R You Out There?: Computational Strategies for Identifying Resistance (R) Genes in *Arabidopsis* and Beyond

Stephanie M. Brandt

Biochem218 Final Project
Spring 2002

I. Introduction

Historical Prospective

One of the first categories of Mendelian traits historically studied was natural variation of disease resistance in plants. This resistance was found to be monogenic, highly specific for conferring resistance to specific pathogens (Holub, 2001). In the mid-1900s, Harold Flor described a "gene for gene" relationship between plant resistance genes and pathogen. He observed that a plant was resistant if it possessed a single resistance gene corresponding to a virulence gene of the pathogen (Flor, 1971).

Monogenic disease resistance was first labeled as an artifact of selective breeding of crops (Crute et al., 1996). However, the recent sequencing of the genome of the wildflower *Arabidopsis thaliana* has revealed a highly polymorphic family of genes that determine genotype-specific resistance to environmental pathogens. So far, over 180 members have been identified (UC Davis R gene database, http://niblrrs.ucdavis.edu/At_Rgenes_Database/index.php3). These resistance genes (so called R genes for short) recognize a pathogen-specific pattern and trigger a hypersensitive response leading to rapid death of infected cells and neighbors (Lam et al., 2001). This theoretically prevents the spread of infection and is an important arm of the plant innate immune defense. Although the R family is quite diverse, recognizing moieties from mildews to viruses to insects to bacteria, similar families of genes have not yet been identified in other organisms that rely heavily on innate defense, namely *Caenhorabditis elegans* and *Drosophila melanogaster* (Holub, 2001).

Rationale

The identification of genotype-specific resistance genes in genetically-tractable model organisms would advance our understanding of host-pathogen interactions and provide target genes for examining naturally occurring susceptibility amongst individuals. The existence of R genes in non-plant organisms seems highly likely. Firstly, resistance genes are an ancient lineage whose divergence has been cited earlier than 200 million years ago (Holub, 2001). Secondly, pathogens, such as *Pseudomonas spp*., infect a broad host range, from plants to humans, supplying a selective pressure for conservation of these genes. Thirdly, secreted products of bacterial type III secretion apparatus are often the targets of R genes (Staskawicz et al., 2001). Many intracellular pathogens that minimize detection by invertebrate/vertebrate pattern-recognition receptors secrete these proteins; thus, likely R protein targets are presented to non-plant organisms. Lastly, proteins containing at least one or more of the conserved domains found in the R family exist in reasonable numbers in non-plant organisms, without assigned functions.

Structure of R genes

In regards to gene organization, R genes tend to cluster in tandem arrays within highly polymorphic loci (Michelmore et al., 1998). This suggests that the proliferative quality of the family may be the result of duplications. The presence of transposon sequence intermingled within these polymorphic loci strengthens this suggestion. In any case, structurally similar R proteins tend to be sprinkled in gene clusters throughout the genome (Holub, 2001).

Three Pfam domains have been identified regularly in the R protein family. The largest structural class contains a leucine-rich repeat domain (LRR) and a nucleotide-binding domain (NB). LRRs come in flavors of 10-40 repeats of approximately twenty-four amino acids that vary in sequence. The nucleotide binding domain is a highly conserved protein motif that is common to many proteins. Lastly, the presence or absence of an amino-terminal Toll/interleukin receptor-like domain (TIR) distinguishes the family into two clades, TIR-class and non-TIR class genes (Meyers et al., 1999). A complete list of the known R genes and their structural characteristics is provided in Appendix, Part A.

Both TIR and LRR domains have been implicated in conferring pathogen specificity (Luck et al., 2000; Caicedo et al., 1999). However, exact residue motifs have not been determined due to a small sampling number of redundant pathogen species-specific proteins. Identifying R genes in multiple organisms that identify the same virulence gene produced by a specific pathogen will help to elucidate the species recognition motifs.

Purpose of Project

The purpose of this study is to analyze the ability of current web accessible protein pattern-recognition algorithms to effectively select R family members within the Arabidopsis genome. A stricter criterion for finding domain-specific members will also be analyzed. The most accurate methods and training sets can then be used to gene mine in other organisms for putative R proteins.

II. Approach

Training Sets

The diversity of the R family of proteins is highlighted by the combination of recognition diversity, shear numbers in a single genome and flexibility of domains. Therefore, the efficacy of each pattern-recognition algorithm will be examined using training sets that include representative proteins chosen randomly from the entire family as compared to smaller training sets representing subgroups (by domains) within the family. The training sets are outlined below:

	Domain(s) Subgroups	R Family (n)	Training Set (n)
I	All	182	91
II	LRR + TIR	89	45
III	LRR	45	23
IV	TIR	26	13
V	NB	21	11

Measuring Efficacy

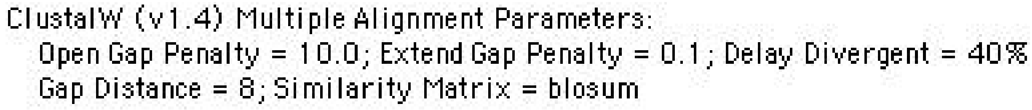
Each method was tested for its ability to generate positive hits within the *Arabidopsis thaliana* genome as stored within the NCBI Genbank database unless otherwise stated. A list of true positives divided into domain subgroups appears in Appendix, Part B. The ability of the training set to select R family members as a whole and, secondly, members within its subgroup was determined. For each training set, the total number of true positive hits to false positive hits was calculated and expressed as an ROC curve. The first one hundred ranked hits were analyzed for true R family hits; the first fifty ranked hits were analyzed for true subdivision hits. A significance cut-off was set at p-values<1.00, however all hits examined met this requirement.

Algorithms Tested

i. ClustalW Multiple Alignment Consensus Sequence
ii. Cobbler Sequence from MOTIF with Embedded BLOCKS
iii. Hidden Markov Model
iv. e-Matrix

i. ClustalW Multiple Alignment Consensus Sequence

Training sets were aligned by ClustalW multiple alignment using MacVector 7.0 with the following constraints:



ClustalW (v1.4) Multiple Alignment Parameters:
Open Gap Penalty = 10.0; Extend Gap Penalty = 0.1; Delay Divergent = 40%
Gap Distance = 8; Similarity Matrix = blosum

Consensus sequences were generated by MacVector 7.0. These contain the most frequent residues represented by the training set at each amino acid position. They appear in Appendix, Part C. These consensus sequences were then used to scan the Arabidopsis genome using the Smith-Waterman protein query vs. Arabidopsis database located at the DeCypher Machine (http://decypher.stanford.edu/index_by_algo.htm).

ii. Cobbler Sequence from MOTIF with Embedded BLOCKS

BLOCKS were generated for each training set using Block Maker located at the following URL: http://blocks.fhcrc.org/blocks/blockmkr/make_blocks.html. The output automatically assigned a COBBLER sequence, appearing in Appendix, Part E. This is a single sequence selected from a set of blocks that is enriched by replacing the conserved regions delineated by the blocks by consensus residues derived from the blocks. This embedded consensus was then used to query the Smith-Waterman protein vs. Arabidopsis database located at the DeCypher Machine.

iii. Hidden Markov Model

HMMs were generated using the DeCypher Machine. Training set sequences were first multiply aligned using ClustalW located on the DeCypher Machine with the same constraints as those used to generate the multiple alignment with the MacVector program. From the alignment, an HMM was built. This was queried against the Arabidopsis protein database.

iv. e-Matrix

Because the publicly accessible e-Matrix maker only scans the Swiss-Prot database, I enlisted the help of Jimmy Huang who was capable of scanning e-matrices against an Arabidopsis database located at Stanford at the following URL: http://baggage.stanford.edu/group/arabprotein/. He generated e-matrices using BLOCK aligned sequences I sent him. He scanned them against the Arabidopsis database with an expectancy threshold of .1 so as to ensure no false positive hits.

III. Results

ClustalW Multiple Alignment Consensus Sequence

The ability of each training set to identify R family members as a whole is depicted in Appendix, Part D, Figure 1. As could be predicted, the consensus sequence from the training set that contained proteins with neither a TIR nor LRR domain, but had a very common NB domain, was unable to effectively identify R family members. Surprisingly, the training set containing proteins with only a TIR domain was extremely effective at identifying R family proteins. The TIR domain is present in 64% of the R family proteins, second to the LRR domain that is present in 74%. Yet, the TIR training set is much more effective than the LRR alone set or the LRR + TIR set. This is most likely attributed to conservation within the domain itself. LRR domains are known to exhibit a fair amount of sequence flexibility within the repeats. The TIR domain is much better conserved. Highlighting the utility of preparing training sets based on conserved domains, the training set that included randomly selected proteins from the family as a whole performed poorly at selecting R family members as compared to the subgroup training sets. It consistently identified more false positives than true positives.

The ability of each training set to identify its own subgroup members is depicted in Appendix, Part D, Figure 2. In contrast to the TIR training set's stellar ability to find R family members, it is not a good set for identifying proteins that only contain TIR domains; most of its strongest hits are proteins that contain both a TIR and LRR domain. The best training set for finding members of its own subgroup was the LRR + TIR training set whose true positive hits were 91% from its own subgroup. This was not unexpected since this training set contained the largest number of input sequences. However, I was surprised and pleased that the LRR alone training set, with only 13 input sequences was capable of identifying members of its own subgroup 76% of the time.

Using a consensus sequence to identify novel family members is by no means a stringent criterion. However, if this method were to be used, I would recommend using the TIR consensus sequence to identify R family members in general. This will identify both TIR and LRR + TIR containing proteins. I would then use the LRR consensus sequence to identify LRR alone containing proteins. This will generate a fair amount of false positives. Proteins without NB domains can be dismissed. As a second weeding step, proteins that rank high as false positives from the NB scan can be dismissed. It appears that by refining the training sets to searches with multiple subgroups, false negatives can be minimized while allowing for coverage of the diversity within the family.

Note: Due to its small training set size and the poor performance of the NB training set, it was excluded from further analysis.

