Identify Legumes Specific Genes
DREU 2010 Final Report

Yingxu Liu
Computer Science
Winona State University
Winona, MN

Mentor: Jessica Schlueter
Bioinformatics and Genomics
University of North Carolina at Charlotte
Charlotte, NC

Mentor: Shannon D Schlueter
Bioinformatics and Genomics
University of North Carolina at Charlotte
Charlotte, NC

Abstract

The Fabaceae has been the target of many genomic studies as the third largest plant family. We used Vmatch software to compare unigene sets from Medicago truncatula, Glycine max and Lotus japonicus to GenBank's nonredundant and EST databases. This process was part of a new pipeline to identify genes, which unique to all legumes with respect to all other know sequences.

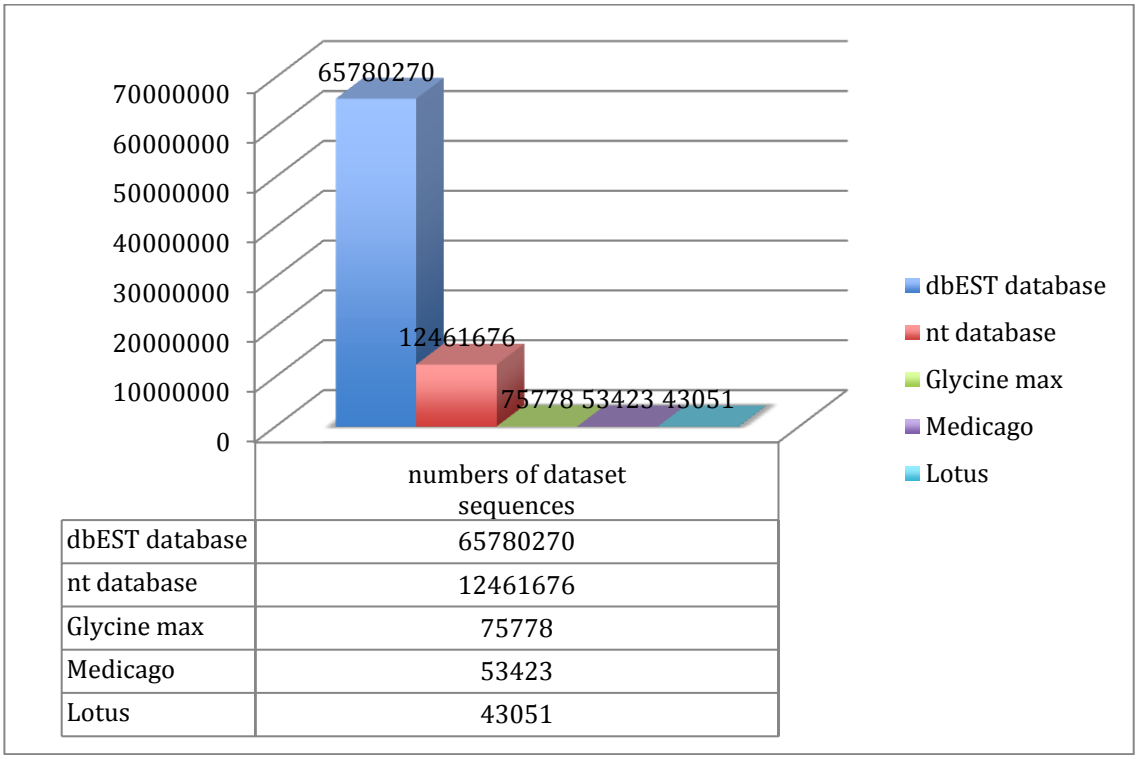
1. Introduction

Legumes play an essential role in human health. They are significant source of protein, oil and mineral nutrients. Legumes are common in daily life, like pea, common bean, lentil, soybean, and other food legumes. They serve as dietary protein for people, are used as fodder for animal feed as well as industrial uses. Further more, legumes produce a number of important compounds called phytochemicals such as isoflavonoids that have huge impact on human health [1].

Legumes are unique among other plants in that they form a symbiotic relationship with soil microbes through the formation of nitrogen-fixing nodules. These nodules form via interaction with rhizobia to obtain nutrients and are an efficient way to acquisition minerals from the soil.

