Summer 2010

Computational methods for the discovery and analysis of genes and other functional DNA sequences

Cyriac Kandoth

Department: Computer Science

COMPUTATIONAL METHODS FOR THE DISCOVERY AND ANALYSIS OF GENES AND OTHER FUNCTIONAL DNA SEQUENCES

by

CYRIAC KANDOTH

A DISSERTATION

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

2010

Approved by:

Fikret Ercal, Co-Advisor
Ronald L Frank, Co-Advisor
Jennifer Leopold
Sriram Chellappan
Sanjay Madria

PUBLICATION DISSERTATION OPTION

This dissertation consists of the following three articles that were published, or submitted for publication, as mentioned below. Each paper has been prepared in the style required by their respective journals.

Pages 9 to 28 - "Validation of an NSP-based (negative selection pattern) gene family identification strategy" was published in *BMC Bioinformatics* 2008, 9 (Suppl 9): S2, as part of the peer-reviewed proceedings from the fifth annual conference of the MidSouth Computational Biology and Bioinformatics Society, MCBIOS 2008.

Pages 29 to 41 - "Automation of an NSP-based (negative selection pattern) gene family identification strategy" was published in *Intelligent Engineering Systems through Artificial Neural Networks*, 18: 319-326, as part of the proceedings from Artificial Neural Network in Engineering, ANNIE 2008.

Pages 42 to 62 - "A Framework for Automated Enrichment of Functionally Significant Inverted Repeats in Whole Genomes" was accepted for an oral presentation at the seventh annual conference of the MidSouth Computational Biology and Bioinformatics Society, MCBIOS 2010, and accepted for publication in *BMC Bioinformatics* as peer-reviewed proceedings from the same conference.

ABSTRACT

The need for automating genome analysis is a result of the tremendous amount of genomic data. As of today, a high-throughput DNA sequencing machine can run millions of sequencing reactions in parallel, and it is becoming faster and cheaper to sequence the entire genome of an organism. Public databases containing genomic data are growing exponentially, and hence the rise in demand for intuitive automated methods of DNA analysis and subsequent gene identification. However, the complexity of gene organization makes automation a challenging task, and smart algorithm design and parallelization are necessary to perform accurate analyses in reasonable amounts of time. This work describes two such automated methods for the identification of novel genes within given DNA sequences. The first method utilizes negative selection patterns as an evolutionary rationale for the identification of additional members of a gene family. As input it requires a known protein coding gene in that family. The second method is a massively parallel data mining algorithm that searches a whole genome for inverted repeats (palindromic sequences) and identifies potential precursors of non-coding RNA genes. Both methods were validated successfully on the fully sequenced and well studied plant species, *Arabidopsis thaliana*.

ACKNOWLEDGMENTS

I am extremely grateful to both Dr. Fikret Ercal and Dr. Ronald L. Frank for their invaluable guidance throughout my term at Missouri S&T. Dr. Ercal, for guiding me through the politics of academia, and for always making time to help troubleshoot a problem. Dr. Frank, for sharing his knowledge and experience in Evolutionary Biology, and his seemingly limitless patience in doing so. I am also grateful to Professors Jennifer Leopold, Sriram Chellappan, and Sanjay Madria, for their valuable guidance and feedback, not just as the members of the advisory committee.

I will not forget the many moments spared by other faculty members, graduate students, friends and colleagues at the Department of Computer Science, in particular Dr. Daniel Tauritz, Dr. Bruce McMillin, Dr. Ali Hurson, Clayton Price, Waraporn Viyanon, Lee Leong, Dylan McDonald, Roy Cabaniss, Krzysztof Charatonik, Rhonda Grayson, and Dawn Davis. They, and many others have all contributed to my work in many little ways, whether they know it or not.

Drs. Ercal, Frank, and Hurson, all deserve gratitude for sponsoring or authorizing my trips to Bioinformatics conferences and events. Finally, I am indebted to my parents and siblings for their unwavering support throughout my research efforts.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Page

LIST OF TABLES

Page

PAPER 1

PAPER 2

PAPER 3

1. INTRODUCTION

The study of genetics has become so intertwined with computing, that its progress in the last two decades approached the predictions of Moore's law, which estimates a doubling in electronic transistor densities every eighteen months. The computing power and storage capacities required by bioinformatics is unprecedented. The human genome project, a decade long international effort to determine the complete DNA sequence of human beings, involved at least a thousand processors running in parallel, several supercomputers, tens of gigabytes of memory, and hundreds of terabytes of storage [1]. The outcome in 2001, of a parallel effort by Venter et. al. [1] was a 14.8 billion base-pair consensus sequence generated from 27,271,853 short sequence reads from the DNA of five individuals. This accounts for approximately 2.91 billion base pairs per individual, which is known as the genome of that individual, their complete genetic information.

Today, a decade after sequencing the first human genome, progress in DNA sequencing technology and genome assembly has taken a breathtaking pace. The cost of DNA sequencing has dropped 14,000-fold between 1999 and 2009. Whole genome sequencing, as opposed to sequencing specific regions of interest, is expected to drop below $1,000 per genome in the next three to five years making it possible to sequence and store the complete genome sequences for each of us. Along with appropriate privacy protections, these can be stored in our medical records where it will be quickly available to guide prevention strategies or medication choice [2].

Once a genome is fully sequenced, the next step is to find and map onto it the various functional and hereditary subsequences, the genes. This is known as gene identification which not only involves finding the functionally significant subsequences within a genome, but also their functionality in the cellular level, and how that functionality is activated and regulated within the organism. The complexity of gene organization is such that there could be multiple such subsequences contributing to a certain gene, or a single subsequence involved in the expression of multiple genes. There are therefore a multitude of methods for gene identification within a genome since these

methods are typically highly specific to certain genes or types of genes. A protein-coding gene, for example, is first transcribed (copied) from its corresponding subsequences in genomic DNA (also known as the coding region) into a messenger RNA (mRNA) sequence which carries the blueprint for the amino-acids that make up the protein. Based on surrounding molecular interactions and their own chemical properties, the amino-acids arrange themselves in a unique three-dimensional structure that gives the protein a specific cellular level functionality. Ignoring the effects of surrounding molecular interactions, it is thus reasonable to assume a direct correlation between genomic subsequence and the functionality of a protein. In an evolutionary context, the more critical a protein is to an organism's survival, the more likely that the corresponding genomic sequence will remain unchanged over the generations.

Despite the significance of proteins in cellular processes and pathways, only around 1.5% of the human genome contains protein coding genes while the remaining includes non-coding RNA genes and regulatory sequences that get transcribed into ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (miRNA), small interfering RNA (siRNA), etc. and a substantial portion of unidentified DNA with no known function. This "dark matter" of the genome could be the remnants of genes from ancestral organisms that no longer provide any sort of functionality, or it could also contain hundreds of novel genes waiting to be identified. This dissertation is a study and advancement of the techniques for identification of novel genes from genomic DNA and transcribed RNA sequences.

2. REVIEW OF LITERATURE

2.1. GENE IDENTIFICATION

Computational analysis of genomic data, using automated techniques developed for bioinformatics, came about as a result of necessity. The progress of the Human Genome Project is a good example [3]. When it started in the 1990s, the identification of genes was a slow and tedious process. It usually involved matching known genes from other species against sequences from the human genome. By 2001, the entire 3 billion nucleotides were sequenced, but the processes used in locating the genes became numerous and elaborate, requiring both "wet lab" techniques (in a laboratory) and "in-silico" methods (on a computer) because of the enormity of the genomic data. Most of these techniques were based on biological or chemical properties that were specific to the genes identified and limited the number of novel genes found. These limitations led to a hunt for more general non-specific techniques that make use of the high resolution DNA sequences from various genomes stored at public-access databases.

