GBuilder——An Application for the Visualization and Integration of EST Cluster Data

Juha Muilu, Patricia Rodriguez-Tomé and Alan Robinson

Genome Res. 2001 11: 179-184
Access the most recent version at doi:10.1101/gr.157501

Resource

GBuilder—An Application for the Visualization and Integration of EST Cluster Data

Juha Muilu,[1] Patricia Rodriguez-Tomé, and Alan Robinson

European Bioinformatics Institute, European Molecular Biology Laboratory Outstation–Hinxton, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

This paper presents a network-centric DNA sequence visualization and analysis tool called GBuilder. The tool is an easy-to-use Java application that can be used to analyze DNA sequence clusters and assemblies. The emphasis is on the analysis of EST data, where these highly redundant collections of low-quality and often alternatively spliced or chimeric sequence data are difficult to explore. The tool has the capacity to visualize similarities or dissimilarities between sequences at the level of the nucleotide base or annotation in many ways. Sequences may also be edited manually. The novel feature of GBuilder is its ability to access different data sources and analysis applications available on the Internet and to integrate these results and functionality back into itself. External resources such as EST cluster databases and conventional command-line analysis applications are integrated and accessed using CORBA (Common Object Request Broker Architecture), which provides a standard implementation independent protocol for integration. New CORBA services can be integrated immediately if they use a known interface described using the Interface Definition Language.

Expressed sequence tags (ESTs) are short (200 bp–500 bp) DNA sequences generated from the 3′ and 5′ ends of randomly selected cDNA clones. The purpose of EST sequencing is to scan rapidly for expressed genes and to provide a tag for each gene (Adams et al. 1991; Gerhold et al. 1996). Due to the fast sequencing process, the number of sequenced ESTs is increasing rapidly and tens of thousands of new sequences are submitted into public databases every month. (http://industry.ebi.ac.uk/~muilu/EST/Database/Monitor).

The EST databases contain a potential wealth of valuable information about expressed genes. For example, ESTs can be used to find genes on a genome; they provide information about different splicing events and polymorphisms (Burke et al. 1998; Gautheret et al. 1998; Picoult-Newberg et al. 1999). They also give an indication about the differing expression levels of genes in the different tissues from which they are derived (Schmitt et al. 1999; Tanabe et al. 1999).

The problems with collections of EST sequences that makes them difficult to use are that these data are often highly redundant, incomplete, and low in quality. Clustering and assembly methods can alleviate the problem of redundancy and improve the quality and length of consensus sequence; however, these techniques tend to lead to different views on the data, depending on the exact methods and criteria used (Miller et al. 1999). The informatics of these large and diverse data are not trivial. For the data to be useful to end-users, there is a need to visualize the data in many ways, as well as to access and merge information from different sources and analysis applications.

The interface to remote services can be implemented using CORBA (Common Object Request Broker Architecture; Siegel 2000), which provides standardized means to integrate computational resources on different machines over computer networks (e.g. the Internet or Intranets). Computer applications can access these resources if they implement the programming interface of a CORBA server, specified using the Interface Definition Language (IDL). The European Bioinformatics Institute (EBI) has been using CORBA since 1996 to resolve some of the current information technology problems in bioinformatics. We are not aware of any general-purpose visualization tool for EST data which also has the capability to integrate different data sets and analysis methods using CORBA. There are a number of useful programs that have been designed mainly for specific applications, such as CRAW (Chou and Burke 1999) for splice variant discovery, Consed for sequence finishing (Gordon et al. 1998), and ESTBlast (Gill et al. 1997) for on-the-fly sequence assembly.