"Legumes are the third-largest family of flowering plants, constituting about 8% of all angiosperms" [2]. The legumes encompass nearly 20,000 species and 700 genera. Of which there are 3 fully sequenced species, Medicago truncatula, soybean (Glycine max), and Lotus japonicus. In addition, there are extensive expressed sequence resources across a variety of species. A legume specific genes analysis in 2004, they used nearly 700,000 nucleotide sequences in dbEST. However, currently there are nearly 65,780,270 nucleotide sequences include in dbEST, and 12,461,676

sequences belong to nt from GenBank. Building upon 2004 analysis that used a small set of ESTs, we chose to reconsider this study using huge datasets of ESTs. Further more, developing a new pipeline that will allow us to identify genes that are unique to all legumes with respect to all other know sequences.



	numbers of dataset sequences
dbEST database	65780270
nt database	12461676
Glycine max	75778
Medicago	53423
Lotus	43051

Figure1. Number of all sequence datasets.

2. Methods

Installed database. Database was used for high efficient analyze legume specific genes, so installed mySQL database.

Downloaded dbEST file. Through National Center for Biotechnology Information (NCBI) ftp server (ftp://ftp.ncbi.nih.gov/genbank/), there were dbEST zipped files from gbest1.seq.gz to gbest413.seq.gz. All the dbEST files were from that dataset available 6/18/10. Perl scripts were used to capture unique locus and taxon number per file then stored to database. Backed up all dbEST sequence information to database was for comparative analysis of expressed sequences.

Due to 413 dbEST files totally, concatenated files to a bigger one was efficient for sequence matching. First 30 files were each 5 concatenated together and the last were 10 as a group. Therefore there were 44 sequence files in the end.

Prepared legumes gene model. There are three gene models sets for legumes, Glycine max, Lotus japonicus, and Medicago truncatula. Those gene models are from genomic sequences, they are abinitio and sequence based methods. At the beginning, only used Glycine max gene model to Vmatch against dbEST file.

Downloaded Glycine max gene model from http://www.phytozome.net/soybean website.

Perl script were used for capture Glyma number among all the sequence, and stored into database.

Set up Vmatch software. Since large scales sequences matching tasks need to be done, Vmatch was chosen as a versatile software tool to solve this problem. The original analysis in 2004 used Blast which given the large datasets would be a very time consuming process to identify matches. Vmatch trades off huge memory space for extremely fast speed. Vmatch needs to pre-process the sequences using mkvtree program and store information into an index structure file then using vmatch for sequence match.

There were flags requirement for mkvtree program control persistent index. Set up DNA sequences as input file, prefix length for bucket sort as 5, all index tables as output files, and specifiy database file as dbEST files. After that configured flags for vmatch program. Set up match length as 300, specify xdrop value for edit distance extension as 100, show match sequence description as 18 or 22, query file to be match as legumes gene model, and subject file for match as dbEST files. Figure 2 shows flags for mkvtree and vmatch programs.

```
mkvtree -dna -pl 5 -allout -db gb1
vmatch -l 300 -exdrop 100 -showdesc 22 -q Lotus_CDS.fa gb1 > gblotus1.vmatch
vmatch -l 300 -exdrop 100 -showdesc 22 -q Mt3.0_cds_20090702_NAMED.fa gb1 > gbmedicago1.vmatch
```

Figure2. Flags for mkvtree and vmatch programs.

Analysis match result. Perl scripts were used to parse the Vmatch reports and dumped to mySQL database. The vmatch result is showed in figure 3, and after Perl scripts process is showed in figure 4. The matching result include subject length, subject number, subject relative position, type, query length, query number, query relative position, distance value, E-value, score value, and percent identity.

I used mySQL database query language to process Vmatch result analysis described below. If a Glyma gene model matched to any dbEST sequence with an E-value more significant than 10^{-4}, it was considered to be non-specific match and legume specific genes. If E-value between 10^{-4} and 10^{-8} was belong to a boundary range which need further analysis to decide homology or not. If Glyma genomic sequence not exist in the Vmatch result were treated as unique legumes genes.