In order to find general identifiable rules within the enormous complexity of gene organization, and subsequently apply them in gene identification techniques, the best option is to use evolutionary rationales [4]. Unfortunately, this does not entirely widen the generalization because different classes of organisms have evolved very differently and developed their own evolutionary mechanisms. So, a technique based on evolutionary rationales is not entirely universal, but is sufficiently so. For instance, most plants have evolved using the same common mechanisms and several global gene identification techniques based only on these mechanisms can be formulated. But such techniques cannot be directly extended for mammal genomes since they have evolved very differently from plants.

The first paper presented in this dissertation is a Perl-based validation of a novel gene family identification technique, where families of protein coding genes related by functionality can be automatically grouped together, with the potential for identification of novel related genes. It was developed by Frank et. al. in 2006 and does not require

whole genome sequences [5]. Instead, it utilizes expressed sequence tags (ESTs) which are short, unedited, randomly selected single-pass sequences that can be easily and cheaply obtained from messenger RNA, the precursors to proteins. Since they are obtained from mRNA, it is already known that they contribute to a protein's three-dimensional structure and hence is functionally significant enough to be a gene. In fact, ESTs are primarily used as markers to pin-point the location of specific protein coding genes within a genome. The technique developed by Frank et. al. is a novel repurposing of these readily available sequences.

The second paper presents a Perl and C++-based automation and validation of the same gene family identification technique that is now fully automated allowing for rapid automated identification of novel genes and gene families. This allowed the technique to be automated and validated on 10 gene families of Arabidopsis thaliana, a well studied and fully sequenced plant species, not including the 5 gene families that were already manually validated in the first paper. Both papers also review a host of other identification techniques for protein coding genes and gene families.

2.2. NON-CODING GENES

Genes can be broadly categorized as protein coding or non-coding depending on whether the final product is a protein or not. Proteins have long been known as critical to cellular level functionality in organisms. But in the last decade, non-coding RNA (ncRNA) sequences have been found to be essential to regulation of gene expression. They were once considered insignificant in comparison to protein coding sequences. But since then, a variety of new types of ncRNA genes have been discovered, each of them revealing new biological roles and cellular mechanisms. Therefore, the identification of ncRNA has significant importance to the biological and medical community. To date, the genomes of numerous organisms have been fully sequenced, making it possible to perform genome-wide computational analyses.

The third paper in this dissertation studies various methods of identifying non-coding RNA sequences from genomic DNA and proposes a fast automated framework

for the identification of non-coding genes. Two novel filters for the identification of non-coding RNA among genomic inverted repeats were implemented and the ability to attach additional Perl-based filters to the automation was included. This enabled the framework to be easily expanded upon as new identifiable characteristics and categories of non-coding RNA are discovered.

2.3. DYNAMIC PROGRAMMING

Dynamic programming (DP) is perhaps the most popular programming method in Bioinformatics. Hundreds of related problems like sequence comparison, gene recognition, and RNA secondary structure prediction are solved by ever new variants of dynamic programming [6]. DP is a technique that can potentially search an exponential search space in polynomial time by dividing the problem into subproblems such that subproblem optimality holds i.e. the optimal solution to the problem can be defined in terms of optimal solutions to its subproblems. In Bioinformatics applications, DP is most commonly used for sequence alignment. The Needleman-Wunsch [7] algorithm, published in 1971, was the first application of dynamic programming for sequence alignment and comparison. As a global alignment algorithm, it attempts to find the best alignment between two given sequences along their entire lengths. This limits its application to sequence comparison between two sequences of relatively equal lengths. A local alignment algorithm, on the other hand, attempts to find the two longest aligning subsequences within two given sequences. This would allow alignment and comparison of very short sequences against much larger sequences, which is a much more common challenge posed by genomics. The Smith-Waterman algorithm [8], is a variation of Needleman-Wunsch that can perform such local alignments. Like the Needleman-Wunsch, it is guaranteed to find the optimal alignment with respect to the scoring system being used, which includes a substitution matrix (numeric scores for matches and penalties for mismatches) and a gap-penalty scheme (penalizes, but allows insertion of gaps between characters in a sequence for an overall better alignment).

The first two papers in the dissertation implicitly use DP in their contig assembly and sequence comparison stages. Well established DP-based techniques and programs are

used. However, in the third paper, a fast and parallelizable DP-based algorithm was designed and implemented to enable the detection of partially palindromic sequences (inverted repeats) in genomic DNA. This was accomplished by creating a unique variation of the Smith-Waterman dynamic programming algorithm.

3. CONCLUSION

The first two papers demonstrate a novel method of gene family identification that performs well in distinguishing contigs that represent real genes from contigs that are artifacts of sequence assembly. Almost every ORF predicted to be a distinct gene family member, matches a known protein coding sequence in the same gene family. A distinctive feature of this method is its use of only EST data which is easily sequenced from cheaply available complementary DNA (cDNA), instead of more expensive high quality whole-genome sequences. However, high-quality whole genome sequences are quickly becoming available in public-access databases and ESTs are used simply as markers to protein coding genes that can then be mapped onto whole genome data. This diverted research in automated gene identification techniques towards finding non-coding RNA genes that cannot normally use ESTs from protein coding messenger RNA as markers.

The third paper presents a Perl-based framework for the multitude of non-coding RNA identification methods, with the computationally intensive preprocessing steps parallelized using C++. The study reveals that partially symmetric inverted repeats (IRs) though abundant in genomic DNA, are easily distinguishable from the IRs of known non-coding RNA and can be filtered out using simple generic characteristics of ncRNA. It is then reasonable to assume that more accurate filters that are highly specific to certain kinds of ncRNA will retain a smaller final list of IRs that can then be further analyzed using wet lab techniques such as *northern blotting* [9] to identify novel ncRNA genes. However, the final set of IR candidates presented in the paper is too large to warrant further analysis, and additional filters are required to enrich the set with those that are most likely to be functional ncRNA. The irScan software framework is designed to be easily expandable with such additional filtering criteria by anyone with experience in the Perl programming language and a suitable filter that characterizes ncRNA.

BIBLIOGRAPHY

[1] J.C. Venter, et. al., "The Sequence of the Human Genome," *Science* 291, 1304, 2001

[2] F. Collins, "Has the revolution arrived?," *Nature* 464, 674-675, 2010

[3] Human Genome Program, U.S. Department of Energy: *DOE Human Genome Program Contractor-Grantee Workshop IV*, 1994

[4] C. Kandoth, "A Quantitative Study of Gene Identification Techniques Based on Evolutionary Rationales," MS Thesis, University of Missouri - Rolla, 2007

[5] R.L. Frank, A. Mane, F. Ercal, "An Automated Method for Rapid Identification of Putative Gene Family Members in Plants," *BMC Bioinformatics*, 7:S19, 2006

[6] R. Giegerich, "A systematic approach to dynamic programming in bioinformatics," *Bioinformatics*, 16(8):665-77, 2000

[7] S.B. Needleman, C.D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology* 48 (3):443–53, 1970

[8] T.F. Smith, M.S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology* 147: 195–197, 1981

[9] Y. Meng, F. Huang, Q. Shi, J. Cao, D. Chen, J. Zhang, J. Ni, P. Wu, M. Chen: "Genome-wide survey of rice microRNAs and microRNA-target pairs in the root of a novel auxin-resistant mutant," *Planta* 230:883-898, 2009

PAPER

I. VALIDATION OF AN NSP-BASED (NEGATIVE SELECTION PATTERN) GENE FAMILY IDENTIFICATION STRATEGY

Ronald L Frank[1], Cyriac Kandoth[2], Fikret Ercal[2]

[1]Department of Biological Sciences, Missouri University of Science and Technology, Rolla MO, 65401, USA

[2]Department of Computer Science, Missouri University of Science and Technology, Rolla MO, 65401, USA

ABSTRACT

Background - Gene family identification from ESTs can be a valuable resource for analysis of genome evolution but presents unique challenges in organisms for which the entire genome is not yet sequenced. We have developed a novel gene family identification method based on negative selection patterns (NSP) between family members to screen EST-generated contigs. This strategy was tested on five known gene families in Arabidopsis to see if individual paralogs could be identified with accuracy from EST data alone when compared to the actual gene sequences in this fully sequenced genome.