RESULTS

Database Interface

External applications and databases are accessed from GBuilder through CORBA interfaces. Currently, interfaces to the EuroGeneIndex (Parsons and Rodriguez-Tome 2000), Radiation Hybrid (Rodriguez-Tome and Lijnzaad 2000), Radiation Hybrid Allocation (http://corba/RHdb/Rhalloc), and EMEST (http://industry.ebi.ac.uk/~muilu/EST/Database) databases are provided. There is also an interface for a simple server

[1]**Corresponding author.**
E-MAIL muilu@ebi.ac.uk; FAX 44-1-1223-494468.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.157501.

that may provide generic name-value information to GBuilder (http://industry.ebi.ac.uk/~muilu/GBuilder). GBuilder may also read name-value information from a Java properties file (Flanagan 1999) so that it is possible to use GBuilder without setting up database services. Table 1 shows the information each CORBA server can provide currently to GBuilder. Using the existing interfaces, it is straightforward for a programmer to add new databases and property information servers to GBuilder by specifying the type and location of the service in a configuration file (URL of the IOR). GBuilder provides a database browser, which allows navigation of different CORBA-compliant sites and databases.

External Applications

The external analysis applications are made CORBA-compliant using the AppLab application (Senger 1999), which provides a CORBA server for executing conventional command line applications and sending the results back to the client. Command line parameters and input and output data are described using a GCG-compliant (http://www.gcg.com/products/wispackage.html) description language, which is then used to generate meta-data that the clients can use for building a user interface to the program. An example of a configuration file is shown in Box 1. GBuilder uses a ready-made client-side graphical user interface provided by the AppLab package for setting command line parameters, as well as starting and cancelling program execution. An important feature of this generic interface is that it is independent of server-side applications and, thus, there is no need to modify the client program if new applications are installed on the server. Information about CORBA resources available to GBuilder is read from a configuration file, which is specified as a URL to GBuilder. Available applications are found by making a query to the server and, thus, the list of available applications is always up-to-date. At press time, the following programs are available to GBuilder using the AppLab server located at the EBI: CAP2 (Huang 1996) and CAP3 (Huang and Madan 1999) multiple alignment programs, CLEANUP (Grillo et al. 1996) for removing redundant sequences, and NCBI's DUST for masking out low-complexity regions.

GBuilder can import sequences in FASTA and ACE format as produced by CAP3. Sequences may be exported and stored as a FASTA file or as serialized Java objects. The latter allows the storage of the complete state of the sequence objects, including any annotations. Optional positional information about assembled sequences may be stored in the FASTA header after the accession number. This format is chosen because of its simplicity and the lack of an alternative at the time GBuilder was developed. A future release of GBuilder will utilize the Biomolecular Sequence Analysis specification (BSA), adopted by the Object Management Group (OMG; http://www.omg.org/cgi-bin/doc?lifesci/99–12–01.pdf). The BSA specification provides standardized descriptions for biological objects, as well as for sequence-analysis services.

Visualization Features

The visualization capabilities of GBuilder fall into two categories: Visualization of similarities between sequences at the nucleotide level, and at the annotation level. In the former case, the nucleotide similarities can be highlighted between adjacent sequences or between selected sequences, as shown in Figure 1. To make the navigation and detection of problematic regions easier on a large assembly, a user may open another window that is linked to the main window by a movable lens. The lens represents the area shown in the new window and can be zoomed in to display sequences at single-nucleotide resolution while maintaining the global view in the main window.

It is also possible to visualize common substrings between two sequences, which is useful for finding alternative-spliced ESTs as shown in Figure 2. The substrings are found by using Java implementation (http://www.accessone.com/~lorre/pages/boyermoore.html) of the Boyer-Moore algorithm (Gusfield 1997), which is applied over the whole sequence using nonoverlapping subsequent substrings taken from the first sequence. This brute-force algorithm works reasonably well with sequences of the length of typical ESTs. It is possible to increase and decrease the length of the substring to get the required level of accuracy and also increase execution speed.

Besides using the substrings, it is possible to visualize how sequences are linked together in a clustering process or visualize the pairwise alignments used in the clustering if this information is stored in a database, as they are for the EST EuroGeneIndex cluster databases (Parsons and Rodriguez-Tome 2000). An example of this is shown in Figure 3. This feature is useful in cases in which an EST super cluster is broken into a number of subclusters during a multiple-sequence alignment. This often happens because of either alternative splicing (Burke et al. 1998; Thanaraj 1999) or low sequence quality (Aaronson et al. 1996). Locating the linking sequences between the subclusters and then visualizing the pairwise alignment or the substring similarity may help to identify the reasons behind the subclusters.