Cobbler Sequence from MOTIF with Embedded BLOCKS

The ability of each training set to identify R family members as a whole is depicted in Appendix, Part F, Figure 3. The use of COBBLER consensus sequence with embedded BLOCKS greatly improved the ability of all training sets to identify R family members. The most improved training set was the LRR subgroup. Interestingly, the ability of the training set formed form the R family as a whole also improved dramatically.

The ability of each training set to identify its own subdivision members is depicted in Appendix, Part F, Figure 4. Once again, almost all training sets exhibited a marked improvement in their ability to identify proteins of their specific subgroup. Remarkably, the LRR + TIR domain identified 100% of its first 50 positive hits to be members of its own subgroup.

Hidden Markov Model

In an effort to improve the ability of the LRR and TIR subgroups to identify members of their own subgroup, I created HMM positional scoring matrices for these training sets and accessed their ability to identify members of their own subgroup. The results are depicted in Appendix, Part G, Figures 5-6, as a comparison between three query methods: ClustalW consensus sequence, COBBLER sequence with embedded BLOCKS and HMM. Although accurate HMMs require a minimum of at least 100 sequences, an HMM for the LRR training set, that contains only 23 input sequences, was the best method for finding LRR domain members. Unfortunately, an HMM search did not improve the ability to identify TIR domain members.

e-Matrix

In yet a final effort to improve the ability of the TIR subgroup training set to identify members of the TIR subgroup, I created e-matrices from BLOCKS aligned sequences. Hoping that the NB domain might be preventing TIR alone hits because of a stricter homology to NB domains within the family as a whole, I created an e-matrix for each block. Four blocks (A-D) were generated, as seen in Appendix, Part H. As seen below, after scanning each block against the Swiss-Prot database with an expectancy threshold of 1, two of the blocks represented particular domains:

TIR_B

1. TLR2_HUMAN Toll-like receptor 2 precursor (Toll/interleukin 1
   692-SIEKSHKTVFVLSENFVKSEWCKYELDFSH-721

2. TLR2_MACFA Toll-like receptor 2 precursor.
   692-SIEKSHKTVFVLSENFVKSEWCKYELDFSH-721

3. TLR2_MOUSE Toll-like receptor 2 precursor.
   692-SIEKSHKTVFVLSENFVRSEWCKYELDFSH-721

4. TLR2_BOVIN Toll-like receptor 2 precursor.
   692-SIEKSHKTIFVLSENFVKSEWCKYELDFSH-721

5. TLR2_CRIGR Toll-like receptor 2 precursor.
   692-SIEKSHKTLFVLSENFVRSEWCKYELDFSH-721

TIR_D

1. APRD_PSEAE Alkaline protease secretion ATP-binding protein ap
   360-SVVGVIGPSGSGKSSLARVVL-380

2. YXEO_BACSU Probable amino-acid ABC transporter ATP-binding pr
   28-EVVAIIGPSGSGKSTLLRCLN-48

3. NIH2_METBA Nitrogenase iron protein 2 (EC 1.18.6.1) (Nitrogen
   2-RQIAIYGKGGIGKSTTTQNLT-22

4. YCKI_BACSU Probable amino-acid ABC transporter ATP-binding pr
   28-KVIAILGPSGSGKTTLLRCLN-48

5. NIF4_CLOPA Nitrogenase iron protein 4 (EC 1.18.6.1) (Nitrogen
   2-RQVAIYGKGGIGKSTTTQNLT-22

TIR_B appears to be a block recognizing the TIR domain. TIR_D appears to be a block recognizing the NB domain. Unfortunately, as depicted in Appendix, Part I, Figure 7, the creation of e-matrices for specific blocks was no better, and in some cases worse, than Cobbler consensus with embedded BLOCKS at identifying R family proteins containing only a TIR domain.

IV. Discussion

ClustalW alignment finds the best global alignment for a set of input sequences. A global alignment refers to the best match over the total length of the sequences. How well the consensus sequence from this alignment will be a reflection of any given input sequence is dependent on the number of input sequences and the total residue diversity. The sensitivity of a given multiple sequence alignment can be improved through sequence weighting, position specific gap penalties and weight matrix choice. In this report, to maintain a fair comparison of each of the training sets, these parameters were unchanged. But, conceivably, increasing the stringency of any of these parameters should generate stricter multiple alignments and subsequent consensus sequences that could improve the ability of each training set to identify true positive hits.

As the parameters were set, the TIR domain ClustalW consensus sequence most accurately reflected conserved domains within the family, TIR and NB. This training set was not effective at identifying TIR domain alone containing proteins because it strongly identified LRR + TIR proteins as well. Bringing up the important point that even good training sets will be inherently biased towards the recognition of a subset of proteins. The ability to have full-coverage of a family of proteins by recognizing subsets within and developing training sets for recognition of underrepresented domains/motifs is important. Particularly in the case of ClustalW for which the consensus sequence is representative of over represented residues.

An alternate method to increasing the stringency of a ClustalW alignment, is the use of the BLOCKmaker. This produces multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. For each training set, different BLOCK sequences were generated and embedded in a Cobbler consensus sequence. These queries were extremely efficient at identifying true positives. The diverse nature of the BLOCKS amongst the training sets lent confidence that coverage of the diversity in the family would be achieved. Sure enough, the BLOCKs-built trainers were much better at identifying subgroup members. The laggard ability of the LRR and TIR trainers to identify members of their subgroup is mostly due to the fact that highly conserved regions are common to their subgroup and the LRR + TIR subgroup. By using methods that do not focus solely on highly conserved regions, but on statistical representation of a particular amino acid at every position in a group of proteins, perhaps the subtle differences between these three subgroups can be teased apart.

One such algorithm for doing just this is the Hidden Markov Model (HMM). This is used to perform sensitive database searching using statistical descriptions of a family's consensus sequence. HMMs estimate the true frequency of a residue at a given position in the alignment from its observed frequency. This means that a profile HMM derived from only 10 to 20 aligned sequences can be of equivalent quality to a standard profile created from 40 to 50 aligned sequences (Profile Hidden Markov Analysis, http://www.sacs.ucsf.edu/Documentation/gcghelp/hmmanalysis.html). Non-TIR class LRR containing proteins are theorized to be more ancient than TIR-class LRR proteins and are phylogenetically more related to each other than any other LRR containing protein in the TIR class. This provides encouragement that HMM, even with a small training set of 23, will locate subtle residue preferences. The HMM in fact does improve the efficacy of the LRR trainer by 13%, such that 86% of its true positive hits are non-TIR LRR proteins. Unfortunately, the TIR proteins are phylogenetically closely related to TIR-class LRR proteins. As expected, the HMM procedure does not significantly enhance the ability of the TIR trainer to recognize members of its subgroup. This was also true for a second positional specific scoring matrix method, e-Matrix that estimates the frequency of a residue at a given position in the alignment based on minimal-risk scoring matrices .

Thus, it appears that in the case of distinguishing two very similar subgroups of proteins between which the similarities far outweigh the differences, the four methods utilized fall short. However, a subtractive comparison of position specific scoring matrices might generate specificity for each group. By allowing similar scores at a shared position between two position specific-scoring matrices to cancel out, the differences between the two sets will remain. When scanning a database, the hits would be required to fulfill the first “unsubtracted” matrix and then the second “subtracted” matrix. This would hopefully define the two subgroups.

Each method utilized in this study was a progressive step towards a more sophisticated and specific means for identifying R family proteins. Although not perfect, these methods were adequate. In the future, I hope to use these methods to identify R proteins in other organisms. Having a battery of methods with increasing stringency/specificity will increase the chances of producing statistically significant hits in divergent organisms. Possessing subgroup specific training sets will enhance family coverage and allow for identification of R proteins in organisms in which one or more subgroups have been selected out.

The identification of R genes in non-plant organisms will be a great step forward for the role of innate detection in immunity. While genetic screens are rapidly being performed to find such things, the power of computational biology will hopefully expedite this process and generate feasible targets to pursue in the future.

V. Citations

Caicedo, A. et al. Diversity and molecular evolution of the RPS2 resistance gene in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **96**, 302-306 (1999).

Crute, I.R., Pink, D.A. Genetics and utilization of pathogen resistance in plants. *Plant Cell* **8**, 2747-2755 (1996).

Flor, H.H. The current status of the gene for gene concept. *Annu. Rev. Phytopathol.* **9**. 275-296 (1971).

Holub, E.B. The arms race is ancient history in *Arabidopsis*, the wildflower. *Nature Genet.* **2**. 516-527 (2001).

Lam, E., Kato, N., et al. Programmed cell death, mitochondria and the plant hypersensitive response. *Nature*. **411**. 796-815 (2001).

Luck, J.E., et al. Regions outside of the leucine-rich repeats of flax rust resistance proteins play a role in specificity determination. *Plant Cell*. **12**. 1367-1377 (2000).

Meyers, B. et al. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.* **20**. 317-332 (1999).

Michelmore, R.W., Meyers, B.C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**. 1113-1130 (1998).

Staskawicz, B.J., Mudgett, M.B., et al. Plant disease-resistance proteins and the gene-for-gene concept. *Science*. **292**. 2285-2289 (2001).

