

Integration of data management and analysis for genome research

Volker Brendel

Department of Zoology & Genetics and Department of Statistics
Iowa State University
2112 Molecular Biology Building
Ames, Iowa 50011-3260
U.S.A.
vbrendel@iastate.edu

Abstract: Technological advances in genome research have produced unprecedented volumes of genetic and molecular data that now provide the context for any biological research. However, data access, curation, and analysis have remained challenging areas for continued research and development and often prove to be the bottleneck for scientific progress.

Many a paper in bioinformatics or even in general molecular biology these days start out just like the abstract above, with an acknowledgement of the explosive growth of molecular sequence, structure, and expression data. What fit nicely within the printed pages of a thin booklet only 25 years ago now comprises large and increasingly complex databases that are Web-accessible to the public. Figure 1 shows the growth on one major molecular sequence repository - GenBank, maintained at the U.S. National Center of Biotechnology Information (NCBI). The slope of the curve is indeed impressive. However, the actual size of the data sets would seem to be easily dwarfed by database sizes in other commercial, governmental, or even other research fields. What then is the real problem, if any, facing the biology community?

I think there are many aspects for consideration. One important facet is that the molecular databases themselves have evolved over the years, and surely many details of database design should have been done differently in hindsight. However, the rapid pace of new data acquisition has so far prevented any major re-design and re-construction of the databases the community is accustomed to. Another critical point is that the data derive from a large variety of sources and are intrinsically heterogenous. There are no uniform standards for data quality and annotation. In these notes I shall not further discuss the challenges faced by the large database providers, but rather I shall review the problem first from the point of a user and then suggest some approaches we have pursued to provide intermediate solutions.