Annotations can be visualized using color codes, sorting sequences according to a property, or drawing a line between sequences sharing common annotation. Text annotations may be displayed alongside the sequence. Visualizing the annotation for ESTs available from the CORBA servers allows users to quickly find out, for example, which ESTs are derived from the

Box 1. Application Configuration File for the CAP3 Program

```
Launcher cap3launcher

Main {


   File
      id = input_sequences
      prompt = Input sequences in FastA format
   End

   File
     id = fasta_out
     filetype = output
     display = -
     answer = no
   End

   File
     id = cap3_out
     filetype = stdout
     display = -
   End

}

Optional {
   Range
     qualifier = a
     scalemin  = 11
     default   = 20
     prompt    = Band expansion size
   End

 …..

}
```

The first line defines the name of the script used to execute CAP3. The input data and output data are described in the "Main" section. The input is a FASTA file containing the sequences. The program can product two outputs: The first is the sequence assembly (results of CAP3 are converted into fasta format by the launcher script) and the second is the standard output of the program. All command line parameters are optional and are given in an "Options" section. In the configuration file it is possible to specify, for example, default values and the value range.

same clone, their putative function, whether they are allocated for radiation hybrid mapping, or how the sequences are clustered in other systems such as UniGene (Schuler 1997). The comparison between other cluster databases or functional and sequence-quality annotation is valuable, for example, in confirming the splice variants. Color coding and sorting sequences by clone library is useful in finding sequence variations between libraries or to highlight tissue-specific expression of a putative gene represented by a cluster.

In addition to the visualization features, GBuilder can operate on sequences in a number of ways. Because EST sequences are low in quality and often the errors cannot be fixed automatically, manual editing is necessary to fix problems and improve the quality of the assembly and consensus sequence. Thus, in GBuilder, it is possible to edit bases in a simple editor, move sequences in an assembly, delete sequences, or make reverse-complements of sequences.

DISCUSSION

Using the GBuilder visualization tool, with integrated access to different data sources and analysis applications, has proven useful in exploring complex sequence data, such as collections of EST sequences. The tool helps to validate the rationales behind EST clustering and also allows the identification of alternatively spliced gene products, as well as artifacts in sequences. Combining the visualization features with the ability to access different databases and analysis applications brings a new synergy for dynamic resources available to users over the Internet.

This integration is made possible using CORBA, which provides a standard and an implementation independent layer between clients and services. Using known program interface specifications based on IDL, it is possible to publish new services to users automatically and transparently, which may then be integrated into tools such as GBuilder. These services may be located on different sites on the Internet or the internal Intranets.

A problem with the network-centric approach is the vulnerability of network connections. This can be overcome by installing critical services locally. If more services become available on different sites over the Internet, the increased volume can overcome the possible low quality of connections.

Future Directions

Work is underway to make the GBuilder compliant with the Biomolecular sequence analysis specification (http://industry.ebi.ac.uk/openBSA), which provides definitions for basic biological objects and analysis mechanisms. The OMG specification is the result of collaboration between industrial and academic organizations and provides a solid framework for extending the integration between different commercial and/or noncommercial analysis resources.

The goal is to add, for example, new analysis applications and new databases such as STACK (Miller et al. 1999), which is a good resource for alternative gene form discovery (Burke et al. 1998), and UniGene (Boguski and Schuler 1995).

METHODS

GBuilder is written in Java JDK1.1, which allows it to run on multiple platforms as a standalone application or as an applet within Internet browsers. The CORBA functionality is independent of the ORB software used; GBuilder has been tested with OrbixWeb (http://www.iona.com/products/orbixweb), ORBacus (http://www.ooc.com/ob), and JavaORB (http://www.multimania.com/dogweb/). The IDLs used by GBuilder