Appendix: Part A

R Family Protein Members

http://niblrrs.ucdavis.edu/At_RGenes/RGenes_Database/index.php3

Protein_ID	TIR_Pfam	NBS_Pfam	LRR_Pfam
At1g10920	-	NB-ARC	LRR
At1g12210	-	NB-ARC	LRR
At1g12220	-	NB-ARC	LRR
At1g12280	-	NB-ARC	LRR
At1g12290	-	NB-ARC	LRR
At1g15890	-	NB-ARC	LRR
At1g17600	TIR	NB-ARC	LRR
At1g17610	TIR	NB-ARC	-
At1g27170	TIR	NB-ARC	LRR
At1g27180	TIR	NB-ARC	LRR
At1g31540	TIR	NB-ARC	LRR
At1g33560	-	NB-ARC	LRR
At1g47370	TIR	-	-
At1g50180	-	NB-ARC	-
At1g51480	-	NB-ARC	LRR
At1g52660	-	NB-ARC	-
At1g53350	-	NB-ARC	LRR
At1g53360	-	-	-
At1g56510	TIR	NB-ARC	LRR
At1g56520	TIR	NB-ARC	LRR
At1g56540	TIR	NB-ARC	LRR
At1g57650	-	NB-ARC	-
At1g57670	TIR	-	-
At1g58390	-	NB-ARC	-
At1g58400	-	NB-ARC	-
At1g58410	-	NB-ARC	LRR
At1g59620	-	NB-ARC	-
At1g59780	-	NB-ARC	-
At1g61180	-	NB-ARC	LRR
At1g61190	-	NB-ARC	LRR
At1g61300	-	NB-ARC	LRR
At1g61310	-	NB-ARC	LRR
At1g62630	-	NB-ARC	LRR
At1g63350	-	NB-ARC	-
At1g63360	-	NB-ARC	LRR
At1g63730	TIR	NB-ARC	LRR
At1g63740	TIR	NB-ARC	LRR
At1g63750	TIR	NB-ARC	LRR
At1g63870	TIR	NB-ARC	LRR
At1g63880	TIR	NB-ARC	LRR
At1g64070	TIR	NB-ARC	LRR
At1g65850	TIR	NB-ARC	LRR
At1g66090	TIR	NB-ARC	-
At1g69550	TIR	NB-ARC	LRR
At1g72840	TIR	NB-ARC	LRR
At1g72850	TIR	NB-ARC	-
At1g72860	TIR	NB-ARC	LRR
At1g72870	TIR	NB-ARC	-
At1g72890	TIR	NB-ARC	-
At1g72900	TIR	NB-ARC	-
At1g72910	TIR	NB-ARC	-
At1g72920	TIR	NB-ARC	-
At1g72930	TIR	-	-
At1g72940	TIR	NB-ARC	-
At1g72950	TIR	NB-ARC	-
At2g14080	TIR	NB-ARC	LRR
At2g16870	TIR	NB-ARC	LRR
At2g17050	TIR	NB-ARC	LRR
At2g17060	TIR	NB-ARC	LRR
AT3g04210	TIR	NB-ARC	-
AT3g04220	TIR	NB-ARC	LRR
AT3g07040	-	NB-ARC	LRR
AT3g14460	-	NB-ARC	LRR
AT3g14470	-	NB-ARC	LRR
AT3g15700	-	NB-ARC	-
AT3g23270	-	-	-
AT3g26470	-	-	-
AT3g44400	TIR	NB-ARC	LRR
AT3g44480	TIR	NB-ARC	LRR
AT3g44630	TIR	NB-ARC	LRR
AT3g44670	TIR	NB-ARC	LRR
AT3g46530	-	NB-ARC	LRR
AT3g46710	-	NB-ARC	LRR
AT3g46730	-	NB-ARC	LRR
AT3g50950	-	NB-ARC	-
AT3g51560	TIR	NB-ARC	LRR
AT3g51570	TIR	NB-ARC	LRR
AT4g04110	TIR	-	-
AT4g08450	TIR	NB-ARC	LRR
AT4g08470	-	-	-
AT4g08480	-	-	-
AT4g08500	-	-	-
AT4g09360	-	NB-ARC	LRR
AT4g09420	TIR	NB-ARC	-
AT4g09430	TIR	NB-ARC	LRR
AT4g10780	-	NB-ARC	LRR
AT4g11170	TIR	NB-ARC	LRR
AT4g12010	TIR	NB-ARC	LRR
AT4g12020	TIR	NB-ARC	LRR
AT4g14370	TIR	NB-ARC	LRR
AT4g14610	-	NB-ARC	LRR
AT4g16860	TIR	NB-ARC	LRR
AT4g16890	TIR	NB-ARC	LRR
AT4g16900	TIR	NB-ARC	LRR
AT4g16920	TIR	NB-ARC	LRR
AT4g16940	TIR	NB-ARC	LRR
AT4g16950	TIR	NB-ARC	LRR
AT4g16960	TIR	NB-ARC	LRR
AT4g16990	TIR	NB-ARC	-
AT4g19050	-	-	LRR
AT4g19060	-	NB-ARC	-
AT4g19500	TIR	NB-ARC	LRR
AT4g19510	TIR	NB-ARC	LRR
AT4g19520	TIR	NB-ARC	LRR
AT4g19530	TIR	NB-ARC	LRR
AT4g19920	TIR	-	-
AT4g23510	TIR	-	-
AT4g26090	-	NB-ARC	LRR
AT4g27190	-	NB-ARC	LRR
AT4g27220	-	NB-ARC	LRR
AT4g33300	-	NB-ARC	LRR
AT4g35470	-	-	LRR
AT4g36140	TIR	NB-ARC	LRR
AT4g36150	TIR	NB-ARC	LRR
AT5g04720	-	NB-ARC	LRR
AT5g05400	-	NB-ARC	LRR
AT5g11250	TIR	NB-ARC	LRR
AT5g17680	TIR	NB-ARC	LRR
AT5g17880	TIR	NB-ARC	LRR
AT5g17890	TIR	NB-ARC	LRR
AT5g17970	TIR	NB-ARC	LRR
AT5g18350	TIR	NB-ARC	LRR
AT5g18360	TIR	NB-ARC	LRR
AT5g18370	TIR	NB-ARC	LRR
AT5g22690	TIR	NB-ARC	LRR
AT5g35450	-	NB-ARC	LRR
AT5g36930	TIR	NB-ARC	LRR
AT5g38340	TIR	NB-ARC	LRR
AT5g38350	-	NB-ARC	LRR
AT5g38850	TIR	NB-ARC	LRR
AT5g40060	TIR	NB-ARC	LRR
AT5g40090	TIR	NB-ARC	-
AT5g40100	TIR	NB-ARC	LRR
AT5g40910	TIR	NB-ARC	LRR
AT5g40920	TIR	NB-ARC	LRR
AT5g41540	TIR	NB-ARC	LRR
AT5g41550	TIR	NB-ARC	LRR
AT5g41740	TIR	NB-ARC	LRR
AT5g41750	TIR	NB-ARC	LRR
AT5g43470	-	NB-ARC	LRR
AT5g43730	-	NB-ARC	LRR
AT5g43740	-	NB-ARC	LRR
AT5g44510	TIR	NB-ARC	LRR
AT5g44870	TIR	NB-ARC	LRR
AT5g45000	TIR	-	-
AT5g45050	TIR	NB-ARC	LRR
AT5g45060	TIR	NB-ARC	LRR
AT5g45070	TIR	-	-
AT5g45200	TIR	NB-ARC	LRR
AT5g45210	TIR	NB-ARC	LRR
AT5g45220	TIR	-	-
AT5g45230	TIR	NB-ARC	LRR
AT5g45240	TIR	NB-ARC	LRR
AT5g45250	TIR	NB-ARC	LRR
AT5g45260	TIR	NB-ARC	LRR
AT5g45440	-	NB-ARC	-
AT5g45490	-	NB-ARC	-
AT5g45510	-	-	LRR
AT5g45520	-	NB-ARC	-
AT5g46260	TIR	NB-ARC	LRR
AT5g46270	TIR	NB-ARC	LRR
AT5g46450	TIR	NB-ARC	LRR
AT5g46470	TIR	NB-ARC	LRR
AT5g46480	TIR	-	-
AT5g46490	TIR	NB-ARC	LRR
AT5g46510	TIR	NB-ARC	LRR
AT5g46520	TIR	NB-ARC	LRR
AT5g47250	-	NB-ARC	LRR
AT5g47260	-	NB-ARC	LRR
AT5g47280	-	NB-ARC	LRR
AT5g48620	-	NB-ARC	LRR
AT5g48770	TIR	NB-ARC	LRR
AT5g48780	TIR	NB-ARC	-
AT5g49140	TIR	NB-ARC	LRR
AT5g51630	TIR	NB-ARC	LRR
AT5g58120	TIR	NB-ARC	LRR
AT5g63020	-	NB-ARC	LRR
AT5g66630	-	NB-ARC	-
AT5g66890	-	NB-ARC	-
AT5g66900	-	NB-ARC	LRR
AT5g66910	-	NB-ARC	LRR