Growth of GenBank

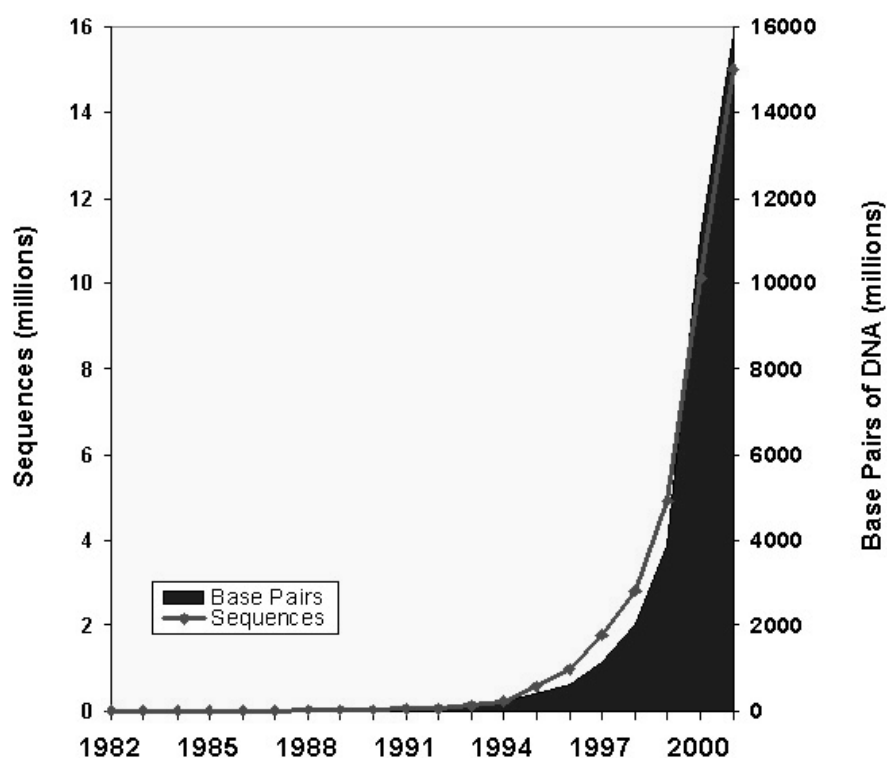


Figure 1: Molecular Database Growth

(from www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

1 Some Comments on the Current Status of Molecular Databases

1.1 Existing Data Retrieval Systems

Two of the centralized, prominent Web access points to existing sequence and molecular biology data are the Entrez system at NCBI (www.ncbi.nlm.nih.gov/Entrez/) and the SRS system developed in Europe (www.embl-heidelberg.de:80/srs5/). Both systems are limited in three ways: 1) data are provided in flat file formats only, often requiring much user manipulation after receipt; 2) retrievals are inflexible, and do not


```

LOCUS      ATKIN2      880 bp      DNA      PLN      23-JUL-1992
CDS        join(104..160,320..390,504..579)

EST Accession 3450035:

Exon 1     78     160 ( 83 n); cDNA      1     80 ( 80 n); score: 0.867
Intron 1   161     321 ( 161 n); Pd: 0.976 (s: 0.90), Pa: 0.972 (s: 1.00)
Exon 2    322     390 ( 69 n); cDNA     81    149 ( 69 n); score: 0.971
Intron 2   391     504 ( 114 n); Pd: 0.999 (s: 0.96), Pa: 0.964 (s: 0.98)
Exon 3    505     785 ( 281 n); cDNA    150   429 ( 280 n); score: 0.996

Alignment (genomic DNA sequence = upper lines):

TTTTACAAGA AAAAAATATC TGAAAAATGT CAGAGACCAA CAAGAATGCC TTCCAAGCCG      137
||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
TTTTCCAAGG AAA-AAATTC TGAAAAAT-T CNGGGACC-A CNAGAATGCC TTCCAAGCCC      57

GTCAGGCCGC TGGCAAAGCT GAGGTACTCT TTCTCTCTTA GAACAGAGTA CTGATAGATT      197
||||| ||||| ||||| ||||| |||||
GTCAGGCCGC TGGCCAAGCT GAG..... .. 80

/////

ATAGGAGAAG AGCAATGTTT TGCTGGACAA GGCCAAGGAT GCTGCTGCTG CAGCTGGAGC      377
||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
...GAGAAG AGCAATGTTT TGCTGGACAA GGCCAAGGAT GCTGCTGCTG CAGCTGGAGN      136

TTCCGCGCAA CAGGTAAACG ATCTATACAC ACATTATGAC ATTTATGTAA AGAATGAAAA      437
||||| ||||| |||||
TTCCGCNCAA CAG..... .. 149

/////

GTTATAGGCG GGAAAGAGTA TATCGGATGC GGCAGTGGGA GGTGTTAACT TCGTGAAGGA      557
||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
.....GCG GGAAAGAGTA TATCGGATGC GGCAGTGGGA GGTGTTAAC- TCGTGAAGGA      201

CAAGACCGGC CTGAACAAGT AGCGATCCGA GTCAACTTTG GGAGTTATAA TTCCCTTTT      617
||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
CAAGACCGGC CTGAACAAGT AGCGATCCGA GTCAACTTTG GGAGTTATAA TTCCCTTTT      261

/////

>Pcorrect (gi|399298|sp|P31169|KIN2_ARATH)
MSETNKNAFQ AGQAAGKAEE KSNVLLDKAK DAAAAAGASA QQAGKSISDA AVGGVNFVKD KTGLNK

>Pfalse (gi|16354|emb|CAA44171.1)
MSETNKNAFQ AGQAAGKAER RRAMFCWTRP RMLLLQLELP RNRAGKSISD AAVGGVNFVK DKTGLNK

```

Figure 2: Example of an erroneous GenBank annotation. The GenBank CDS gives incorrect assignment of both acceptor sites (319 should be 321, 503 should be 504), as pointed out by Korning et al. [KHRB96]. Spliced alignment with an *Arabidopsis* EST by the GeneSeqer program [UB00] proves the correct assignment (identities between the genomic DNA, upper lines, and EST, lower lines, are indicated by |; positions of the rightmost residues in each sequence block are given on the right; introns are indicated by ... for brevity, some sequence segments are replaced by /////). The erroneous intron assignment led to an incorrect protein sequence prediction (Pfalse). Both the incorrect sequence and the correct protein sequence (Pcorrect) persist in the NCBI non-redundant protein database under different accessions.

