 (3') OR:=      Homo sapiens LIB:=      NCI_CGAP_Kid3 CH:=      qk21

1  cons for 1

-3 R.C.AA464502 |  CL:= IMAGE:810200 (3') OR:=      Homo sapiens LIB:=      Soares ovary tumor NbHOT CH:
-4 R.C.T59937 |  CL:= IMAGE:79425 (3') OR:=      Homo sapiens LIB:=      Stratagene lung (#937210) CH:=
```

Figure 4 Alternative splice forms of the RAD1/REC1 gene are placed in the same cluster by d2_cluster, and the splice variants are separated into distinct subclusters by CRAW.

surely examples of where the multiple alignment algorithm and CRAW failed to identify existing sequence identities. Therefore, improvements to the multiple alignment would most likely result in an even lower upper bound.

To formulate bounds for the type I error (or the probability of not joining sequences that belong together), we perform an all versus all comparison of the cluster members with a Smith–Waterman algorithm. If one accepts that Smith–Waterman is an absolute method of identifying pairwise sequence overlap within defined constraints, then the true cases of type I error would be a subset of all intercluster similarities identified by Smith–Waterman. Fifty-one intercluster matches exist for the rat data set. Because type I errors are a subset of this, the type I error rate is <0.4% (=100 * 51/13,755). As in the type II error analysis, this bound could be sharpened if we were to inspect the 51 cases individually.

DISCUSSION

We have characterized d2_cluster and described the algorithm in terms that should be familiar to statisticians, computer scientists, and biologists alike. It has been shown that d2_cluster performs quite favorably to, and is consistent with, current EST clustering methods. In an empirical study based on >40,000 available rat EST and mRNA sequences, d2_cluster produced results that were between 83% and 90% identical with UniGene although d2_cluster created 10% fewer clusters and 20% fewer singleton clusters than UniGene. Three different measures of joining strength are used to compare the overall sensitivity of d2_cluster and UniGene, and these numbers are averaged to provide an estimate that d2_cluster joins sequences at a rate ~13% higher than UniGene. It remains undetermined, however, if this higher join rate results in more accurate index classes or is simply due to the joining of paralogous genes or other phenomena. Additionally, the absolute correctness of groups formed by d2_cluster has been quantified, and the sensitivity and selectivity are shown to be >99% (i.e., type I and type II error rates are bounded above by 1%).

d2_cluster has found application in the STACK project (Hide et al. 1997) in which ESTs are hierarchically clustered within tissue and arbitrary source categories. d2_cluster is set to join ESTs that are >96% identical over a window of 150 bases. More detail on STACK is given elsewhere (R.T. Miller, A.G. Christoffels, C. Gopalakrishnan, J. Burke, A.A. Ptitsyn, T.R. Broveak, and W.A. Hide, in prep.).

Unfortunately, space limitations prevent elaboration of the many other tools for sequence clustering that have been developed to cluster DNA sequence or to remove redundancies from sequence sets (Houlgatte et al. 1995; Parsons 1995; Grillo et al. 1996; Gill et al. 1997; Eckman et al. 1998; Yee and Conklin 1998; Pietu et al. 1999). Significant research has also been put into the grouping of protein sequence and the determination of domain boundaries (Sonnhammer and Kahn 1994; Worley et al. 1995; Sonnhammer et al. 1997; Gracy and Argos 1998a,b).