Results - The NSP method uniquely identified family members in all the gene families tested. Two members of the FtsH gene family, three members each of the PAL, RF1, and ribosomal L6 gene families, and four members of the CAD gene family were correctly identified. Additionally all ESTs from the representative contigs when checked against MapViewer data successfully identify the gene locus predicted.

Conclusions - We demonstrate the effectiveness of the NSP strategy in identifying specific gene family members in Arabidopsis using only EST data and we describe how this strategy can be used to identify many gene families in agronomically important crop species where they are as yet undiscovered.

1. BACKGROUND

A significant proportion of genes that make up a genome are part of larger families of related genes resulting from duplications of individual genes [1], genomic segments [2], or even whole genomes ([3, 4]. The accumulation of mutations in duplicates (paralogs) leads to either loss of function for one (death), altered function (subfunctionalization), or a new function (neofunctionalization). The study of the molecular processes by which functional innovation occurs interests not only evolutionary biologists, but protein engineers and medical and agricultural biologists. A clearer understanding of the extent to which gene families contribute to the selected traits in our most important crop species will help guide decisions regarding future improvements. Many studies are aimed at the diversity of function, expression, and regulation among gene family members in many species [reviewed in 5]. Others have spawned computational methods to analyze and predict the evolution of gene families in a phylogenetic context [6] or determine clinically relevant sites in a protein sequence where amino acid replacements are likely to have a significant effect on phenotype, including those that may cause genetic diseases [7].

Therefore, it is not surprising that research aimed at the identification of specific gene families and their constituent members has proliferated in last few decades. Although experimental approaches using degenerate primers for PCR and oligofingerprinting [8] and cDNA library screening [9] generally produce the most reliable results, they can be time consuming and labor-intensive. Many strategies of gene family identification are computational approaches that take advantage of database mining and analysis tools to increase the capability and improve the efficiency of dealing with large amounts of sequenced data. Naturally, if a significant amount of a genome is sequenced computational methods can be somewhat more exhaustive in their search and identification [10, 11, 12, 13, 14, 15]. However, complete genomic data is available for only a limited number of species. Expressed sequence tags (ESTs) on the other hand, are short, unedited, randomly selected single-pass sequences. They can be easily and

inexpensively obtained directly from cDNA libraries. Although they were initially used for human gene discovery [16, 17], exponential growth in the generation and accumulation of EST data for many diverse organisms has occurred in the last decade. The National Center for Biotechnology Information (NCBI) has a database for ESTs from over 1300 species totaling more than 48 million ESTs (as of December 14, 2007). Sixty-three species have more than 100,000 ESTs in the database making computational analyses more fruitful but complex. Because the number of ESTs in databases is increasing, computational techniques, including BLAST and its variants for comparative analysis and CAP3 [18] for sequence assembly, can be used to speed up gene or gene family identification processes and improve the feasibility of extracting meaningful information from a large and redundant database [19] when parameters are properly selected. These EST-based gene family identification strategies are valuable in species without fully sequenced genomes [20, 21]. Caution must be exercised when assembling contigs from EST sequences because contigs not representative of real genes can result from chimera formation during cDNA cloning, errors in single-pass high-throughput sequencing of ESTs, or similarity between protein domains of unrelated sequences. Our group has developed a simple but novel method using evidence of negative selection pressure during divergence of the coding sequences to filter artifactual contigs from those potentially representing actual gene family members. Molecular evolution researchers studying divergence between well-characterized orthologs or paralogs often employ an estimation of the number of synonymous base substitutions per synonymous site versus the number of nonsynonymous base substitutions per nonsynonymous site [22, 23]. A dS/dN ratio > 1 indicates purifying or negative selection (lower fitness) that tends to keep amino acid sequences the same if changes were deleterious. A ratio equal to 1 indicates changes that were neutral to fitness, while a dS/dN ratio < 1 would indicate adaptive or positive selection presumably because natural selection favored the amino acid changes. Differences between contigs that are artifactual should be proportionally distributed among synonymous and nonsynonymous sites, whereas differences between contigs that represent paralogs will often exhibit negative selection, dS/dN > 1.

We understand that negative selection may not be uniform over entire coding regions even assuming that purifying selection was at work in a given gene family. And not all gene families will exhibit negative selection between members. However, we believe that the number of gene families that can be detected by this approach is significant. Evidence has been found for a model whereby complementary deleterious mutations in regulatory elements between duplicate genes partitions the original function resulting in sub-functions [24]. It has also been discovered that the number of shared regulatory elements between duplicated genes in yeast decreases with evolutionary time [25]. The age of the duplicates was estimated by the accumulation of synonymous substitutions in the coding regions. Clearly, some forms of subfunctionalization can occur by changes in regulatory elements whereby some degree of negative selection has maintained protein function. Coding regions of paralogs that have subfunctionalized via changes to regulatory elements should exhibit a bias toward synonymous substitutions. In plants, a significantly greater proportion of genes belong to gene families than in animals or other major taxa [26]. Either gene duplication events have been more common in plants, or more duplicates have been retained during the evolutionary history of plants [27]. If this is the case, there should exist a significant number of gene families that can be identified by a bias toward synonymous substitutions when contigs are assembled from a significantly large database of ESTs. We have demonstrated previously that a simple strategy to detect negative selection patterns (NSP) among assembled ESTs provides a good screen for real versus artifactual contigs [28]. We have modified the filtering criterion to an empirically determined dS/dN threshold and decided to test the negative selection pattern (NSP) strategy on an EST database for which a large percentage of the ESTs have already been mapped to a fully-sequenced genome, *Arabidopsis thaliana*.

In this article we demonstrate the NSP strategy and report how well it was able to identify ESTs representing distinct family members in a genome where it is testable.

2. METHODS

2.1. GENE FAMILY IDENTIFICATION BY NSP METHOD

The five gene families chosen to validate the NSP strategy were, eukaryotic release factor 1 (*RF1*), ribosomal protein L6 (*L6*), cinnamyl alcohol dehydrogenase (*CAD*), phenylalanine ammonia-lyase (*PAL*), and an FtsH protease (*FtsH*). One member of the selected gene family was chosen as query for a tblastn search of the *Arabidopsis thaliana* dbEST. All hits with an E value < $1x10^{-10}$ (maximum of 150 sequences) were selected and the resulting EST sequences were assembled using a contig assembly program (AssemblyLIGN, Oxford Molecular) with 100% match over a minimum 100 nucleotide overlap. The largest open reading frame greater than 100 codons was identified in each resulting non-singleton contig (MacVector, Accelrys). Open reading frames were translated and the resulting polypeptides aligned using ClustalX. The PAL2NAL program [29] produced a codon alignment of all contig open reading frames, and the SNAP program [30] at http://www.hiv.lanl.gov was used to calculate dS/dN for all pairwise comparisons of contig open reading frames.

The empirically determined threshold for dS/dN was set to 2.00 and all pairs of contigs with a dS/dN ratio greater than this were classified as putative paralogs. A graph was constructed using vertices to represent contigs, and edges to represent whether pairs of contigs are putative paralogs. In such a graph, the largest fully connected sub-graph (the maximum clique) will be made up of vertices that represent markers (contigs) to the members of the same gene family as the query protein. This sub-graph was determined using a brute-force algorithm. A brute-force algorithm works by checking every possible sub-graph for connectedness. This operation is computationally expensive, and its time complexity increases exponentially, as the factorial of the number of vertices. Fortunately, the contigs that these vertices represent are usually quite few in number. Some contigs can also be excluded from the graph since they do not pass the dS/dN threshold to pair with any other contig. This can be observed in Figure 1 where only 5 pairwise comparisons of contigs obtained a dS/dN of more than 2.00.