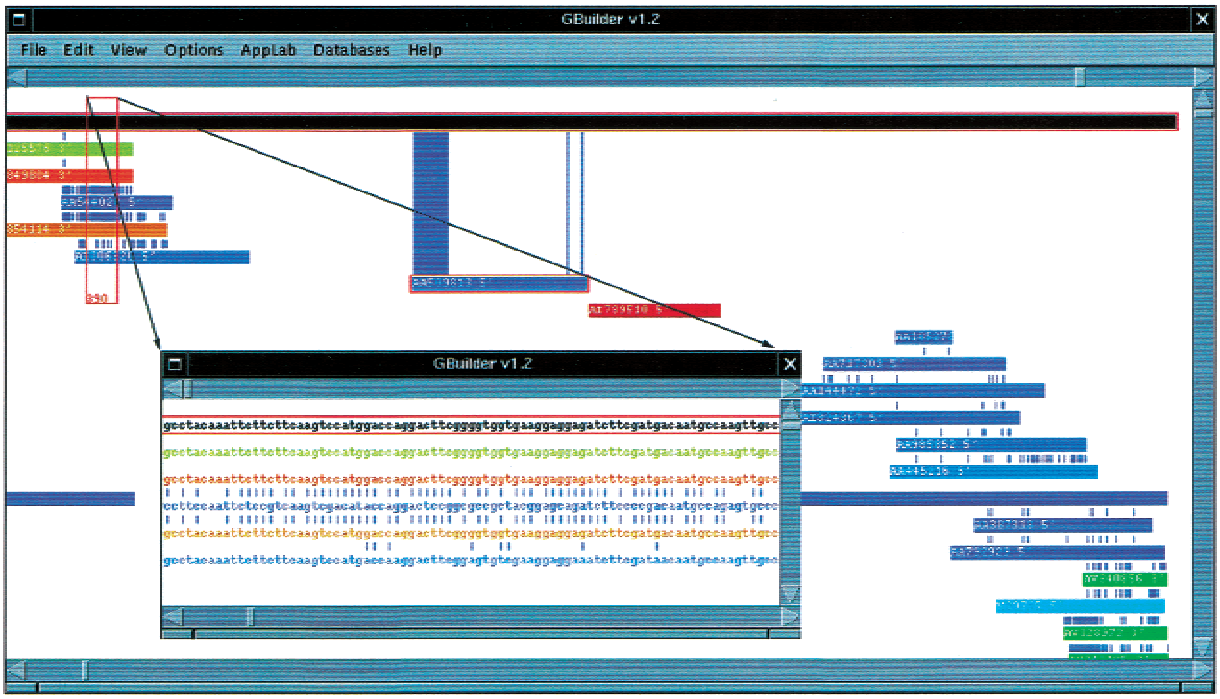

Figure 1 EuroGeneIndex super cluster from the mouse EST/mRNA database (accession no. U10115). Dissimilarities between adjacent and selected sequences are highlighted. A portion from the Main window, as indicated by the rectangle, can be shown in the other window. Sequences are color coded by clone library.