Appendix: Part B

True Positives for Each Domain Subdivision

LRR+TIR	LRR	TIR	NB
At1g17600.swi	At1g10920 .swi	At1g17610.swi	At1g50180.swi
At1g27170.swi	At1g12210.swi	At1g47370.swi	At1g52660.swi
At1g27180.swi	At1g12220.swi	At1g57670.swi	At1g57650.swi
At1g31540.swi	At1g12280.swi	At1g66090.swi	At1g58390.swi
At1g56510.swi	At1g12290.swi	At1g72840.swi	At1g58400.swi
At1g56520.swi	At1g15890.swi	At1g72850.swi	At1g59620.swi
At1g56540.swi	At1g33560.swi	At1g72870.swi	At1g59780.swi
At1g63730.swi	At1g51480.swi	At1g72890.swi	At1g63350.swi
At1g63740.swi	At1g53350.swi	At1g72900.swi	At3g15700.swi
At1g63750.swi	At1g58410.swi	At1g72910.swi	At3g23270.swi
At1g63870.swi	At1g61180.swi	At1g72920.swi	At3g26470.swi
At1g63880.swi	At1g61190.swi	At1g72930.swi	At3g50950.swi
At1g64070.swi	At1g61300.swi	At1g72940.swi	At4g08470.swi
At1g65850.swi	At1g61310.swi	At1g72950.swi	At4g08480.swi
At1g69550.swi	At1g62630.swi	At3g04210.swi	At4g08500.swi
At1g72860.swi	At1g63360.swi	At4g04110.swi	At4g19060.swi
At2g14080.swi	At3g07040.swi	At4g09420.swi	At5g45440.swi
At2g16870.swi	At3g14460.swi	At4g16990.swi	At5g45490.swi
At2g17050.swi	At3g14470.swi	At4g19920.swi	At5g45520.swi
At2g17060.swi	At3g46530.swi	At4g23510.swi	At5g66630.swi
At3g04220.swi	At3g46710.swi	At5g40090.swi	At5g66890.swi
At3g25510.swi	At3g46730.swi	At5g45000.swi	
At3g44400.swi	At4g09360.swi	At5g45070.swi	
At3g44480.swi	At4g10780.swi	At5g45220.swi	
At3g44630.swi	At4g14610.swi	At5g46480.swi	
At3g44670.swi	AT4g19050.swi	At5g48780.swi	
At3g51560.swi	At4g26090.swi		
At3g51570.swi	At4g27190.swi		
At4g08450.swi	At4g27220.swi		
At4g09430.swi	At4g33300.swi		
At4g11170.swi	At4g35470.swi		
At4g12010.swi	At5g04720.swi		
At4g12020.swi	At5g05400.swi		
At4g14370.swi	At5g35450.swi		
At4g16860.swi	At5g43470.swi		
At4g16890.swi	At5g43730.swi		
At4g16900.swi	At5g43740.swi		
At4g16920.swi	At5g45510.swi		
At4g16940.swi	At5g47250.swi		
At4g16950.swi	At5g47260.swi		
At4g16960.swi	At5g47280.swi		
At4g19500.swi	At5g48620.swi		
At4g19510.swi	At5g63020.swi		
At4g19520.swi	At5g66900.swi		
At4g19530.swi	At5g66910.swi		
At4g36140.swi			
At4g36150.swi			
At5g11250.swi			
At5g17680.swi			
At5g17880.swi			
At5g17890.swi			
At5g17970.swi			
At5g18350.swi			
At5g18360.swi			
At5g18370.swi			
At5g22690.swi			
At5g36930.swi			
At5g38340.swi			
At5g38350.swi			
At5g38850.swi			
At5g40060.swi			
At5g40100.swi			
At5g40910.swi			
At5g40920.swi			
At5g41540.swi			
At5g41550.swi			
At5g41740.swi			
At5g41750.swi			
At5g44510.swi			
At5g44870.swi			
At5g45050.swi			
At5g45060.swi			
At5g45200.swi			
At5g45210.swi			
At5g45230.swi			
At5g45240.swi			
At5g45250.swi			
At5g45260.swi			
At5g46260.swi			
At5g46270.swi			
At5g46450.swi			
At5g46470.swi			
At5g46490.swi			
At5g46510.swi			
At5g46520.swi			
At5g48770.swi			
At5g49140.swi			
At5g51630.swi			
At5g58120.swi			

Appendix: Part C

raining Set	ClustalW Consensus Sequence
ll	MSEKEELPLTLTSIGAATATSDYHQRVGSSGEGISSSSSDVDPRFMQNSPTGLMISQSSSMCTVPPGMAATP SSGSGLSQQLNNSSSSKLCQVEGCQKGARDASGRCISHGGGRRCQKPDCQKGAEGKTVYCKAHGGGRRC YLGCTKGAEGSTDFCIAHGGGRRCNHEDCTRSAWGRTEFCVKHGGGARCKTYGCGKSASGPLPFCRAHG KKCSHEDCTGFARGRSGLCLMHGGGKRCQRENCTKSAEGLSGLCISHGGGRRCQSIGCTKGAKGSKMFCK CITKRPLTIDGGGNMGGVTTGDALNYLKAVKDKFEDSEKYDTFLEVLNDCKHQGVDTSGVIARLKDLFKGHD LGFNTYLSKEYQITILPEDDFPIDFLDKVEGPYEMTYQQAQTVQANANMQPQTEYPSSSAVQSFSSGQPQIPT PDSSLLAKSNTSGITIIEHMSQQPLNVDKQVNDGYNWQKYGQKKVKGSKFPLSYYKCTYLGCPSKRKVERSL QVAEIVYKDRHNHEPPNQGKDGSTTYLSGSSTHINCMSSELTASQFSSNKTKIEQQEAASLATTIEYMSEASD EDSNGETSEGEKDEDEPEPKRRITEVQVSELADASDRTVREPRVIFQTTSEVDNLDDGYRWRKYGQKVVKG YPRFSSSKDYDVVIRYGRADISNEDFISHLRASLCRRGISVYEKFNEVDALPKCRVLIIVLTSTYVPSNLLNILEH TEDRVVYPIFYRLSPYDFVCNSKNYERFYLQDEPKKWQAALKEITQMPGYTLTDKSESELIDEIVRDALKVLCS KVNMIGMDMQVEEIGGESSTFRLNDEVFSSSKSTSSSPSSSSSVFSFGDVPRFSHIFDLHLAISVSYSWRLW DYIRHVFINDTGTSAIEGIFLDMLNLKFDANPNVFYCLELIGVSFPQGLEYLPSKLRLLHVFNLTQYPVQGFFWK KCEWALCGDLIIRLINIYFSKYYTPEILVGELRTGIGEEAGIGKTTALFREFFANLPMLLLKVLVLDDVLWFFGGS TTKGLGLYVLAFAFNFLNIQSFSILRVLCLVLPPILLLLVGLKEWLISLSYDLSPKFLIASELQDLPFFILGLLLVLLL
RR + TIR	MSEKEELPLTLTSIGAATATSDYHQRVGSSGEGISSSSSDVDPRFMQNSPTGLMISQSSSMCTVPPGMAATP SSGSGLSQQLGSRNDRSAEVSSEDKRRSSSSPGGGNMGGSSSSPSSSYDVFLEVLNDCKHQGVDTSGVIA KSFGDVRFLSHLKRGPEDDFIFDGPYEMTYQQAQTVQANANMQPQTEYPSSSAIERIPLAISLLAKSNTSRIAI SQQPLNVDKQVNNYASSWGQKKVKGSKFPLSYYKCTYLGCPSKRKVERSLDGQVAEIVYKDRHNHEPPNQ DGSTTYLELVEICSSNKTKIEQQEAAVPVFYVDPSDNEEDSNGETSEGEKDEDEPEPKRRITEVQVSELADAS TVRKQRVIFQTTSEVDNLDDGYRWRKYGQKVVGPYPRFGKDYDVVIRYGRADISNEDFISHLRASLCRRGIS EKFNEVDALPKCRVLIIVLTSTYVPSNLLNILEHQHTEDRVVYPIFKTSPYDFVCEEQWALVNIGWEAEGINREK DDMIEIADVSKLPDFGPFGEVGEHLLLDSRMGIWGGIGKTTIARALKSFFLPMPDDKLLQLSILILGARLKKVLIV DVDQLALWFGGSKSRIITTDLHNIYVPALFCAFPGFLAPYVLAGLPLLVLGSLRGKIEWLRLLDILSYDLKKFLIA FNVLLLGLLKSLIKQIVQVLLLNKFSHVRHTKRNKEMHLLGREIVRSLWPRFLIVELTGTVGILDDIAFMNLFLYK LLKWNYCPLKSLPSTFKAEYLVNLIMKYSKLEWWTLPRGSLKKMDLGCSNNLKEIPDLSLAINLEPPRKLRLLQ NVELEGKSLEGMCNLEYLSVDWSSMEGTQGLYLPLRLLWPLPFPLVLMSLEKLWGTTNSSLLKLSLKEPDLL TNLELLCSLVELPSSSQLQNLRVVNLSGCTEIKCFSGVPPNIEELHLQGTRIREIPIFNATHPPKVKLDRKKLWN NLKLLLRRCSLSEFLDVSGLKLEKLFLCLLPNTLYLKQTGIRSIPTVTFSPQDNSFIYDHKDLLCLPNLSLLLCSL SIKTPKLPVLSMSECEDLQAFPTYIESLLFPISGNLLGCRIVEDCLWNKNLPLTGLDYLTIEPSCLLPLLLLSCLPL
RR	MVQHWDGDPAVRTKSHIEMLHPACEMSWNLKQHDDTAASLMVIQSPTSDPVFNPVHTIKQFGERYTQDFVI RDKSFGVLIMMYILNLLEDLSNDVEELWLLSSSSNRLCNSQLCYGVLEVLFERLQTQEVGLLDSGLGMGGVG TLNEFDIWVVSTGEHAEGEAQILWEELLKFVLLDDWVDLIGPPDLGKFTTRSVMVCLAWLFKVLEFSILAVAKC PLALVIGMKQEWLEFSDMILPLKSYDLKTILDCFLYSCLFPEDIELIWCEGFIRGGLVRLLLANLMHDVDPREMA WSGEVGTRAGGIPWRRTPKKFFSLLTLLLRLSIEELKALTKLNTLEVSGASLSFFMLVLDLSTKIELPSCLNLQN LLPEISLLYLLSTSGLRTCFDNADILPNLLLFGGLHELKELSHLRGTLRISELQNVAFASEASLLSGILVALLLRPH TFCIESYQGGASDSFFGITELLLLGENNSRGVPFQSLQILKFYGMPRWDEWICPELEDGIFPCLQKLIIQRCPSL KKFPEGLPSSTEVTISDCPLRAVSGGENSFRLLSNPKSDASTSAQPGFASSSQSNLQTEDFDQYETQLGSLP FEEPAVISARYSGYISDIPSTLSPYMSRTSLVPDPKNEGSILPGSSLLSSEISPQNLQSLFLILSHPPTTLKFANLL VIKKFLLLTFSIHAGLGDDRILITFPQGGLPLLCPKLIFPSNLRTLCISLCDEWGLRDLENLRNLEIDGGNLTYLNL HLEFLIYVDHSDVPPPHEC
IR	MDSYFFLGTVVVVALAIYTLLGTISFMVYRKFRLHQEKNNISSCFSFSMSSSSSVFLSFGDRFVSHLLIFDEER GIAIESAVVSYASWCLELILVIPIFYVDPDVRQGFGFRWRALGSSRNLIIIETVSVLGLMTDIDMKYKLNSLSGSV GFGLERYALEATSVMDDMSNSLVGHMKLLLVGMVIGCLLVLFFVLKQSAYTSTRIGIWGGAISAGKIADNVTAY PRDTYTFHFRPRPLTPPVLLTRKFQVVLVDVDQAWFPGSRITLLGYEVLGLALQFAFPFELRAVAGEPMIKRW GSWLLKNPFYIRAFLKNCDADVGIESSSLAMLVNLLRLPFRFTRPRLTSLVLRKLLEDLKSWLFQTSLLSGLSG DHLSFSSVQKTAHQSVTHLLNSGFFGLKSLDIKRFSYRLDPVNFSCLSFADFPCLTELKLINLNIEDIPEDICQL ETLDLGGNDFVYLPTSMGQLAMLKYLSLSNCRRLKALPQLSQVERLVLSGCVKLGSLMGILGAGRYNLLDFC KCKSLGSLMGILSVEKSAPGRNELLELSLENCKSLVSLSEELSHFTKLTYLDLSSLEFRRIPTSIRELSFMRTLY NCNKIFSLTDLPESLKYLYAHGCESLEHVNFSSNHSFNHLDFSHCISLECISDLVRDFMNEEYSQEAPFRLVCI YSIASTNNMRTSWREPMRIKLPKIKAAPKLVGFFVQIMVVCEKPFHLQFPAFSYNWDCEGSRLYRINLKPNLY SEMMEDNNNRPYKWHHLVIVQIPTGIISAEIDEVQFESHIQVPEPGDEIILCGVEHVGFVLK
IB	MADPASCEVDSWVRVRGSLKRIGKDQGEIDRPISIDFVPSMDRRGGSGPDDDEVGLEFTGMGGLFVRQGE IPKLVSDEELDWLIPGSGAKFDLKCGGLVGLLHDWGVRVLYLPDEGVPDLAPVEEGDRTPTLVDGEEVDRLD KDCLKEEEKKTSPSEESSSPGDQKGIGVTGSSVSLWWKLLELDKRVESGKVVEDLPEILLNILKVEKEGDREG HDEVESSIDVLPKEKVNLEEKHDSDGLLKGTEIERFQKADGLDVDKGKLKVILEPPDGSKSKIVDILSKKQYSR TPPILEKDVLRKWCRTKKGKEQFPTNKPPSWLNPDLKNLIKLSRIGELKPDLKVPSTYPNSM