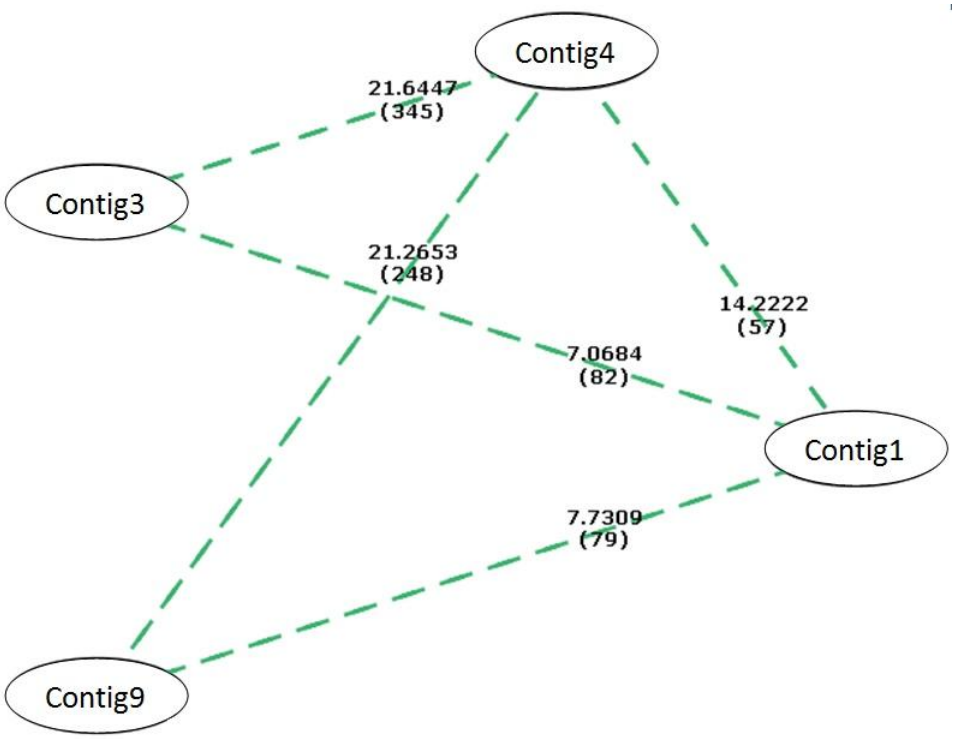

Figure 1. Graph representing potential paralogs with dS/dN >= 2

Edges are labeled with the dS/dN ratios, followed by the number of substitutions (Sd+Sn) seen. An edge indicates dS/dN>=2.00; No edge indicates dS/dN<2.00 OR dS/dN=NA

Figure 1 shows the dS/dN ratios between contigs generated using the PAL1 gene as the protein query. Note that there are two maximum cliques in this graph. When there are more than one maximum cliques, we arbitrarily choose one of these cliques. The contigs represented by the vertices belonging to this clique are then identified as members of the same gene family. Any vertices that are not part of this clique are classified as either a possible marker to a distinct gene, or as a duplicate marker to an identified gene family member (in the maximum clique) which was different enough to be assembled into a different contig. In the case of Contig3 and Contig9 from Figure 1, it was found that these contigs were extremely similar to each other. They were later found to be duplicate markers to the PAL4 gene.

The representative contig for each putative gene family member identified was then compared to each of the actual gene family member sequences (NCBI) using bl2seq [31] to determine how closely contigs filtered through NSP represented the gene family. Either all or a subset of ESTs from each NSP-identified contig were checked on MapViewer (NCBI) to determine if ESTs from the same contig mapped to different gene family members or if ESTs from different contigs mapped to the same gene family member.

3. RESULTS

3.1. PHENYLALANINE AMMONIA-LYASE GENE FAMILY

The tblastn search of using *AtPAL1* protein as query resulted in ESTs and contigs reported previously [28]. Here we report the refinement of using dS/dN ratio rather than a tally of 1st, 2nd, and 3rd position differences as well as the MapViewer results that validate the accuracy of gene family member identification. The dS/dN data for the assembled contigs are shown in Table 1 and the resulting maximum clique graph indicating putative paralog relationships is shown in Figure 1. The 2.0 dS/dN threshold was established empirically by dS/dN measurements among actual members of several Arabidopsis gene families. Pairwise comparison of contigs 1, 3, and 4 with the actual Arabidopsis gene sequences, reported previously [28] indicate that these three contigs represent AtPAL1, AtPAL4, and AtPAL2, respectively with greater than 96% similarity. The contigs selected by NSP as representative of real gene family members were further validated by checking to see if each EST comprising a single contig is assigned to a single gene family member on the Arabidopsis genome by NCBI MapViewer. Table 2 shows that all ESTs that comprise a single contig map to the same gene locus and confirms that contigs 1, 3, and 4 represent the PAL1, PAL4, and PAL2 genes of Arabidopsis, respectively.

Table 1. dS/dN calculations for phenylalanine ammonia-lyase (PAL) contigs

Comparison		Sd[a]	Sn	S	N	ps	pn	ds	dn	ds/dn	ps/pn
Contig1	Contig4	38.50	18.50	62.83	219.17	0.61	0.08	1.27	0.09	14.22	7.26
Contig1	Contig3	42.17	39.83	65.17	216.83	0.65	0.18	1.49	0.21	7.07	3.52
Contig1	Contig9	42.33	36.67	65.50	216.50	0.65	0.17	1.48	0.19	7.73	3.82
Contig1	Contig6	52.50	138.50	63.33	197.67	0.83	0.70	NA	0.00	NA	1.18
Contig1	Contig8	53.50	138.50	63.33	197.67	0.84	0.70	NA	0.00	NA	1.21
Contig1	Contig7	45.17	153.83	62.83	210.17	0.72	0.73	2.39	2.80	0.85	0.98
Contig4	Contig3	211.83	133.17	286.33	985.67	0.74	0.14	3.22	0.15	21.64	5.48
Contig4	Contig9	143.00	105.00	191.67	639.33	0.75	0.16	3.94	0.19	21.27	4.54
Contig4	Contig6	152.50	490.50	201.50	665.50	0.76	0.74	NA	0.00	NA	1.03
Contig4	Contig8	80.33	225.67	99.83	314.17	0.80	0.72	NA	0.00	NA	1.12
Contig4	Contig7	65.00	233.00	93.17	326.83	0.70	0.71	2.00	2.25	0.89	0.98
Contig3	Contig9	1.50	14.50	197.17	633.83	0.01	0.02	0.01	0.02	0.33	0.33
Contig3	Contig6	160.50	486.50	206.83	660.17	0.78	0.74	NA	0.00	NA	1.05
Contig3	Contig8	81.83	222.17	102.83	311.17	0.80	0.71	NA	0.00	NA	1.11
Contig3	Contig7	70.50	228.50	96.00	324.00	0.73	0.71	2.90	2.11	1.37	1.04
Contig9	Contig6	150.00	454.00	196.17	613.83	0.76	0.74	NA	0.00	NA	1.03
Contig9	Contig8	81.33	220.67	103.17	310.83	0.79	0.71	NA	0.00	NA	1.11
Contig9	Contig7	71.67	229.33	96.33	323.67	0.74	0.71	3.61	2.17	1.66	1.05
Contig6	Contig8	2.00	4.00	108.33	305.67	0.02	0.01	0.02	0.01	1.42	1.41
Contig6	Contig7	68.33	200.67	97.50	301.50	0.70	0.67	2.04	1.64	1.25	1.05
Contig8	Contig7	68.33	199.67	97.50	301.50	0.70	0.66	2.04	1.61	1.27	1.06

SNAP output results for all 21 pairwise comparisons of 7 contigs in which an ORF was identified. A ds/dn value greater than 2.00 was chosen as threshold to indicate contigs that potentially represent distinct gene family members.

a – See http://www.hiv.lanl.gov for explanation of abbreviations and calculations.

For the following four additional NSP-identified gene families, only the validating data is shown, not the dS/dN data or maximum clique graphs.