significant similarity to cyanobacterial proteins; and (4) determine possible roles for the identified proteins in biochemical pathways associated with the chloroplast. The intersection of sequence sets identified by (2)-(4) should provide a highly reliable set of definite chloroplast proteins. Novel methods for signal peptide identification can then be developed with this set as the positive training set. This approach is easily stated but its execution currently involves extensive programming and scripting.

```

Query= gi|539663|pir|A42701 PR264/SC35 protein -human (221 letters)
Database: nr 511,898 sequences; 160,474,304 total letters

Sequences producing significant alignments:

Score      E
(bits)    Value

gi|3929383|sp|Q62093|SFR2_MOUSE SPLICING FACTOR, ARGININE/SERINE... 185 2e-46
gi|266991|sp|P30352|SFR2_CHICK SPLICING FACTOR, ARGININE/SERINE-... 185 2e-46
gi|6755478|ref|NP_035488.1|| splicing factor, arginine/serine-ri... 185 2e-46
gi|4506899|ref|NP_003007.1|| splicing factor, arginine/serine-ri... 182 2e-45
gi|7243688|gb|AAF43415.1|AF232775_1 (AF232775) SR family splicin... 140 1e-32
gi|423485|pir||A46241 interferon response element-binding factor... 124 6e-28
gi|3892187|gb|AAC78303.1| (AF064592) RNA-binding protein [Schist... 117 8e-26
gi|3929375|sp|Q09511|SFR2_CAEL PUTATIVE SPLICING FACTOR, ARGINI... 116 1e-25
gi|7446336|pir||T09704 probable arginine/serine-rich splicing fa... 85 4e-16

>gi|3929383|sp|Q62093|SFR2_MOUSE SPLICING FACTOR, ARGININE/SERINE-RICH 2
(SC-35) (SPLICING COMPONENT, 35 KD) (PR264 PROTEIN)
>gi|1405747|emb|CAA67134.1| (X98511) PR264/SC35 [Mus musculus]

>gi|266991|sp|P30352|SFR2_CHICK SPLICING FACTOR, ARGININE/SERINE-RICH 2
(SC-35) (SPLICING COMPONENT, 35 KD) (PR264 PROTEIN)
>gi|539509|pir||B42701 PR264 protein - chicken
>gi|63752|emb|CAA44306.1| (X62446) PR 264 [Gallus gallus]
>gi|228503|prf||1805195A RNA-binding protein PR264 [Gallus gallus]

>gi|6755478|ref|NP_035488.1|| splicing factor, arginine/serine-rich 2 (SC-35)
>gi|539663|pir||A42701 PR264/SC35 protein - human
>gi|35597|emb|CAA44307.1| (X62447) PR 264 [Homo sapiens]
>gi|455419|emb|CAA53383.1| (X75755) PR264/SC35 [Homo sapiens]
>gi|3335676|gb|AAC71000.1| (AF077858) SC35 [Mus musculus]
>gi|228504|prf||1805195B RNA-binding protein PR264 [Homo sapiens]

>gi|4506899|ref|NP_003007.1|| splicing factor, arginine/serine-rich 2
>gi|266992|sp|Q01130|SFR2_HUMAN SPLICING FACTOR,
ARGININE/SERINE-RICH 2 (SPLICING FACTOR SC35) (SC-35)
(SPLICING COMPONENT, 35 KD) (PR264 PROTEIN)
>gi|420095|pir||A42634 splicing factor SC35 - human
>gi|337926|gb|AAA60306.1| (M90104) splicing factor [Homo sapiens]

```

Figure 3: (Partial) NCBI BLASTP output of a default query with a human SC35 protein sequence. Resolution of the query result into a non-redundant set of SC35-homologs would require much additional work of sequence and annotation comparisons.

1.3.3 Conserved Proteins Between Plants and Fungi but not Animals

Similar to the previous problem, in this example the query involves initially the intersection of two sequence sets (plant and fungal protein sequences). Subsequently, we are interested in the complement of the result of the first intersection intersected with a third set (animal

2 An *Arabidopsis thaliana* Genome Database and Web-Workbench

A common theme across the examples given above is the need to work with a subset or "extract" of data relevant to a particular research question. From many researchers' experience, derivation of such extract can be one of the most time-consuming parts of a project. In our own work with the *Arabidopsis* genome we found data access and scope in the existing databases (TAIR, www.arabidopsis.org, MATDB, mips.gsf.de/proj/thal) too limited. Figure 4 displays one of the entry pages into our local *Arabidopsis* database (AtGDB, zmdb.iastate.edu/PlantGDB/AtGDB), developed on MySQL (www.mysql.org). The page illustrates one of the principles of molecular database interface design, which is to provide intuitive graphics to access the data in addition to command-line type access. In this case, the user can select a chromosomal segment of interest either by clicking on the graphic or by typing numerical coordinates in the toolbar fields.