Appendix: Part D

Figure 1: ROC Comparison of ClustalW Consensus Sequences from Domain Specific Subsets of R Proteins: Ability to Identify R Family Members

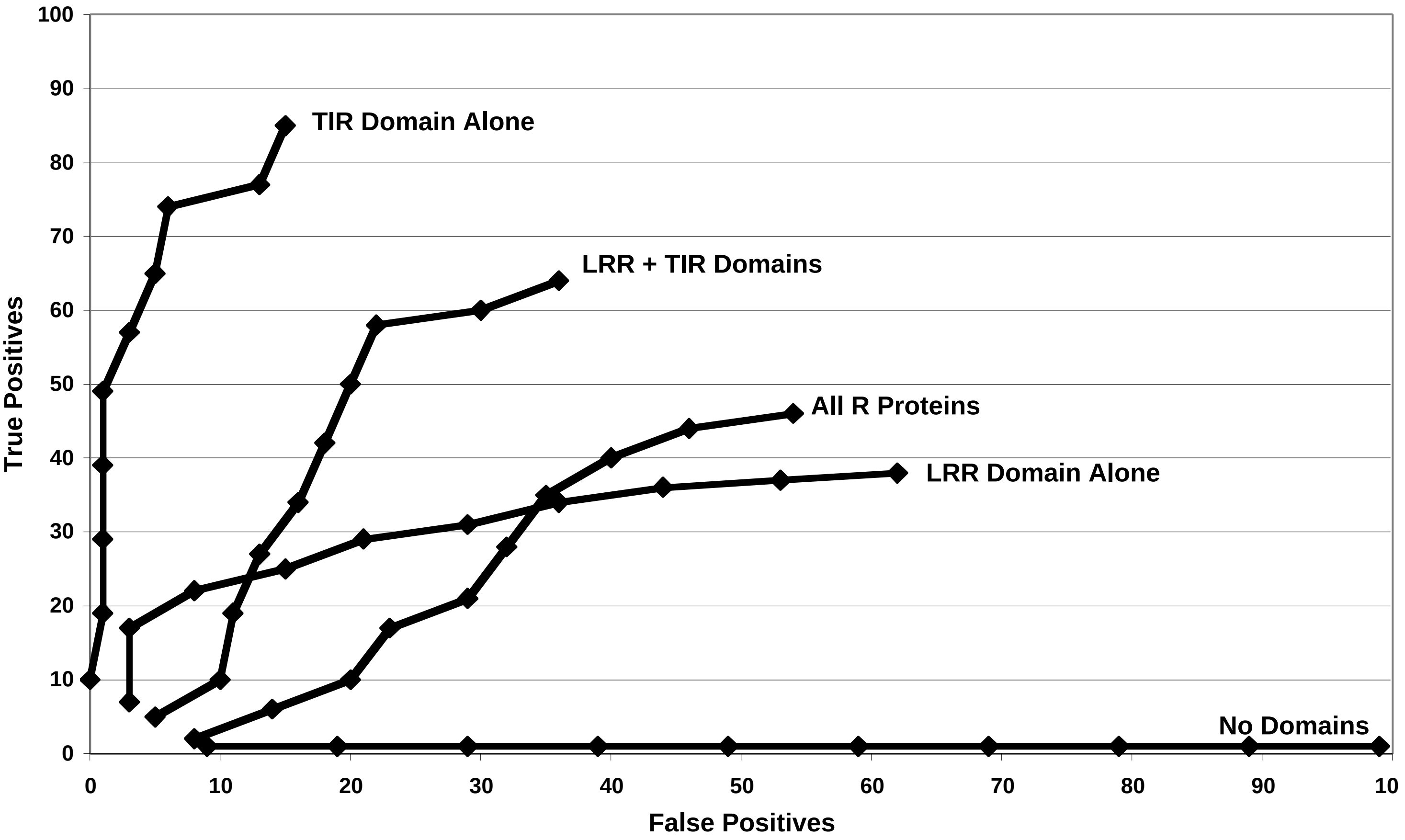
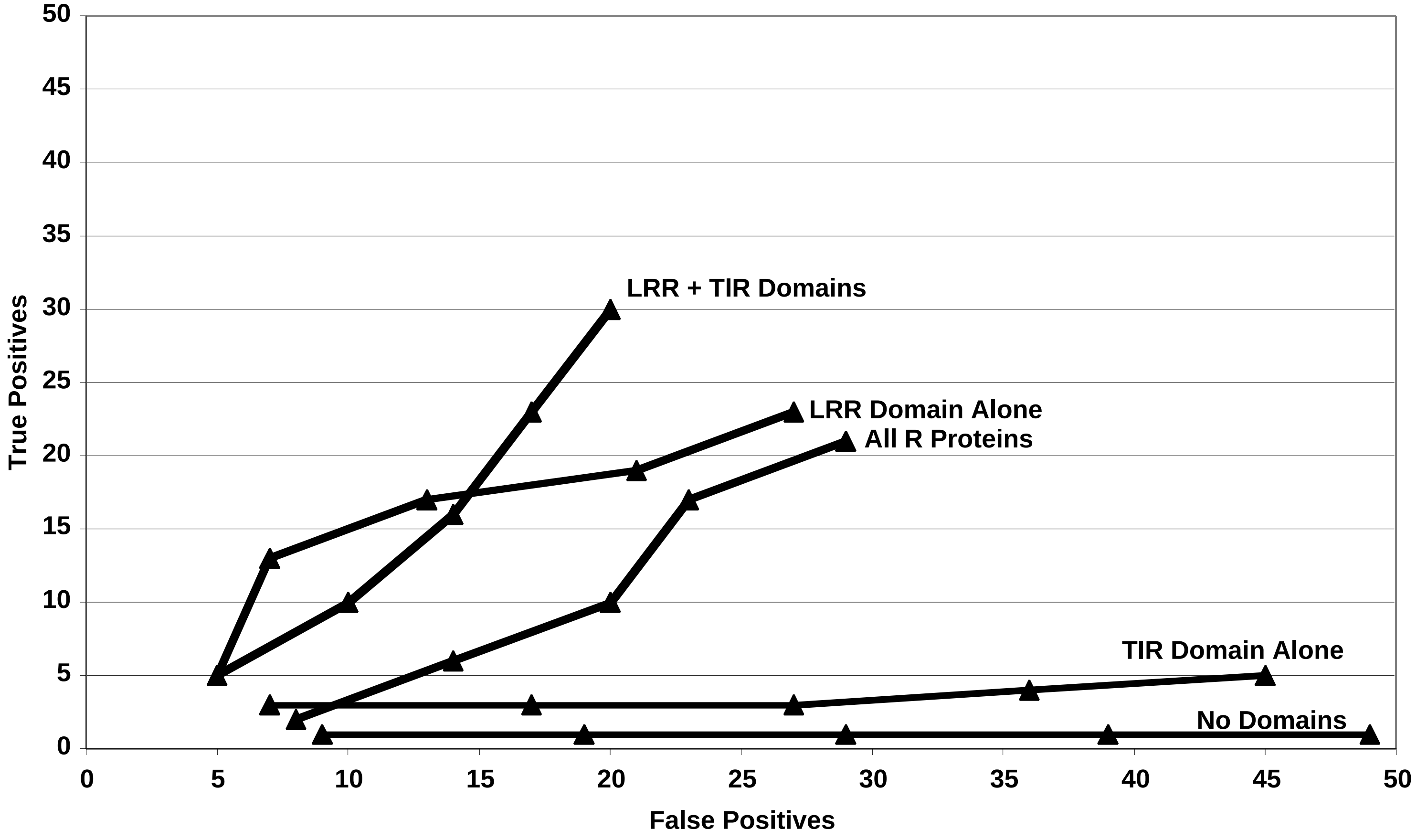


Figure 2: ROC Comparison of ClustalW Consensus Sequences from Domain Specific Subsets of R Proteins: Ability to Identify R Family Domain Members