Table 2. MapViewer locus for ESTs of NSP generated contigs

Putative gene family	Gene group by NSP	Contig	EST accession	MapViewer locus	MapViewer Gene Name
CAD	GeneB	contig3	CK121258	AT4G39330	AtCAD1
			CB074210	AT4G39330	AtCAD1
	GeneC	contig1	BP561562	ELI3-1	AtCAD4
			BP796450	ELI3-1	AtCAD4
			CD530744	ELI3-1	AtCAD4
RF1	GeneA	contig1	AV823314	ERF1-3	AteRF1-3
	GeneB	contig3	AV822373	ERF1-2	AteRF1-2
			BP803175	ERF1-2	AteRF1-2
			Z18188	ERF1-2	AteRF1-2
	GeneC	contig6	AV825957	ERF1-1	AteRF1-1
			BE845168	ERF1-1	AteRF1-1
PAL	GeneA	contig1	8720101	PAL1	AtPAL1
			8736225	PAL1	AtPAL1
	GeneB	contig3	8722848	AT3G10340	AtPAL4
			8723431	AT3G10340	AtPAL4
			8728745	AT3G10340	AtPAL4
			8730514	AT3G10340	AtPAL4
			9780248	AT3G10340	AtPAL4
			9788228	AT3G10340	AtPAL4
	GeneC	contig6	8690351	PAL2	AtPAL2
			8724245	PAL2	AtPAL2
			8725529	PAL2	AtPAL2
			19869024	PAL2	AtPAL2
			19869200	PAL2	AtPAL2
			37426635	PAL2	AtPAL2
	GeneC	contig8	9786707	PAL2	AtPAL2
			37426640	PAL2	AtPAL2
	GeneC	contig4	8719100	PAL2	AtPAL2
			14580232	PAL2	AtPAL2
			19855615	PAL2	AtPAL2
			49165014	PAL2	AtPAL2
			59667557	PAL2	AtPAL2
L6	GeneA	contig1	5761694	AT1G18540	AtL6A
			8724065	AT1G18540	AtL6A
			19802678	AT1G18540	AtL6A
	GeneB	contig4	19868834	AT1G74060	AtL6B
			23303389	AT1G74060	AtL6B
	GeneC	Contig6	8714872	AT1G74050	AtL6C
FtsH	GeneA	contig1	AV518555	VAR2	AtFtsH2
			AV558102	VAR2	AtFtsH2
			AV800962	VAR2	AtFtsH2
			BP785237	VAR2	AtFtsH2
	GeneB	contig6	BP626558	FTSH8	AtFtsH8

Individual ESTs of representative contigs for putative gene family members of the 5 Arabidopsis families tested were located to a specific locus by NCBI MapViewer.

3.2. RIBOSOMAL PROTEIN L6 GENE FAMILY

AtRPL6A protein was used as query for the tblastn search of *A. thaliana* dbEST yielding 150 EST sequences that assembled into eight contigs ranging from 449 to 953 bases and 2 to 36 ESTs each. Following ORF identification the 28 pairwise codon alignments and subsequent dS/dN values were analyzed to sort contigs into putative gene family members. From that analysis contig1, contig3 and contig8 were assigned to putative geneA, contig2, contig4, and contig5 to geneB, and contig6 to geneC. Table 3 shows that each of these contig groups identified, by greater than 98% similarity, a different member of the Arabidopsis ribosomal protein *L6* gene family when aligned to the actual gene sequences.

Table 3. Percent similarity of NSP generated contigs against ribosomal L6 genes

	GeneA			GeneB			GeneC
	contig1	contig3	contig8	contig2	contig4	contig5	contig6
AtL6A	100	98	98	82	83	82	84
AtL6B	83	83	82	99	99	99	93
AtL6C	84	84	83	93	93	93	99

Representative contigs for 3 putative gene family members, GeneA, GeneB, and GeneC, identified by the NSP method were aligned with actual Arabidopsis gene family members and percent similarity determined.

In Table 2 for ribosomal protein L6 it can be seen that all ESTs from the same contig as well as all ESTs from the same gene grouping are assigned to the same gene locus. Also, in no instances did ESTs belonging to different gene groupings by NSP ever map to the same gene locus.

3.3. CINNAMYL ALCOHOL DEHYDROGENASE GENE FAMILY

The tblastn search of using *AtCAD5* protein as query resulted in 150 EST sequences. The ESTs assembled into eight contigs ranging from 592 to 1248 bases and 2 to 21 ESTs each. Following ORF identification the 28 pairwise codon alignments and subsequent dS/dN values were analyzed to sort contigs into putative gene family

members as described above (data not shown). The eight contigs assorted into four groups based on their negative selection pattern with each other contig. These four groups were arbitrarily designated GeneA represented by contig4, contig6, and contig8, GeneB, represented by contig3, contig5, and possibly contig7, GeneC represented by contig1, and GeneD represented by contig2.

The results of the comparison of representative contigs to the actual gene sequences for the CAD gene family of Arabidopsis are shown in Table 4. Each contig group identified, by greater than 99% similarity, a different member of the CAD gene family. MapViewer analysis for the CAD gene family (Table 2) shows that all ESTs from the same contig are assigned to the same gene locus, and no ESTs belonging to different contigs map to the same gene locus. Contigs validated by alignment to actual genes but not shown in Table 2 are comprised of ESTs that have not yet been mapped to specific loci by MapViewer.

Table 4. Percent similarity of NSP generated contigs against CAD genes

	GeneA	GeneB	GeneC	GeneD
	contig8	contig3	contig1	contig2
AtCAD-1	NSS[a]	99	NSS	NSS
AtCAD-2	99	NSS	NSS	NSS
AtCAD-3	NSS	NSS	78	82
AtCAD-4	NSS	76	100	87
AtCAD-5	NSS	72	84	100
AtCAD-6	79	NSS	NSS	NSS
AtCAD-7	NSS	78	72	NSS
AtCAD-8	NSS	NSS	NSS	NSS
AtCAD-9	NSS	NSS	NSS	NSS

Representative contigs for 4 putative gene family members, GeneA, GeneB, GeneC, and GeneD identified by the NSP method were aligned with actual Arabidopsis gene family members and percent similarity determined.

a – No significant similarity as returned by bl2seq program.

3.4. RELEASE FACTOR 1 GENE FAMILY

AtRF1-3 protein was used as query for the tblastn search of *A. thaliana* dbEST yielding 109 EST sequences that assembled into six contigs ranging from 591 to 930 bases and three to 19 ESTs each. Following ORF identification the 15 pairwise codon alignments by the NSP program resulted in three contigs exhibiting NSP. These were arbitrarily assigned as contig1 representing geneA, contig3 representing geneB, and contig6 representing geneC. Each of these contigs identified, by greater than 97% similarity, a different member of the Arabidopsis RF1 gene family when aligned to the actual gene sequences (Table 5). MapViewer results again show that ESTs comprising NSP-selected contigs are unambiguous in the gene locus to which they have been assigned (Table 2).

Table 5. Percent similarity of NSP generated contigs against release factor genes

	GeneA	GeneB	GeneC
	contig1	contig3	contig6
AtRF1-1	82	83	99
AtRF1-2	88	97	83
AtRF1-3	99	85	82

Representative contigs for 3 putative gene family members, GeneA, GeneB, and GeneC, identified by the NSP method were aligned with actual Arabidopsis gene family members and percent similarity determined.

3.5. FTSH PROTEASE GENE FAMILY

The TBLASTN search of using *AtFtsH8* protein as query resulted in 150 EST sequences. The ESTs assembled into six contigs ranging from 526 to 1217 bases and 2 to 33 ESTs each. Following ORF identification the 15 pairwise alignments by the NSP program resulted in two contig groups exhibiting NSP. Contig1 and contig5 represent geneA, and contig3 and contig6 represent geneB. Each of these contig groups identified, by greater than 97% similarity, a different member of the Arabidopsis *FtsH* gene family when aligned to the actual gene sequences, as shown in Table 6. MapViewer results again

show that ESTs comprising NSP-selected contigs are unambiguous in the gene locus to which they have been assigned (Table 2).