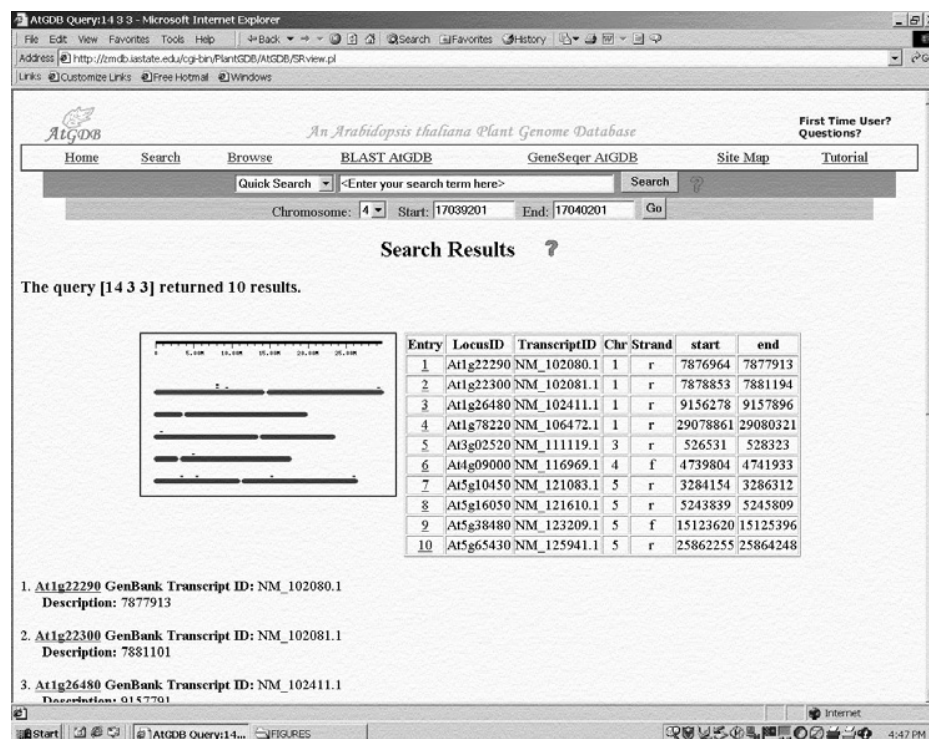


Figure 5: AtGDB - Query results. "14 3 3" refers to a particular gene family. All annotated genes from this family are shown with GenBank gene model code and respective genome location.

Figure 5 shows the results of an alternative access point by querying for terms in the gene definition lines. The output lists all the locations of genes matching the search term,