Appendix: Part E

raining Set	Cobbler Sequence from MOTIF with Embedded Consensus BLOCKS
ll	mlgwlvipwnqiftaacgcfFLSFHGEDVRKTFldalqktmeelkngrddllgrvsieedkglqrlaqvngwlsrvqiveseYASSNWCLDELtgrlcllgycsedcissynygekvskmleevkellskkdfrmvaqeiihkvekkliqttvgldklvemawsslmndeigtlglygmggvgkttlleslnnkfvelesefdvviwvRRVWGMxGKGKTTLilgrlrsdkewereteskkasliynnlerkkfvlllddlwsevdmtkigvppptrengskivfttrstevckhmkadkqikvaclspdeawelfrltvgdiilrshqdipalarivaakchglplalnvigkamscketiqewshainvlnsaghefpgmeerilpilkfsydslkngeiklcflycslfpedseipkekwieywicegfinpnryedggtnhgydiigllvrahllieceltdnvkmhdviremalwinsdfgkqqeticvksgahvrmipndinweivrtmsftctqikkiscrskcpnlstllildnrllvkisnrffrfmpklvvldlsanldliklpeeisnlgslqylnisltgikslpvglkklrkliylnleftgvhgslvgiaatlpnlqvlkffyscvyvddilmkelqdlehlkiltanvkdvtileriqgddrlassirslcledmstprvilstialgglqqlailmcniseiridweskerrelspteilpstgspgfkqlstvyinqlegqrdlswllyaqnlkklevcwspqieeiinkekgmnitklhrdivvpfgnledlalrqmadlteicwnyrtlpnlrksyindcpklpedifvpllpekspsrffff
RR + TIR	masssssrsrtwrYDVFPSFRGEDVRKTFLSHLrkqfsyngismfndqsiersqtivpaltgAIRESRIAIVILSKNYASSSWCLDELVEIMKCredigqivmtvfygvdpsDVRKQTGEFGKAFKKTCKGKTEEKKQKWKEALNDVANIAGEDSRNWDNEAEMIEKiardvsnklnatiswdfedmvgieahlqkmqsllhldyedgaRMVGIWGPPGIGKTTIARALYsrlsssfqltcfmenirgsynsgldeyglklrlqeqllskvlnhdgirinhlgaipERLKNKKVLIILDDVDELEQLEALAGETDWFGPGSRIIVTTQDKELLKAHdvnkkyhvdfptreeackifctyafrrsfapDGFMELAREVVELAGNLPLGLKVLGSYLRGKSKEEWEEQLPRLETSLDrkidgvlrvgydhlceddqflylliafffnyvdddhvkamlvednldvklglktlayksliqisaegnivmhkllqrvgreaiqrqeptkrrilidareicdvlrygkgtsnvsgisfdtsdmsevtisddafkrlhdlrflkvtksrydgkyrmhipagiefpcllrlLHWxxYPWKCLPSEFMPENLVELxMxxSKLEKLWsgtqslrnlknmdlgwspnlkelpdltnatnledlnlnsceslveipssfshlhklknlwmsycinlqvipahmnlvslervtmtgcsrfrkipvisthinyldiahntefevvhasialwcrlhylnmsynenfmglthlpmsltqlilrysdieripdcikalhqlfsldltgcrrlaslpelpgslldleaedcesletvfsplhtprallnftncfklggqarraiirrrseiigkallpgrevpaefdhrakgnsltiilngyrpsydfiqylvcvvispnqeitkisdsstllchtngyifpsyeevyigavskcrkehlfifrsgyylnvdpsgasreivfefssksqdfdiiecgvkiwtaqsiergylvfeddneikhddhtnrvrghykasnvdyksvsrkrprktdlkleiprrrf
RR	maegvvsfgvqklwallnreserlngideqvdglkrqlrglqsllkdadakkhgsdrvrnfledvkdlvfdaediiesyvlnklrgegkgvknhvrrlacfltdrhkvasdiegitkriskvigemqslgiqqqiidggrslslqdiqreirqtfpnssesdlvgveqsveelvgpmveidniqVIGISGMGGIGKTTLARQLFNDDDVKRhfdgfawvcvsqqftqkhvwqrilqelrphdgeilqmdeytiqgklfqlletKRYLIVLDDVWKEEDWDrikevfprkrgwkmlltsrnegvglhadptclsfrarilnpkeswklferivprrneteyeemeaigKKMVKKCNGLPLAIKVLGGLLSNKhtasewkrvsenigaqivgksclddnslnsvyrilslsyedlptdlKHCFLYLGHFPEDYKIktrtlysywaaegiydgltildsgedyleelvrrnlviaeksnlswrlklcqmhdmmrevciskakvenflqiikvptststiiaqspsrsrrltvhsgkafhilghkkkvrsllvlglkedlwiqsasrfqslpllrvldlssvkfeggklpssigglihlrflslhqavvshlpstirnlklmlylnlhvaigvpvhvpnvlkemlelrylslpldmhdktklelgdlvnleylwcfstqhssvtdllrmtklrffgvsfserctfenlssslrqfrkletlsfiysrktymvdyvgefvldfihlkklslgvhlskipdqhqlpphiahiyllfchmeedpmpilekllhlksvelrrkafigrrmvcskggfpqlralqiseqseleewiveegsmpclrdliihscekleelpdglkyvtslkelkiegmkrewkeklvgedyykvqhipdvqffncddeqre
IR	mssatatynydVFLSFRGVDVRRSFISHlykelvgrdirtfkddkelengqmispelilAIEESRIAVVIISKNYASSTWCLDELLKIMdiqknkgsitvmpifygvnpchlrrqigdvaeqfkkhearekdlEKVQKWRRALaaladisgdcsgeddsklvdviadkiskelmivtrisngrnlvgidkhmnelnrlmdlnsnkgkRMVGIWGTAGSGKSTIARYVYqtscqhfdshcflgnvkricqgnyfeshlhkefldniqgensskqslkkqkvllvaddvdkleqldalagdfsgfgpgsvviittkdkqllisygiqlvyeaefltfqkfcrsfrslafkkrddisaafewalyi

Appendix: Part F

Figure 3: ROC Comparison of BLOCKS Consensus Sequence from Domain Specific Subsets of R Proteins: Ability to Identify R Family Members

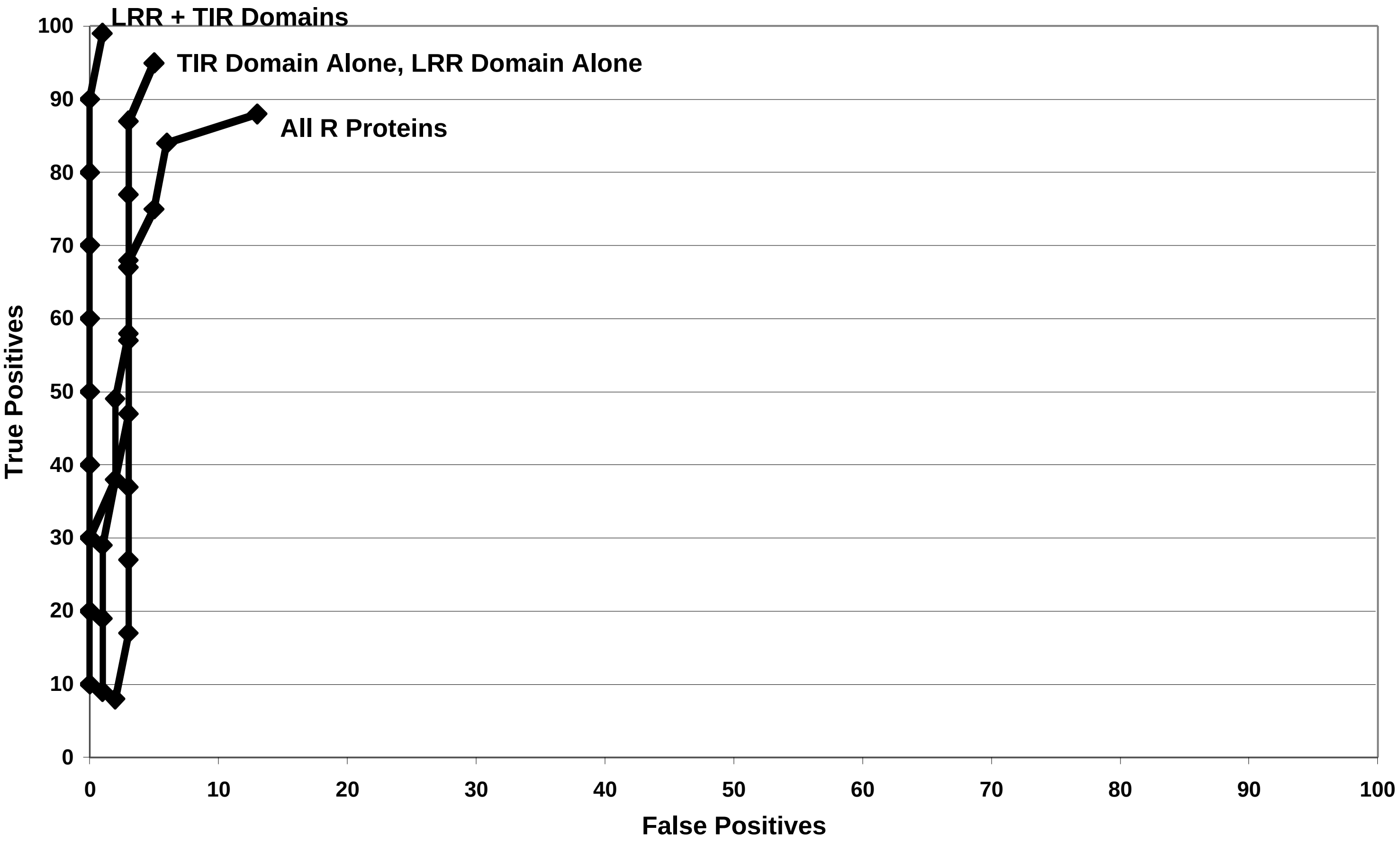
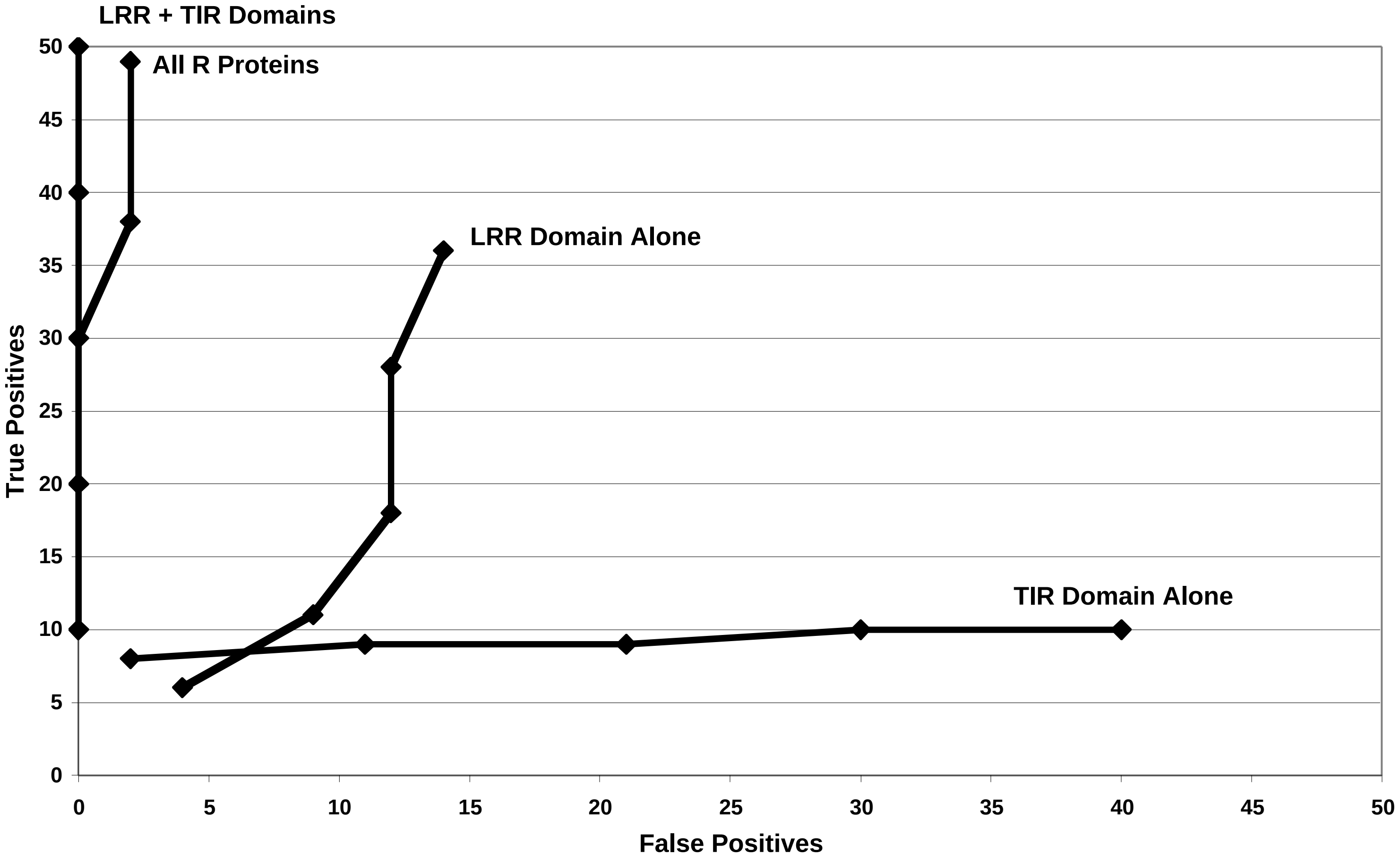