Table 6. Percent similarity of NSP generated contigs against FtsH genes

	GeneA		GeneB	
	contig1	contig5	contig3	contig6
AtFtsH1	NSS	71	73	NSS
AtFtsH2	100	97	86	83
AtFtsH3	NSS	78	70	NSS
AtFtsH4	NSS	79	78	NSS
AtFtsH5	NSS	73	73	NSS
AtFtsH6	72	73	69	77
AtFtsH7	NSS	68	73	NSS
AtFtsH8	88	85	99	100
AtFtsH9	NSS	68	NSS	NSS
AtFtsH10	NSS	75	74	NSS
AtFtsH11	NSS	77	77	NSS
AtFtsH12	NSS	NSS	NSS	NSS

Representative contigs for 2 putative gene family members, GeneA and GeneB, identified by the NSP method were aligned with actual Arabidopsis gene family members and percent similarity determined.

4. DISCUSSION

It has been observed for some time that contig assembly from EST sequences can produce artifactual sequences resulting from relatively high error in EST sequences, chimeras generated in cDNA cloning, and regions of highly conserved domains interspersed in related genes. Therefore, it is necessary that strategies involving the generation of contigs from ESTs employ some criterion for either eliminating unauthentic coding regions or selecting for authentic ones. We have found that contigs representing gene families where the paralogous coding regions have been constrained by negative (purifying) selection pressure can be identified by screening for amino acid substitution patterns indicative of such (NSP, Negative Selection Patterns). However, if differences between contigs are artifacts no pattern among codon positions should be exhibited. If no

negative selection pattern is detected we do not conclude that the contigs necessarily represent the same gene. Our goal is only to identify contigs that represent different genes of the same family. We do not expect that all members of a particular family will be detectable by this method. Other members may be identified with iterative searches using previously identified contigs.

To illustrate that this method can identify members of a gene family with some accuracy using only EST data we tested it on five well-characterized gene families in Arabidopsis. Each case resulted in successful identification of one to three additional gene family members distinct from the member used as initial query. Of the eight initial contigs generated from EST hits when *AtCAD5* was used as query the NSP strategy identified those representing *AtCAD1*, *AtCAD2*, and *AtCAD4*, in addition to one representing *AtCAD5* (Table 4). Moreover, each of these contigs exhibited less than 87% similarity to other actual members of the gene family. No contigs generated at the parameters specified in the assembly program represented *AtCAD3*, 6, 7, 8 or 9. This could be the result of relative expression levels of those genes, limits on the necessary similarity between gene family members, or limitations on the method which are discussed elsewhere [28]. Similarly, when ribosomal protein *L6A* was used as query the NSP strategy identified contigs accurately representing all three genes of the family, *L6A*, *L6B*, and *L6C* (Table 3). Furthermore, all three members of the *RF1* gene family were accurately represented by NSP-screened contigs (Table 5), as were *AtFtsH2* and *AtFtsH8* of that 12-member gene family (Table 6). We previously reported the accurate identification of *AtPAL1*, 2, and 4 of phenylalanine ammonia-lyase gene family and show here further validation that the contigs identified the appropriate gene family members.

In addition, we were able to show that all the ESTs of a single contig defined the same actual gene family member according to MapViewer (Table 2), i.e., all ESTs of a single contig mapped to the same locus, and perhaps more importantly, no ESTs from different contigs of the same gene family ever mapped to the same locus. This would suggest that although the initial assembly of related ESTs may indeed generate non-valid

contigs, screening by NSP allows one to determine which contigs represent real gene loci.

A limitation to the NSP strategy is the fact that only paralogs that exhibit purifying selection can be identified and that selection pattern must be evident in the portion of the coding region reconstructed by contig assembly, roughly the 3' two-thirds of the protein by our experience. For this reason the NSP strategy in it current phase will only identify a subset of gene families. However, when we consider that estimates of the number of gene families in a plant species may be 10-12,000 [32], that subset may comprise a significant portion in which NSP can detect two to three additional family members. Our goal is to broaden the NSP approach to identify as many gene families as possible without sacrificing the accuracy reported here. We have already automated the four basic steps, 1) BLAST collection of related ESTs, 2) contig assembly, 3) ORF identification, and 4) NSP screening of contigs, such that the input is a query protein of a potential gene family member and the output is contigs representing at least two gene family members. Since the query can be an orthologous sequence, we are currently working on identifying, in *Glycine max*, every gene family for which at least one member has been identified in another plant species. The specific objectives for accomplishing this are to:

1) use all known *Glycine max* mRNAs as queries to identify other family members, if any.
2) use mRNAs from related species as queries to identify gene families not identified above.
3) use Arabidopsis gene families as queries (currently about 1000 gene families in TAIR).
4) use other Arabidopsis genes, not currently associated with a family as queries to identify potential genes existing as a family in soybean but not so in Arabidopsis.
5) use all *Glycine max* ESTs not included in contigs from above searches in clustering experiments to potentially identify novel gene families.

Objectives 1-4 above are identical in protocol. They differ only in the species of origin for the protein query. There are currently about 1350 known *Glycine max* gene sequences in the NCBI database, mostly mRNA sequences but some genomic. Some of these already represent multiple members of the same gene family (e.g. glycinin and conglycinin seed storage proteins, uricase, ascorbate peroxidase, lipoxygenase, rubicase small subunit, phosphoenolpyruvate carboxylase, etc) [33]. Objective 1 will use all known genes of soybean as queries to identify other members of the gene family. Objective 2 involves genes from species more closely related to soybean than Arabidopsis. These include other eurosids I and particularly other legumes that have significant sequence data available such as *Pisum sativum, Phaseolus vulgaris,* and *Medicago truncatula.* Objective 3 will involve queries chosen from Arabidopsis genes that are known to exist as part of a gene family. Currently, The Arabidopsis Information Resource (TAIR) has genomic, coding region, and amino acid sequence data for 996 gene families comprised of 8,331 genes. Objective 4 will use as initial queries all remaining Arabidopsis genes not already identified in soybean and not associated with a gene family in TAIR. It is possible that of the remaining 16,000 genes of Arabidopsis there could be some that are associated with a family in *Glycine max*. Objective 5 does not start with a query sequence but rather a set of ESTs clustered by similarity to each other. Several clustering algorithms could be used for this, UniGene (at NCBI), PACE [34], or one developed in our laboratory several years ago. The majority of UniGene clusters are annotated with "strongly similar to," moderately similar to," or "weakly similar to" gene or protein functions of other organisms. Others are labeled simply as "Transcribed locus" to indicate that they represent RNA sequences that do not show similarity to any currently known gene or protein (Build #31 has 6812 such clusters). We have run a few of these clusters through the NSP strategy and found that some will generate contigs that indicate the cluster may represent ESTs from distinct members of a gene family. More work in this direction will allow us to expand the strategy to include identification of yet undiscovered gene families.

5. CONCLUSION

Although the NSP strategy is not a global gene family identification protocol, our tests on the Arabidopsis EST dataset indicate that it performs well in distinguishing contigs that represent real genes from contigs that are artifacts. Every EST tested, from contigs that NSP predicted to be distinct gene family members, mapped to the appropriate gene in Arabidopsis. Further expansion of the strategy to clustered ESTs eliminating the need for individual query sequences and further automation of the steps will allow the identification of a significant proportion of gene families with reliable accuracy.

6. COMPETING INTERESTS

The authors declare that they have no competing interests.

7. AUTHORS' CONTRIBUTIONS

RLF participated in the conception and design of the study, carried out the gene family identification via NSP including BLAST, contig assembly, ORF identification, alignment and dS/dN analysis, and drafted the manuscript. CK developed scripts to construct graphical output of dS/dN results and performed all genome locus identification studies using MapViewer. FE participated in the conception, design, and development of the computational aspects of data generation. All authors read and approved the final manuscript.