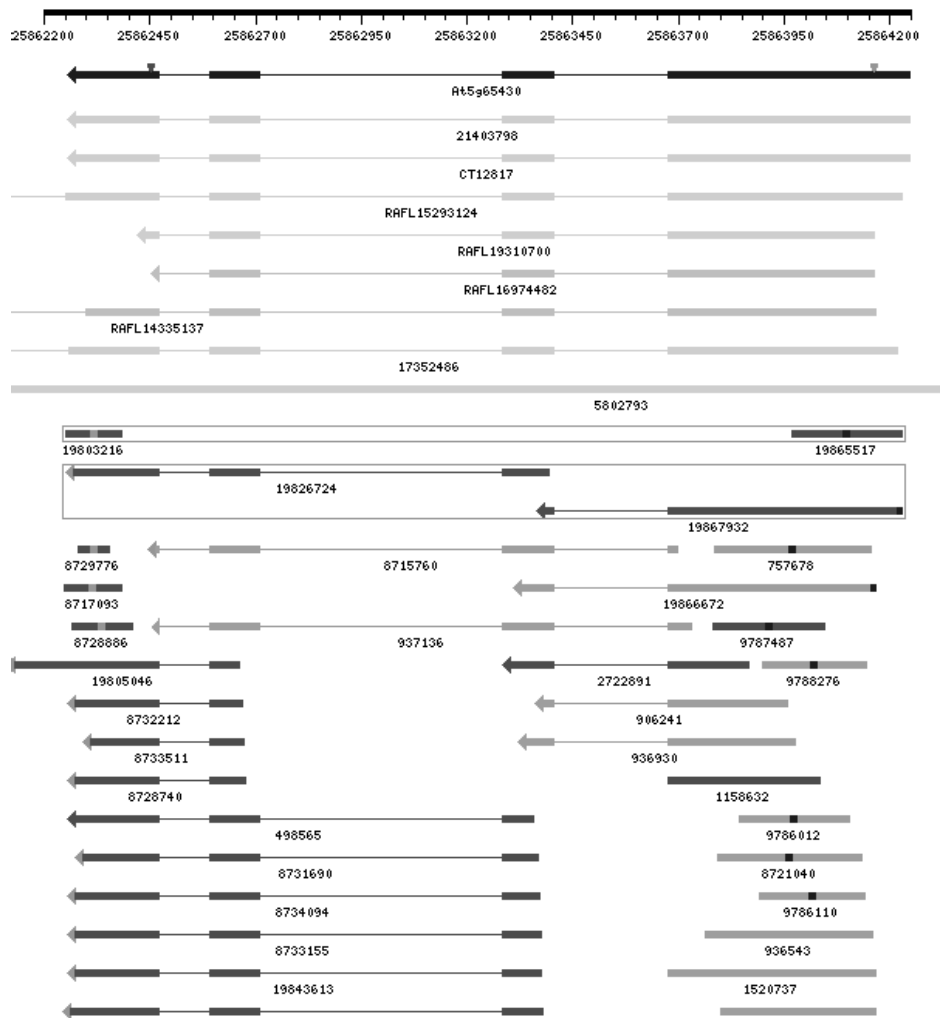


Figure 6: AtGDB - Genome context view for a particular gene selected from the search results in Figure 5. The current gene model (GenBank annotation) is shown on the top in dark blue. Exons are shown as solid boxes, introns as lines. cDNA (light blue and gray) and EST (red and pink) spliced alignments are shown below (the darker colors indicate cognate locations, whereas the lighter colors indicate that the respective sequence have a better match elsewhere in the genome). The green box associates ESTs experimentally known to derive from the same gene.

and the user has again the choice to hone in on a particular gene by either clicking on the graphic or selecting the gene from the table. The listing below provides links to the

GenBank repository.

The core of our database is shown in Figure 6. This schematic summarizes spliced alignment results of the type shown in Figure 2 using all available cDNAs and ESTs matching the selected locus. The display is drawn dynamically from pre-computed spliced alignment coordinates stored in the database. Clicking on a particular sequence will show the record for that sequence (Figure 7) as well as provide links to insert this sequence directly into other tools, e.g. BLAST (Figure 8).

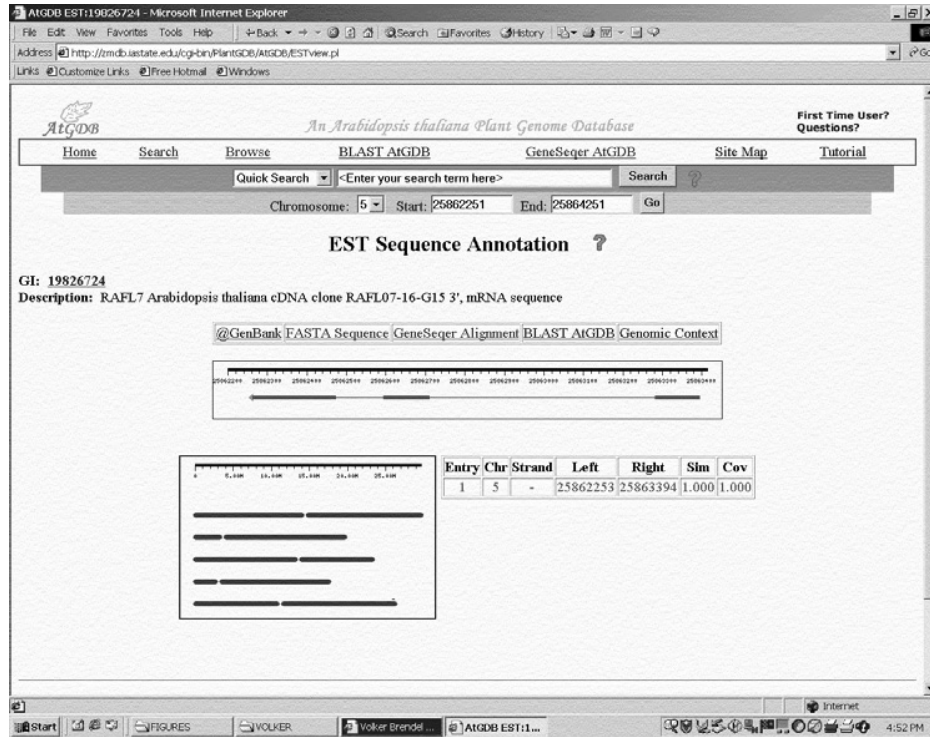


Figure 7: AtGDB - Detail for a particular EST selected from the spliced alignment display in Figure 6. Alignment details and links to other analytical tools are available via buttons.