Figure 4: ROC Comparison of BLOCKS Consensus Sequences from Domain Specific Subsets of R Proteins: Ability to Identify R Family Domain Members



Appendix: Part G

Figure 5: ROC Comparison of Methods Used to Identify LRR Domain Members

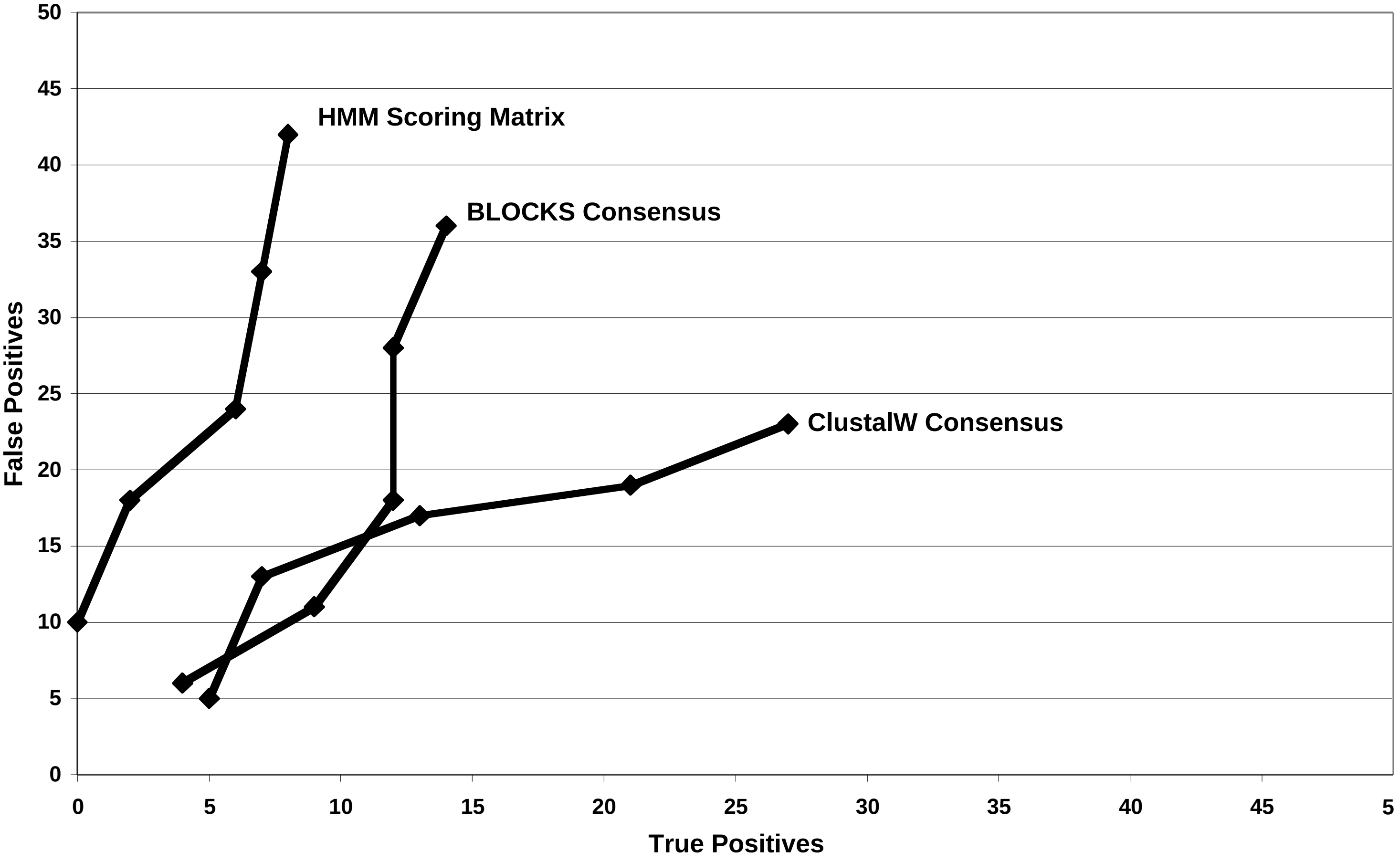
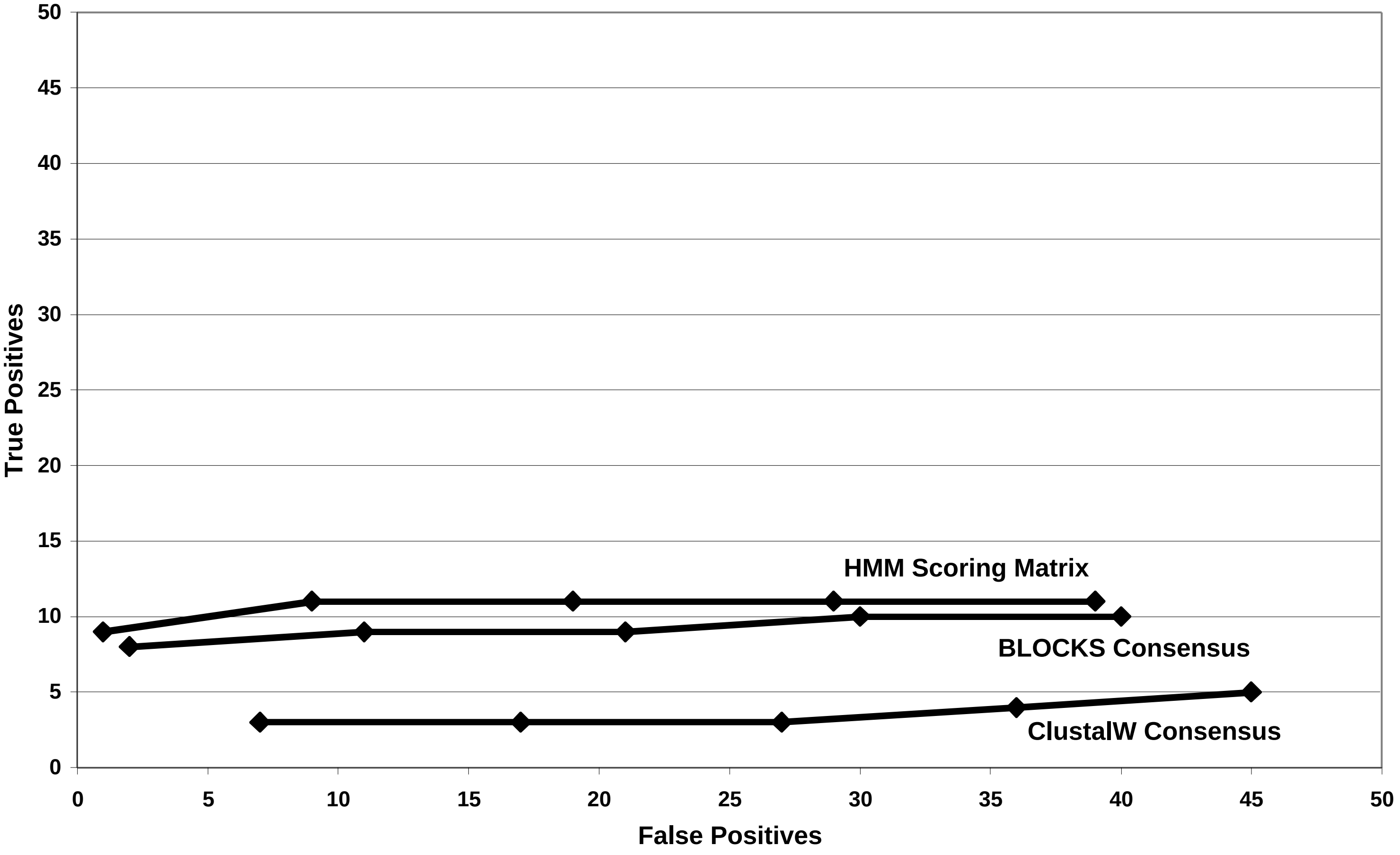