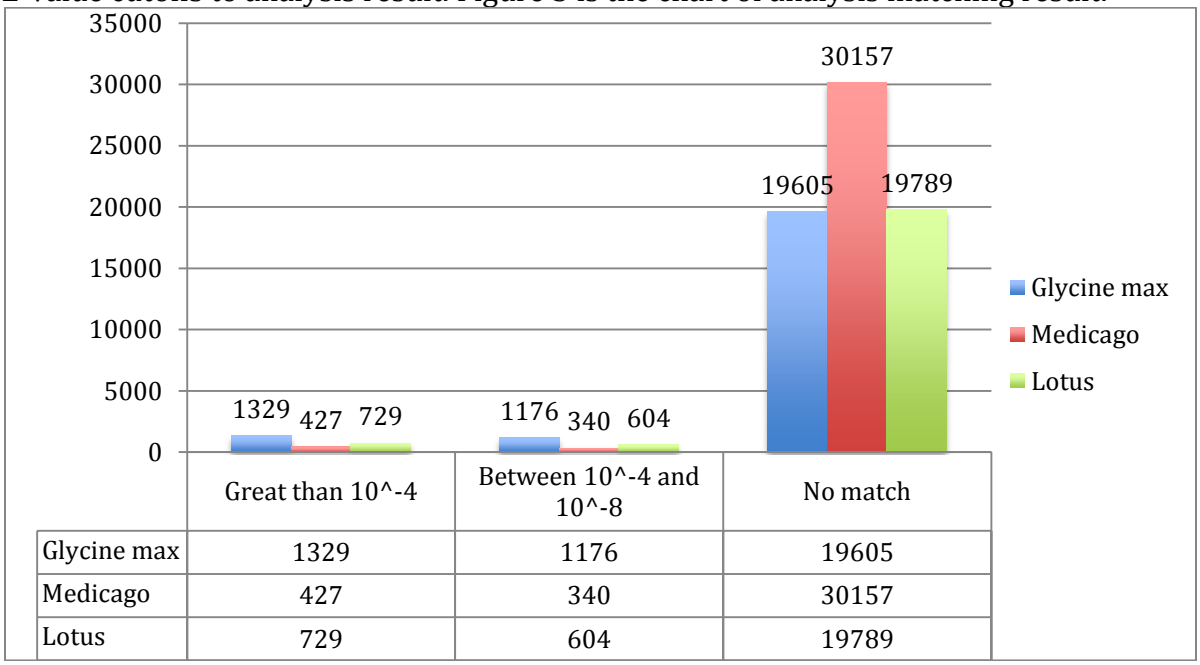
```
# args=-l 300 -exdrop 100 -showdesc 18 -q Glyma1.cds.fa /storage/yliu52/gbest/gb1
  566   AA660554_00440_MtR        28   D   495   Glyma0053s00200.1        2 196    0.00e+00    473    65.37
  563   AA660650_00538_MtR         0   D   486   Glyma0053s00200.1       68 212    0.00e+00    413    62.34
  465   AA824913_CT201.REV         0   D   465   Glyma01g01180.1      795  99    7.94e-88    633    78.71
  465   AA824913_CT201.REV         0   D   465   Glyma01g01180.2      795  99    7.91e-88    633    78.71
  465   AA824913_CT201.REV         0   D   465   Glyma01g01180.3      795  99    6.68e-88    633    78.71
  411   AA752962_97AS1419_         1   D   409   Glyma01g01370.1      814  96    2.44e-65    532    76.64
  438   AA429275_zv50c02.s         0   D   474   Glyma01g02340.1       45 218    0.00e+00    258    54.01
  398   AA825749_od52a07.s         0   D   389   Glyma01g02340.1       20 177    0.00e+00    256    55.53
  416   AA660413_00289_MtR        49   D   408   Glyma01g02720.1        0  83    1.17e-85    575    80.05
  395   AA660551_00437_MtR       107   D   407   Glyma01g02720.1        0 108    7.69e-49    478    73.46
  514   AA660344_00215_MtR        74   D   515   Glyma01g03260.1        0  61   1.06e-170    846    88.16
```