Sequence clustering is of little consequence in and of itself, and the prime motivation is to obtain biological knowledge. Effective sequence clustering is an organizing principle that serves as a starting point for discovery. For example, with all sequence information corresponding to a single gene in a cluster, features such as alternative exons and aberrant splicing, among others, can be modeled with greater ease. It is difficult to tune the parameters of primary sequence clustering such that features like splicing differences and even artifacts, such as chimerism, are accounted for while simultaneously generating proper index classes. Instead, decoupling this feature detection and artifact correction from the clustering step allows these problems to be handled in a more robust fashion. Hence, postprocessing steps, such as CRAW, are used to contrast gene variants and correct for artifact. Figure 4 shows how ESTs and mRNAs from a cluster of human sequences corresponding to human RAD1/REC1 cell-cycle control checkpoint protein are placed in the same cluster, whereas CRAW is used to separate distinct splice variants into separate subclusters. In a similar fashion, CRAW is also used to isolate and correct for chimeric sequence and other artifacts. Full details and additional examples are found in previous work (Burke et al. 1998; Chow and Burke 1999). Work has also been done to associate discovered multiple gene forms with sequence source information to infer state specificity or associate novel exon/UTR usage with disease (Burke et al. 1998; Gautheret et al. 1998). The d20 algorithm and others specified here are available at no cost for university researchers, and commercial licenses are available (details available from authors upon request). The commercial versions of CRAW and CRAWview (www.pangeasystems.com) were used in the preparation of this manuscript.

REFERENCES

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno et al. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634.

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* (Suppl.) **377:** 3–17.

Benson D.A., M.S. Boguski, D.J. Lipman, and J. Ostell. 1994. GenBank. *Nucleic Acids Res.* **22:** 3441–3444.

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

Boguski, M.S., T.M. Lowe, and C.M. Tolstohev. 1993. DbEST: Database for "expressed sequence tags." *Nat. Genet.* **4:** 332–333.

Burke, J.P., H. Wang, W. Hide, and D. Davison. 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* **8:** 276–290.

Chow, A. and J.P. Burke. 1999. CRAWview: For viewing splicing variation, gene families, and polymorphism in clusters of ESTs and full-length sequences. *Bioinformatics* **15:**(5) 376–381.

Eckman, B.A., J.S. Aaronson, J.A. Borkowski, W.J. Bailey, K.O. Elliston, A.R. Williamson, and R.A. Blevins. 1998. The Merck Gene Index Browser: An extensible data integration system for gene finding, gene characterization and EST data mining. *Bioinformatics* **14:** 2–13.

Gautheret, D., O. Poirot, F. Lopez, S. Audic, and J.M. Claverie. 1998. Alternative polyadenylation in human mRNAs: A large-scale analysis by EST clustering. *Genome Res.* **8:** 524–530.

Gill, R.W., T.C. Hodgman, C.B. Littler, M.D. Oxer, D.S. Montgomery, S. Taylor, and P. Sanseau. 1997. A new dynamic tool to perform assembly of expressed sequence tages (ESTs). *Comput. Appl. Biosci.* **13:** 453–457.

Gracy, J. and P. Argos. 1998a. Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics* **14:**(2) 164–173.

———. 1998b. Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics* **14:**(2) 174–187.

Grillo, G., M. Attimonelli, S. Liuini, and G. Pesole. 1996. CLEANUP: A fast computer program for removing redundancies from nucleotide sequence databases. *Comp. Appl. Biosci.* **12:** 1–8.

Hide, W., J. Burke, and D. Davison. 1994. Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J. Comp. Biol.* **1:** 199–215.

Hide, W., J. Burke, A. Christoffels, and R. Miller. 1997. A novel approach towards a comprehensive consensus representation of the expressed human genome. In *Genome informatics 1997* (ed. S. Miyano and T. Takagi), pp. 187–196. Universal Academy Press Inc., Tokyo, Japan.

Houlgatte, R., R. Mariage-Samson, S. Duprat, A. Tesslier, S. Bentolila, B. Lamy, and C. Auffray. 1995. The GenExpress Index: A resource for gene discovery and the genic map of the human genome. *Genome Res.* **5:** 272–304.

Hudson, T.J., L.D. Stein, S.S. Gerety, J. Ma, A.B. Castle, J. Silva, D.K. Slonim, R. Baptista, L. Kruglyak, S. Xu et al. 1995. An STS-based map of the human genome. *Science* **270:** 1945–1954.