8. REFERENCES

1. Taylor JS, Raes J: Duplication and Divergence: The Evolution of New Genes and Old Ideas. *Annual Review of Genetics* 2004, 38:615-643
2. Van de Peer Y, Meyer A: Large-Scale Gene and Ancient Genome Duplications. In *The Evolution of the Genome*. Elsevier Academic Press; 2005:329-368

3. Gregory TR, Mable BK: (2005) Polyploidy in Animals. In *The Evolution of the Genome*. Elsevier Academic Press; 427-517
4. Tate JA, Soltis DE, Soltis PS: Polyploidy in Plants. In *The Evolution of the Genome*. Elsevier Academic Press; 2005:371-426
5. Taylor JS, Raes J: Small-Scale Gene Duplications. In *The Evolution of the Genome*. Elsevier Academic Press; 2005:289-327
6. Bie T, Cristianini N, Demuth JP, Hahn MW: CAFE: A Computational Tool for the Study of Gene Family Evolution. *Bioinformatics Applications Note* 2006, 22(10):1269-1271
7. Gaucher EA, De Kee DW, Benner SA: Application of DETECTER: An Evolutionary Genomic Tool to Analyze Genetic Variations to the Cystic Fibrosis Gene Family. *BMC Genomics* 2006, 7(44)
8. Fuchs T, Malecova B, Linhart C, Sharan R, Khen M, Herwig R, Shmulevich D, Elkon R, Steinfath M, O'Brien JK, Radelof U, Lehrach H, Lancet D, Shamir R: DEFOG: A Practical Scheme for Deciphering Families of Genes. *Genomics* 2002, 80(3):295-302
9. Schwarz RS, Hodes-Villamar L, Fitzpatrick KA, Fain MG, Hughes AL, Cadavid LF: A Gene Family of Putative Immune Recognition Molecules in the Hydroid Hydractinia. *Immunogenetics* 2007, 59(3):233-246
10. Albert VA, Soltis DE, Carlson JE, Farmerie WG, Wall PK, Ilut DC, Solow TM, Mueller LA, Landherr LL, Hu Y, Buzgo M, Kim S, Yoo M-J, Frohlich MW, Perl-Treves R, Schlarbaum SE, Bliss BJ, Zhang X, Tanksley SD, Oppenheimer DG, Soltis PS, Ma H, dePamphilis CW, Leebens-Mack JH: Floral Gene Resources from Basal Angiosperms for Comparative Genomics Research. *BMC Plant Biology* 2005, 5(1):5
11. Cannon SB, Young ND: OrthoParaMap: Distinguishing Orthologs from Paralogs by Integrating Comparative Genome Data and Gene Phylogenies. *BMC Bioinformatics* 2003, 4(35)
12. Liu Q: Computational Identification and Systematic Analysis of the ACR Gene Family in Oryza Sativa. *Journal of Plant Physiology* 2006, 163(4):445-451

13. Nakano T, Suzuki K, Fujimura T, Shinshi H: Genome-Wide Analysis of the ERF Gene Family in Arabidopsis and Rice. *Plant Physiology* (Rockville) 2006, 140(2):411-432
14. Tian C, Wan P, Sun S, Li J, Chen M: Genome-Wide Analysis of the GRAS Gene Family in Rice and Arabidopsis. *Plant Molecular Biology* 2004, 54(4):519-532
15. Zhang G, Wang H, Shi J, Wang X, Zheng H, Wong GK, Clark T, Wang W, Wang J, Kang L: Identification and Characterization of Insect-Specific Proteins by Genome Data Analysis. *BMC Genomics* 2007, 8(93)
16. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC: Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science* 1991, 252(5013):1651-6
17. Adams MD, Dubnick M, Kerlavage AR, Moreno RF, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC: Sequence Identification of 2375 Human Brain Genes. *Nature* 1992, 355:632-634
18. Huang X, Madan A: CAP3: A DNA Sequence Assembly Program. *Genome Research* 1999, 9(9):868-877
19. Nagaraj SH, Gasser RB, Ranganathan S: A Hitchhiker's Guide to Expressed Sequence Tag (EST) Analysis. Briefings in Bioinformatics 2006, 8(1):6-21
20. Brown S, Chang JL, Sadée W, Babbitt PC: A Semiautomated Approach to Gene Discovery through Expressed Sequence Tag Data Mining: Discovery of New Human Transporter Genes. *AAPS PharmSci* 2003, 5(1)
21. Retief JD, Lynch KR, Pearson WR: Panning for Genes: A Visual Strategy for Identifying Novel Gene Orthologs and Paralogs. *Genome Research* 1999, 9:373-382
22. Nei M, Gojobori T: Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 1986, 3(5):418-426
23. Ota T, Nei M: Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol Biol Evol* 1994, 11(4):613-619
24. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of Duplicate Genes by Complementary Degenerative Mutations. *Genetics* 1999, 151:1531-1545

25. Papp B, Pál C, Hurst LD: Evolution of Cis-Regulatory Elements in Duplicated Genes of Yeast. *TRENDS in Genetics* 2003, 19:417-422
26. Lockton S, Gaut BS: Plant Conserved Non-Coding Sequences and Paralogue Evolution. *TRENDS in Genetics* 2005, 21:60-65
27. Shiu SH, Shih MC, Li WH: Transcription Factor Families have Much Higher Expansion Rates in Plants than in Animals. *Plant Physiology* 2005, 139:18-26
28. Frank RL, Mane A, Ercal F: (2006) An Automated Method for Rapid Identification of Putative Gene Family Members in Plants. BMC *Bioinformatics* 7:S19
29. Suyama M, Torrents D, Bork P: PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments. *Nucleic Acids Res* 2006, 34:W609-W612
30. Korber B: HIV Signature and Sequence Variation Analysis. In *Computational Analysis of HIV Molecular Sequences.* Edited by Rodrigo AG, Learn GH Netherlands: Kluwer Academic Publishers, 2000:55-72
31. Tatiana A, Tatusova TL: Blast 2 sequences: A new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 1999, 174:247-250
32. Nelson RT, Shoemaker RC: Identification and Analysis of Gene Families from the Duplicated Genome of Soybean using EST Sequences. *BMC Genomics* 2006, 7(204)
33. Frank RL, Ercal F: Evaluation of *Glycine max* mRNA clusters. *BMC Bioinformatics* 2005, 6(Suppl 2):S7
34. Kalyanaraman A, Aluru S, Kothari S, Brendel V: Efficient clustering of large EST data sets on parallel computers. *Nucleic Acids Res* 2003, 31, 2963–2974

II. AUTOMATION OF AN NSP-BASED (NEGATIVE SELECTION PATTERN) GENE FAMILY IDENTIFICATION STRATEGY

CYRIAC KANDOTH
Computer Science
Missouri S&T
Rolla, Missouri

RONALD L FRANK
Biological Sciences
Missouri S&T
Rolla, Missouri

FIKRET ERCAL
Computer Science
Missouri S&T
Rolla, Missouri

ABSTRACT

Expressed Sequence Tags (ESTs) are short nucleotide subsequences of expressed genes, which can be rapidly generated in large quantities. With improving sequencing technology, the number of ESTs available in open-access databases is exponentially increasing. A method of gene family identification that makes use of this data can be a valuable tool in genome analysis. We have previously demonstrated such a technique which uses negative selection patterns (NSP) between family members to screen out potential paralogs from contigs assembled from ESTs (Frank et al., 2006; Frank et al., 2008). This strategy is now fully automated and tested on 10 gene families in Arabidopsis thaliana to see how the resulting putative paralogs compare with the actual gene sequences in this fully sequenced genome. The automation correctly identifies specific member genes in these families using only EST data. These results suggest that this automated strategy can identify many gene families in species where they are as yet undiscovered.

1. BACKGROUND

Gene duplication is a key evolutionary mechanism because it sets up the foundation for the birth of new genes. Once duplicated, copies of a gene can undergo mutations and diverge to create different genes with possibly related functions. As a result, a significant proportion of genes that make up a genome are part of larger families of related genes. The accumulation of mutations in duplicate genes (paralogs) leads to either loss of function, altered function, or a new function. Many studies are aimed at the

diversity of function, expression, and regulation among gene family members in many species (Taylor and Raes, 2005). Others have spawned computational methods to analyze and predict the evolution of gene families in a phylogenetic context (Bie et al., 2006) or determine clinically relevant sites in a protein sequence where amino acid replacements are likely to have a significant effect on phenotypes.