Figure 6: ROC Comparison of Methods Used to Identify TIR Domain Members



Appendix: Part H

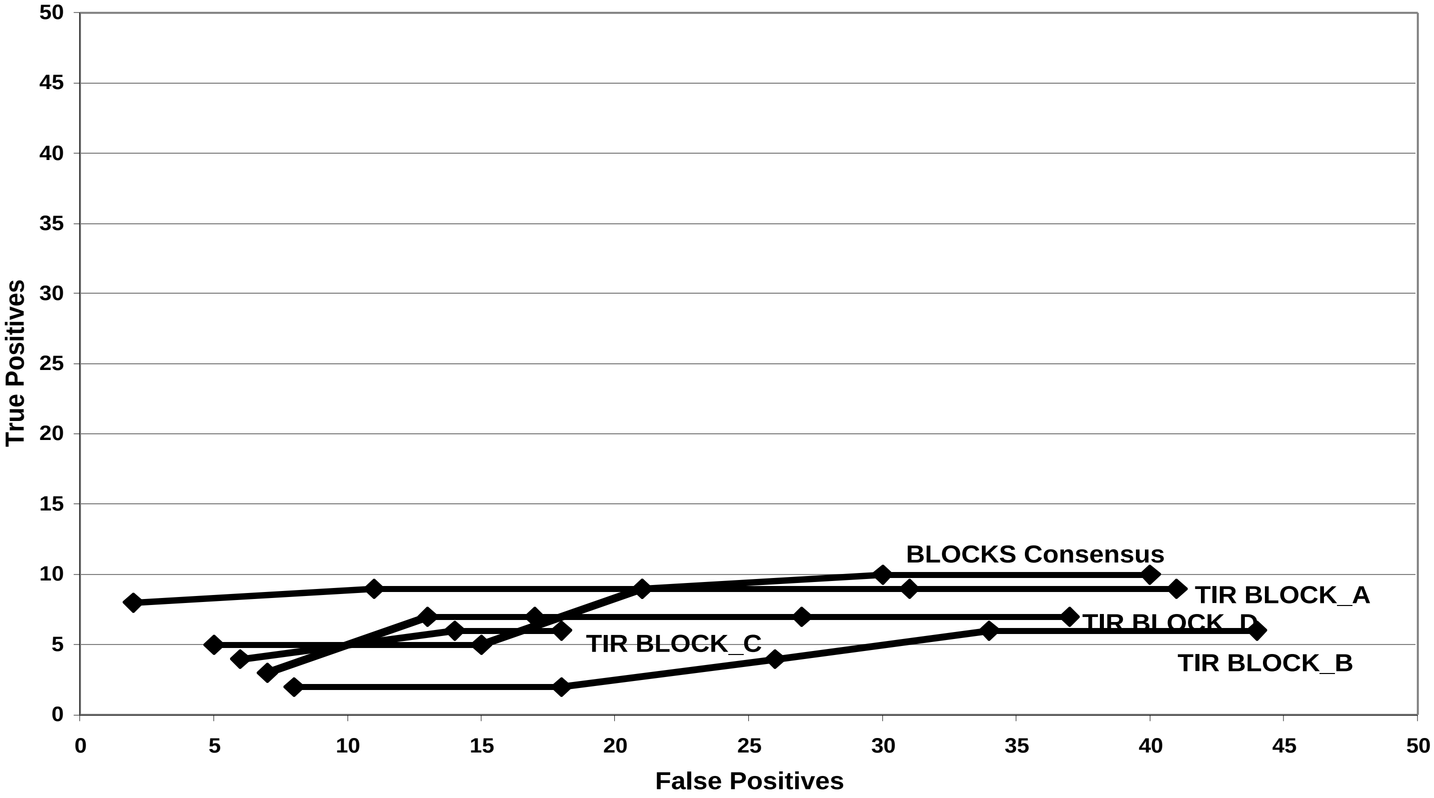
BLOCKS generated for TIR Subgroup by BLOCKmaker:

```
    ID   TIR; BLOCK
    AC   TIR____A; distance from previous
    block=(4,67)
    DE   family
    BL   DRF motif=[10,0,17] motomat=[1,1,-10]
    width=17 seqs=12
    Atg70              (  11) VFINFRGKDIRHGFVSH  37
    Atg71              (  17) VFLSFRGVDTRQTIVSH  38
    Atg72              (  12) VFLSFRREDTGRTFVSH  38
    Atg73              (  12) VFLSFRGPDTRRKFISF  30
    Atg74              (  11) VFLSFRGLDTRRNFISF  25
    Atg75              (  11) VFLSFRGVDTRRNFISF  24
    Atg76              (  68) VFPSFHGADVRKSFLSH  42
    Atg77              (  20) VEAETLVSDLRSSFSEN  93
    Atg79              (  11) SLLLGREVDVFLSFCCQ 100
    Atg80              (  16) VFINFRGKDLRNGFLSF  35
    Atg81              (   5) VFLNFSGEDVRGTFLNH  41
    Atg82              (  14) VFLSFQGLDTRRTFVSH  27
    //
    ID   TIR; BLOCK
    AC   TIR____B; distance from previous
    block=(22,37)
    DE   family
    BL   SVS motif=[10,0,17] motomat=[1,1,-10]
    width=30 seqs=12
    Atg70              (  57) RIEEATIALVILSPRYGESKWCLEELTTIM  83
    Atg71              (  65) AIQTSWFAVVILSENYATSTWCLEELRLIM  58
    Atg72              (  61) AINESRIAVVVISENYVSSVLCLDVLAKII  58
    Atg73              (  60) AIEDSRFAVVVVSVNYAASSWCLDELVKIM  42
    Atg74              (  59) AIEESKFAVVVVSVNYAASPWCLDELVKIM  36
    Atg75              (  59) AIEESKFAVVVVSVNYAASPWCLDELVKIM  36
    Atg76              ( 115) AIRGSRVAIVFLSRKYASSSWCLNELALIM  61
    Atg77              (  59) GIRESKVAVVVISQSYAISAQCLNELQTIV  67
    Atg79              (  65) ALEESRVAVVMTSTTKPCSVGFLEELLVIL 100
    Atg80              (  62) RIQESRVAVVIFSKDYTSSEWCLDELAEIK  57
    Atg81              (  52) AIRESRITVVVFSKNYSSSTWCLNELLQVY  72
    Atg82              (  62) TIGESKVAVVLISVNYASSPLCLDSLLKIL  60
    //
    ID   TIR; BLOCK
    AC   TIR____C; distance from previous
    block=(30,44)
    DE   family
    BL   EWA motif=[10,0,17] motomat=[1,1,-10]
    width=10 seqs=12
    Atg70              ( 129) EEMEKWQVAL  88
    Atg71              ( 138) EKVSKWRRAL  48
    Atg72              ( 131) RTVNRWRDAL  94
    Atg73              ( 134) EKVLKWRQAL  50
    Atg74              ( 132) EKVASWRRAL  44
    Atg75              ( 132) EKVASWRRAL  44
    Atg76              ( 187) DEIGRWRHAL 100
```

```
Atg77              ( 131) EKVQAWMIAL  86
Atg79              ( 125) EKAPSWRTAL  72
Atg80              ( 136) ERTQKWQEAL  91
Atg81              ( 124) ACVYQWRRAL  93
Atg82              ( 133) EKVQTWRQAL  55
//
ID   TIR; BLOCK
AC   TIR____D; distance from previous
block=(29,322)
DE   family
BL   GIG motif=[10,0,17] motomat=[1,1,-10]
width=21 seqs=12
Atg70              ( 303) SVLGIINTVRSGEGREDDCVK  74
Atg71              ( 213) HMIGIWGMGGIGKSTIAKCLY  30
Atg72              ( 275) RTIGIWGFQGVGKTTLAECVF  33
Atg73              ( 208) RMVGIWARGGSCRSALAKYVY  30
Atg74              ( 208) RVVGIWARGYNGRSALAKYVY  24
Atg75              ( 208) RVVGIWARGYNGRSALAKYVY  24
Atg76              ( 262) RMIGIWGPPGIGKTSIARVLF  39
Atg77              ( 227) RLIGICGQGGVGKTTLARYVY  29
Atg79              ( 204) RTIGIWGSAGVGKTTLARYIY  27
Atg80              ( 175) IQKALWQIAMKGNPKVESNSK 100
Atg81              ( 175) EFVGIEDHIAAMNSVDKRXXX  65
Atg82              ( 465) RVVGIWGTGGIGKTTLSRYAY  27
```

Appendix: Part I

Figure 7: ROC Comparison of E-matrices formed from BLOCKS of the TIR Training Set: Ability to Identify TIR Subdomain Members


True Positives
BLOCKS Consensus
TIR BLOCK_A
TIR BLOCK_D
TIR BLOCK_C
TIR BLOCK_B
False Positives