Figure3. Screenshot for Vmatch program result.

```
mysql> select * from estglymavh limit 10;
+-------+----------+-------+------+-------+-------------------+-------+--------+----------+-------+----------+
| sub_l | sub_n    | sub_r | type | que_l | que_n             | que_r | dvalue | evalue   | score | identity |
+-------+----------+-------+------+-------+-------------------+-------+--------+----------+-------+----------+
|   563 | AA660650 |     0 | D    |   486 | Glyma0053s00200.1 |    68 |    212 |        0 |   413 |    62.34 |
|   566 | AA660554 |    28 | D    |   495 | Glyma0053s00200.1 |     2 |    196 |        0 |   473 |    65.37 |
|   465 | AA824913 |     0 | D    |   465 | Glyma01g01180.1   |   795 |     99 | 7.94e-88 |   633 |    78.71 |
|   465 | AA824913 |     0 | D    |   465 | Glyma01g01180.2   |   795 |     99 | 7.91e-88 |   633 |    78.71 |
|   465 | AA824913 |     0 | D    |   465 | Glyma01g01180.3   |   795 |     99 | 6.68e-88 |   633 |    78.71 |
|   411 | AA752962 |     1 | D    |   409 | Glyma01g01370.1   |   814 |     96 | 2.44e-65 |   532 |    76.64 |
|   438 | AA429275 |     0 | D    |   474 | Glyma01g02340.1   |    45 |    218 |        0 |   258 |    54.01 |
|   398 | AA825749 |     0 | D    |   389 | Glyma01g02340.1   |    20 |    177 |        0 |   256 |    55.53 |
|   416 | AA660413 |    49 | D    |   408 | Glyma01g02720.1   |     0 |     83 | 1.17e-85 |   575 |    80.05 |
|   395 | AA660551 |   107 | D    |   407 | Glyma01g02720.1   |     0 |    108 | 7.69e-49 |   478 |    73.46 |
+-------+----------+-------+------+-------+-------------------+-------+--------+----------+-------+----------+
10 rows in set (0.02 sec)
```

Figure4. Screenshot for stored vmatch result in database.

Prepared other legumes gene models. Downloaded the next two legume gene models from http://medicago.org/ and http://www.kazusa.or.jp/lotus/ website, which were Medicago truncatula and Lotus japonicus. Perl scripts parsed those genomic sequences to capture Medicago gene identification number and Lotus gene identification number and backed up to database. I used that two legumes through Vmatch software to match against dbEST, after that stored result and used the same E-value cutoffs to analysis result. Figure 5 is the chart of analysis matching result.



	Great than 10^-4	Between 10^-4 and 10^-8	No match
Glycine max	1329	1176	19605
Medicago	427	340	30157
Lotus	729	604	19789

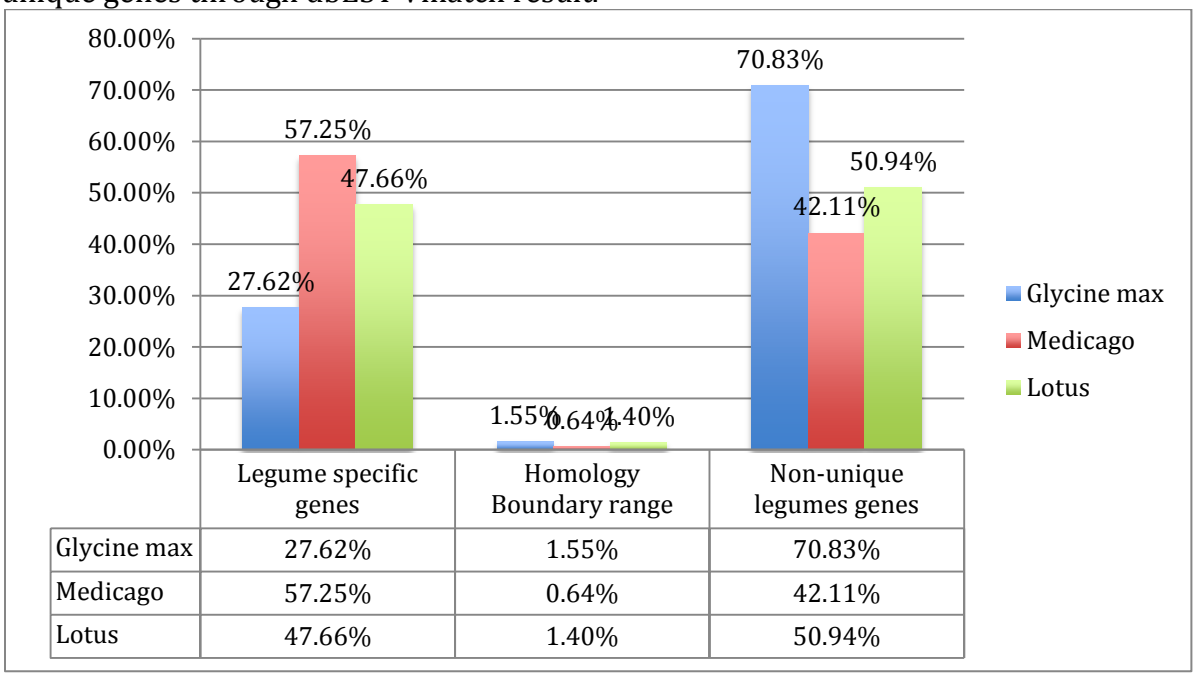
Figure5. Legumes gene models Vmatch against dbEST result.