A detailed description of biological background and questions addressed with the database and its interface are beyond the scope of this discussion. What I hope to convey are a number of design principles that in our view are critical for providing the best possible access to the rich resource of genomic sequence data. One key element is to parse the analytical output of standard research applications on the genome sequences also into the database, in addition to the raw sequence data and annotation. In this way, the full query capabilities of the database software can be applied to the results to quickly provide genome-wide views of the data. For example, it is now trivial to pull out a list of all gene models supported by

full-length cDNA evidence, or to view all matching locations of a particular gene probe, or to select all duplicated gene pairs with different exon numbers, and so forth. A second point is to link the analytical tools directly to the displayed data so that all results can be reproduced by the user, possibly using different parameter choices or additional input data. In this way, for example, the biologist who is expert on a particular subset of genes is empowered to easily check the annotation provide in the database, without the awkward steps of having to download the sequence data, format the data correctly for input into local analytical programs, and then relating the results back to the database source. Ultimately, it would be helpful to design interfaces that will allow expert curation of the database via the Web. It is likely that in another 25 years hence today's achievements will look as insignificant as the small printed collections of sequences a quarter of a century ago.

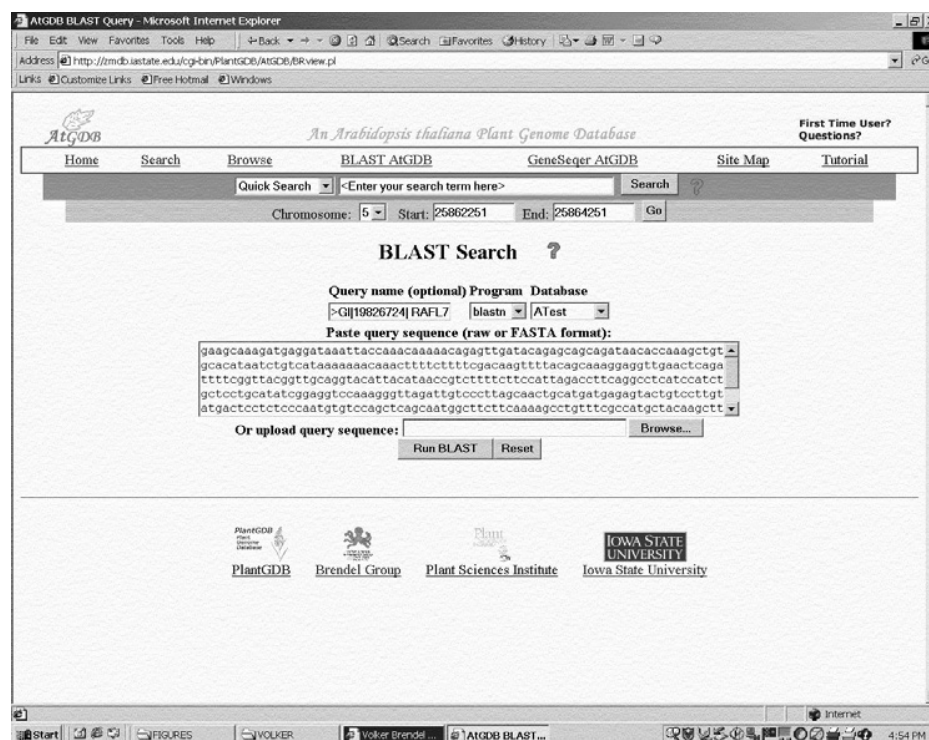


Figure 8: AtGDB - Integrated analytical tools. The EST sequence selected in Figure 7 is pasted into a text input window for a BLAST search against other sequence selections.

I would like to acknowledge the students and staff in my research group at Iowa State University who has friends and collaborators contribute greatly to these emerging ideas and implementations.