Johnson, R.A. and D.W. Wichern. 1994. *Applied multivariate statistical analysis*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.

Matsubara, K. and K. Okubo. 1993. Identification of new genes by systematic analysis of cDNAs and database construction. *Curr. Opin. Biotech.* **4:** 672–677.

Miller, R.T., A.G. Christoffels, J. Burke, B.R. Karlak, A.A. Ptitsyn, T.R. Broveak, and W.A. Hide. 1999. A comprehensive approach to determination of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res.* (this issue).

Okubo, K., H. Hori, R. Matuba, T. Niiyama, and K. Matsubara. 1991. A novel system for large-scale sequencing of cDNA by PCR amplification. *DNA Seq.* **2:** 137–144.

Okubo, K., H. Hori, R. Matuba, T. Niiyama, A. Fukushima, Y. Kiojima, and K. Matsubara. 1992. Large-scale cDNA sequencing analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2:** 173–179.

Okubo, K., J. Yoshii, H. Yokouchi, M. Kameyama, and K. Matsubara. 1994. An expression profile of active genes in human colonic mucosa. *DNA Res.* **1:** 37–45.

Parsons, J.D. 1995. Improved tools for DNA comparison and clustering. *Comp. Appl. Biosci.* **11:** 603–613.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. In *Molecular evolution: Computer analysis of protein and nucleic acid sequences, methods in enzymology* (ed. R.F. Doolittle). Academic Press, San Diego, CA.

Pietu, G., R. Mariage-Samsom, N.A. Fayein, C. Matingou, E. Evenco, R. Houlgatte, C. Decraene, Y. Vandenbrouck, F. Tahi, M. Devegnes et al. 1999. The Genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome Res.* **9(2):** 195–209.

Schuler, G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.

Sonnhammer, E.L.L. and D. Kahn. 1994. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3:** 482–492.

Sonnhammer, E.L.L., S.R. Eddy, and R. Durbin. 1997. Pfam: A comprehensive database of protein families based on seed alignments. *Proteins* **28:** 405–420.

Sutton, G., O. White, M.D. Adams, and A.R. Kerlavage. 1995. TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1:** 9–18.

Torney, D.C., C. Burkes, D. Davidson, and K.M. Sirkin. 1990. *Computation of d2: A measure of sequence dissimilarity, computers and DNA, SFI studies in the sciences of complexity* (ed. G. Bell and T. Marr), vol. VII. Addison-Wesley, New York, NY.

Vasmatzis, G., M. Essand, U. Brinkmann, B. Lee, and I. Pastan. 1998. Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl. Acad. Sci.* **95:**(1) 300–304.

Venter, J.C. 1993. Identification of new human receptor and transporter genes by high throughput cDNA (EST) sequencing. *J. Pharm. Pharmacol.* (Suppl. 1) **45:** 355–360.

White, O. and A.R. Kerlavage. 1996. TDB: New databases for biological discovery. *Methods Enzymol.* **206:** 27–41.

Wilcox, A.S., A.S. Khan, J.A. Hopkins, and J.M. Sikela. 1991. Use of 3′ untranslated sequences of human cDNAs for rapid chromosomal assignment and conversion to STSs: Inplications for an expression map of the genome. *Nucleic Acids Res.* **19:** 1837–1843.

Worley, K.C., B.A. Wiese, and R. Smith. 1995. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity results. *Genome Res.* **5:** 173–184.

Wu, T.J., J.P. Burke, and D.B. Davison. 1997. A measure of DNA sequence dissimilarity based on mahalanobis distance between frequencies of words. *Biometrics* **53:** 1431–1439.

Yee, D.P. and D. Conklin. 1998. Automated clustering and assembly of large EST collections. *Intell. Syst. Mol. Biol.* **6:** 203–211.

Received June 16, 1999; accepted in revised form August 16, 1999.