Downloaded nt file. Through NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/), nt.gz file was available. That sequence file was modified on 7/20/10. Due to big size of that file, a Perl script was used to split every 5000 sequences as a smaller one. Captured nt number by Perl script, and dumped to database.

Redid Vmatch. Program of mkvtree and vmatch were need reconfigured. Flags for mkvtree program were list below. Set up DNA sequences as input file, prefix length for bucket sort as 5, all index tables as output files, and specifiy database file as nt files. Then flags for vmatch program. Set up match length as 300, specify xdrop value for edit distance extension as 100, show match sequence description as 34, query file to be match as Glyma, Medicago, or Lotus gene model, and subject file for match as nt files. Stored result to database, and analysed result used the same method as mentioned before.

Compared dbEST and nt Vmatch.

Nt database Vmatch program is still running. The compare between dbEST and nt are wait to be seen. If matching results are same, means we get the legume specific genes. However, if difference between that results, further analyze need to be continue. Like does non-redundant nt database hide some gene sequences information? Does dbEST matching result have duplication? Figure 6 is identified unique genes through dbEST Vmatch result.



	Legume specific genes	Homology Boundary range	Non-unique legumes genes
Glycine max	27.62%	1.55%	70.83%
Medicago	57.25%	0.64%	42.11%
Lotus	47.66%	1.40%	50.94%

Figure6: Legume specific genes from dbEST Vmatch result.

3. Results and Discussion

Vmatch is a versatile software tool for efficiently solving large sequence matching tasks; it trades off huge space and memory requirements for fast matching speed. As overwhelmingly large datasets are one characteristic of genomic studies, enough memory space to support Vmatch program use needs to be considered. In this study, we were able to utilize the R900 servers in the Bioinformatics department to allow enough space for the creation of the virtual suffix trees.

Perl script and shell script have powerful text processing facilities. Their flexibility and adaptability can easily parse Vmatch results and dump data to the database.

Databases like mySQL have obvious advantages to deal with datasets. E-value cutoffs to find unique genes by the help of query language become more fast and accurate.

4. Future Work

In this paper, we used Vmatch algorithms to compare gene sets from legumes gene models to dbEST and nt sequence files. We identified legume specific genes from all other known sequences.

In future work, we will describe unique legumes genes physical distribution, like are they clustered, are multiple genes of this class occur in homeologous blocks? All in all, we will further finding the distribution of legume-specific genes in the soybean genome relative to polyploidy duplications.

References

[1] Michelle A. Graham, Kevin A.T. Silverstein, Steven B. Cannon, and Kathryn A. VandenBosch. "Computational Identification and Characterization." Plant Physiology 135 (2004): 1179.

[2] Susan R. Singer, Sonja L. Maki, Andrew D. Farmer, Dan Ilut, Gregory D. May, Steven B. Cannon, and Jeff J. Doyle. (2009). Venturing Beyond Beans and Peas: What Can We Learn from Chamaecrista? *Plant Physiology* , 1041.