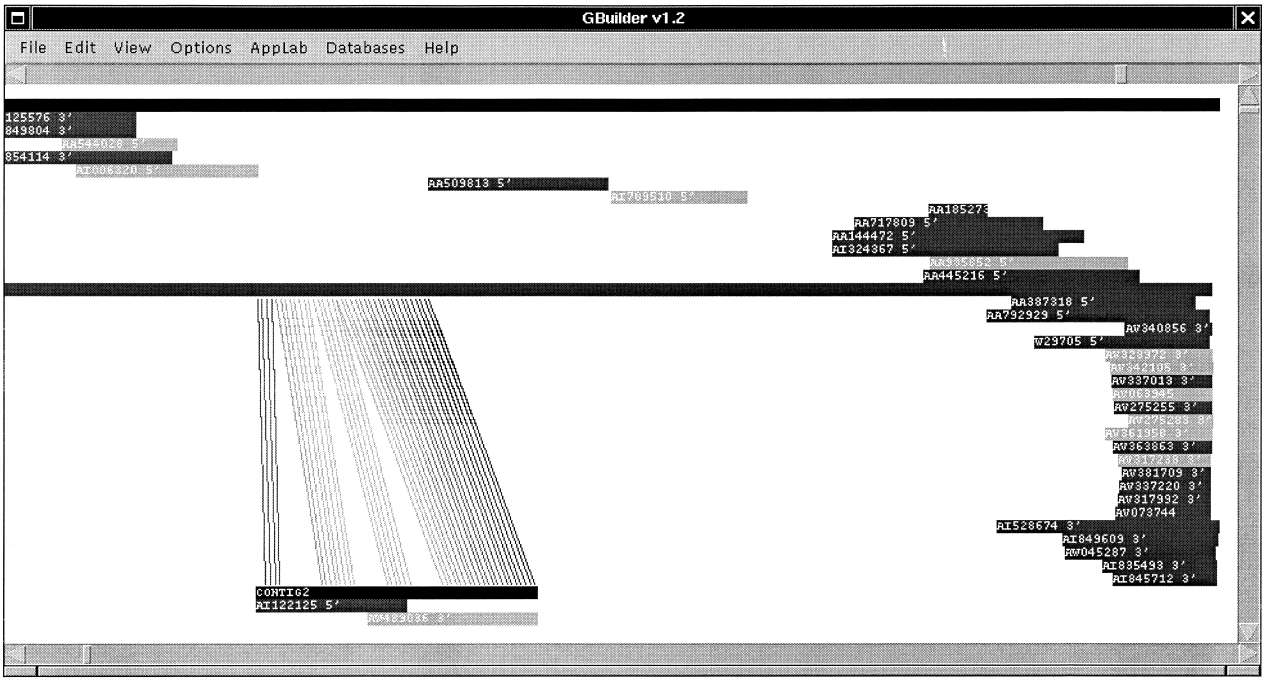

Figure 2 The substring similarity between two subclusters (U10115 super cluster) is shown. The lines' colors are based on their position along the sequence. Sequences are color coded by the UniGene cluster name. Brown sequences are in UniGene cluster (Mm.3400) and green sequences are not in UniGene. The sequences may have been excluded because they are marked as low-quality sequences in dbEST (the quality annotation is shown in Fig. 3).

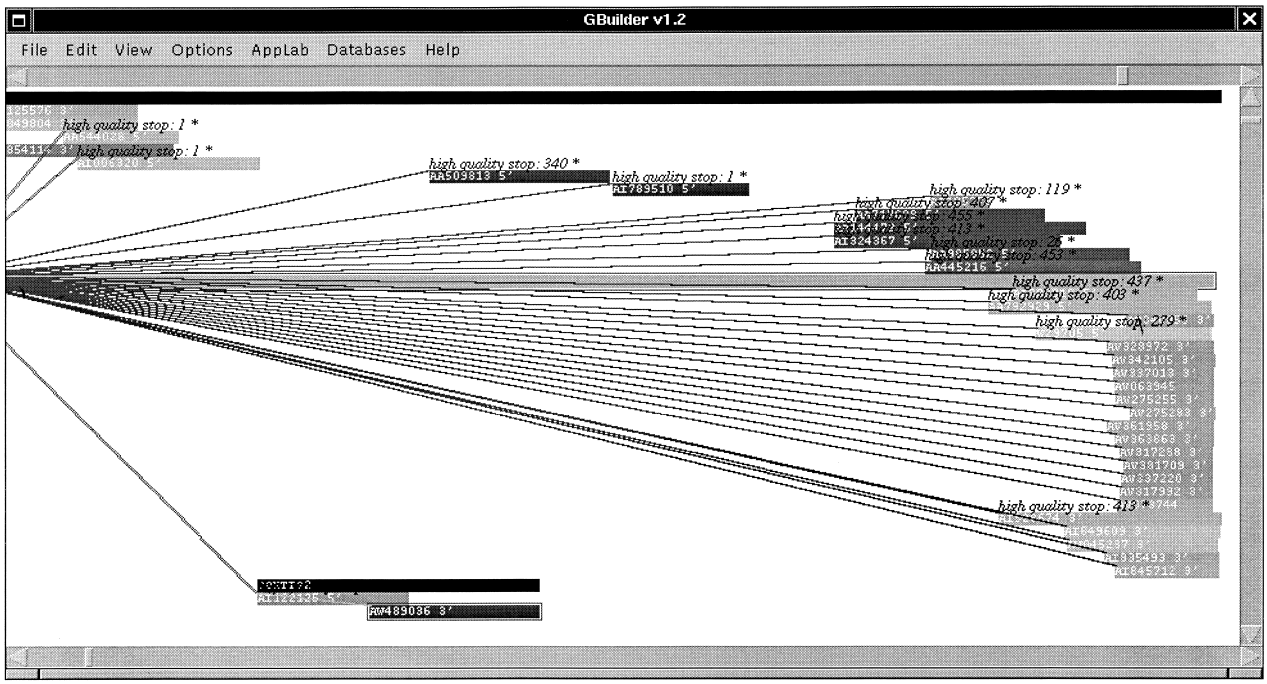

Figure 3 Sequences that are linked with the longest sequence (U10115 mouse EuroGeneIndex super cluster) are connected by drawing a line between them. Sequences are colored by the number of links they have to other sequences. Sequence quality annotation (measured last high quality base) is shown on top of the sequence.

are available from the EBI Web site (http://industry.ebi.ac.uk/~muilu/GBuilder).

Availability

The program and source code is available free of charge from the Web site. The program is licensed under an open source license.

ACKNOWLEDGMENTS

We thank Matteo di Tommaso for starting the project; Adam Butler, Tom Flores, Tom Freeman, Peter Van Heusden, Jeremy Parsons, Martin Senger, Mike Starkey, Alphonse Thanaraj, and Yagnesh Umrania for their valuable discussions; and Doyle Myers for providing the Java implementation of the Boyer-Moore algorithm. The GBuilder program is developed in the Industry Programme of the European Bioinformatics Institute.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S., and Elliston, K.O. 1996. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res* **6:** 829–845.

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

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

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

Flanagan, D. 1999. *Java in a nutshell: A desktop quick reference.* O'Reilly, Sebastopol, CA.

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

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

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

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

Gusfiled, D. 1997. *Algorithms on strings, trees, and sequences*, 1st ed. Cambridge University Press, UK.

Huang, X. 1996. An improved sequence assembly program. *Genomics* **33:** 21–31.

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

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R., and Hide, W.A. 1999. A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res.* **9:** 1143–1155.

Parsons, J.D. and Rodriguez-Tome, P. 2000. JESAM: CORBA software components to create and publish EST alignments and clusters. *Bioinformatics* **16:** 313–325.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., and Boyce-Jacino, M. 1999. Mining SNPs from EST databases. *Genome Res.* **9:** 167–174.

Rodriguez-Tome, P. and Lijnzaad, P. 2000. RHdb: The radiation hybrid database. *Nucleic Acids Res.* **28:** 146–147.

Schmitt, A.O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C.P., Hinzmann, B., and Rosenthal, A. 1999. Exhaustive mining of EST

libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27:** 4251–4260.
Schuler, G.D. 1997. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75:** 694-–698.
Senger, M. 1999. AppLab CORBA-Java based Application Wrapper. *CCP11 Newsl*, Issue 8. http://www.hgmp.mrc.ac.uk/CCP11
Siegel, J. 2000. *CORBA 3 fundamentals and programming*. John Wiley and Sons, New York.
Tanabe, K., Nakagomi, S., Kiryu-Seo, S., Namikawa, K., Imai, Y., Ochi, T., Tohyama, M., and Kiyama, H. 1999. Expressed-sequence-tag approach to identify differentially expressed genes following peripheral nerve axotomy. *Brain Res. Mol. Brain Res.* **64:** 34–40.
Thanaraj, T.A. 1999. A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures. *Nucleic Acids Res.* **27:** 2627–2637.

WWW RESOURCES

http://corba/RHdb/Rhalloc
http://industry.ebi.ac.uk/~muilu/EST/Database
http://industry.ebi.ac.uk/~muilu/EST/Database/Monitor
http://industry.ebi.ac.uk/~muilu/EST/Gbuilder
http://industry.ebi.ac.uk/openBSA
http://www.accessone.com/~lorre/pages/boyermoore.html
http://www.gcg.com/products/wis-package.html
http://www.iona.com/products/orbixweb
http://www.multimania.com/dogweb/
http://www.omg.org/cgi-bin/doc?lifesci/99–12–01.pdf
http://www.ooc.com/ob

Received July 31, 2000; accepted in revised form October 27, 2000.