Expressed sequence tags (ESTs) are short, unedited, randomly selected single-pass sequences. They can be easily and inexpensively obtained directly from cDNA libraries. Because the number of ESTs in publicly available databases is increasing, computational techniques, including BLAST and its variants for comparative analysis and CAP3 (Huang and Madan, 1999) for sequence assembly, can be used to speed up gene or gene family identification processes and improve the feasibility of extracting meaningful information from a large and redundant database. These EST-based gene family identification strategies are valuable in species without fully sequenced genomes (Brown et al., 2003; Retief et al., 1999). Caution must be exercised when assembling contigs from EST sequences because contigs not representative of real genes can result from chimera formation during cDNA cloning, errors in single-pass high-throughput sequencing of ESTs, or similarity between protein domains of unrelated sequences. Our group has developed a simple but novel method using evidence of negative selection pressure during divergence of the coding sequences to filter out such artifactual contigs from those potentially representing actual gene family members. Molecular evolution researchers studying divergence between well-characterized orthologs or paralogs often employ an estimation of the number of synonymous base substitutions per synonymous site (dS) versus the number of nonsynonymous base substitutions per nonsynonymous site (dN) (Nei and Gojobori, 1986; Ota and Nei, 1994). A dS/dN ratio > 1 indicates purifying or negative selection (lower fitness) that tends to keep amino acid sequences the same if changes were deleterious. A ratio equal to 1 indicates changes that were neutral to fitness, while a dS/dN ratio < 1 would indicate adaptive or positive selection presumably because natural selection favored the amino acid changes. Differences between contigs that are artifactual should be proportionally distributed among

synonymous and nonsynonymous sites, whereas differences between contigs that represent paralogs will often exhibit negative selection, dS/dN > 1.

We understand that negative selection may not be uniform over entire coding regions even assuming that purifying selection was at work in a given gene family. And not all gene families will exhibit strong negative selection between members. However, we believe that the number of gene families that can be detected by this approach is significant (Frank et al., 2008). We have demonstrated previously that a simple strategy to detect negative selection patterns (NSP) among assembled ESTs provides a good screen for real versus artifactual contigs (Frank et al., 2006). We have modified the filtering criterion to an empirically determined dS/dN threshold and tested the negative selection pattern (NSP) strategy on an EST database for which a large percentage of the ESTs have already been mapped to a fully-sequenced genome, *Arabidopsis thaliana*. The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org) provides a comprehensive open-access listing of all the known gene families in *Arabidopsis thaliana*. This allowed us to create an automated script that would test our NSP strategy on all known *Arabidopsis* gene families. In this article, we demonstrate how the NSP strategy was able to identify open reading frames (ORFs, in contigs assembled from ESTs) that represent distinct family members.

2. METHODS

2.1. GENE FAMILY IDENTIFICATION USING NSP

From the hundreds of gene families listed in TAIR, 10 distinct gene families were arbitrarily chosen to verify the NSP strategy. One member out of each family was chosen as the query protein sequence for NCBI's tblastn search through the *Arabidopsis thaliana* dbEST database. All the hits that were returned using an Expect threshold of 10 (a maximum of 250 sequences) were fetched (using NCBI Entrez) and assembled using Huang and Madan's CAP3 (1999). The parameters used by CAP3 are crucial in determining the quality of the resulting contigs. Of particular importance, are the three parameters -o (overlap length cutoff), -y (clipping range), and -p (overlap percent identity

cutoff). Ideally, -p should be a 100% and -y should be zero (only a minimum of 6 is allowed by CAP3). However, low-quality ESTs make this impractical. Hence, an automation was devised which continually tried different combinations of -y, and -p until the longest possible contigs could be constructed with the highest possible value of -p. The overlap length cutoff (-o) was kept fixed at its lowest allowed value of 21 bases.

The contigs obtained were nucleotide sequences made up of 5 characters (A, C, G, T, and N). An 'N' in the sequence is meant to indicate an unidentified nucleotide. Although CAP3 truncates its consensus sequences by removing gaps, it was observed that it often does not have enough redundant ESTs to confirm the absence of a gap. In such cases, it inserts an 'N' where there might not be a nucleotide at all. Since the presence of an incorrectly inserted 'N' adversely affects the ClustalW alignments (performed in a later step), we truncate all CAP3's final contigs by removing 'N's. Each contig is then submitted to NCBI's ORF Finder and the largest open reading frame is chosen. If no ORF is found on a contig, it is discarded. The ORFs are then converted to their corresponding amino-acid sequences while still preserving the original nucleotide sequences in separate files. Since the ORFs (and the contigs they came from) are usually shorter subsequences of a full length coding region, each ORF is aligned against the query protein sequence using ClustalW (Larkin et al., 2007) in order to find their positions relative to each other (see Figure 1).

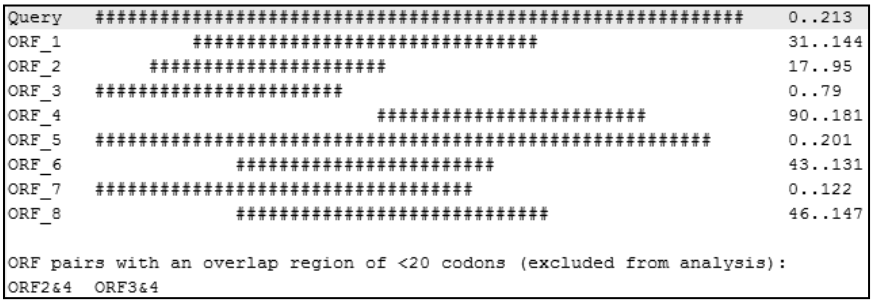

Figure 1. Pair-wise alignment of individual ORFs against the query protein: RPL19A

Based on this multiple pair-wise alignment, it is possible to tell which ORFs belong to the same region and hence might display an NSP against each other. All pairs of ORFs that show an overlap of more than 20 codons are then aligned using ClustalW. All ClustalW alignments are made using their default parameters, except that the scoring matrix is BLOSUM30 and the gap extension penalty is set to 5. These amino-acid alignments are then run through Pal2Nal (Suyama et al., 2006) which converts them into codon alignments using the original nucleotide sequences. Each alignment is then run through the SNAP.pl program (Korber, 2000) which produces the required dS/dN values.

The empirically determined threshold for dS/dN is set to 2.00 and all pairs of ORFs with a dS/dN ratio greater than this are classified as putative paralogs. A graph is constructed using vertices to represent ORFs, and edges to represent whether pairs of ORFs are putative paralogs. In such a graph, the largest fully connected sub-graph (the maximum clique) will be made up of vertices that represent markers (ORFs) to the members of the same gene family as the query protein. This sub-graph was determined using a brute-force algorithm. A brute-force algorithm works by checking every possible sub-graph for connectedness. This operation is computationally expensive, and its time complexity increases exponentially, as the factorial of the number of vertices. Fortunately, the ORFs that these vertices represent are usually quite few in number. Some ORFs can also be excluded from the graph since they do not pass the dS/dN threshold to pair with any other ORF. This can be observed in Figure 2 where only ORF6 and ORF8 (out of 8 ORFs in total) did not score a dS/dN > 2.00 against any other ORF.

Figure 2 shows the dS/dN ratios between contigs generated using the RPL19A gene as the protein query. Note that there are two maximum cliques in this graph. When there are more than one maximum cliques, we select the clique that contains the query (i.e. Query, ORF_5, and ORF_7). If more than one maximum clique contains the query or if none of the maximum cliques contain the query, then we arbitrarily choose one of the maximum cliques. The contigs represented by the vertices belonging to this clique are then identified as members of the same gene family. All the remaining ORFs are

classified as possibly a distinct gene, or as a duplicate marker to an identified gene family member (in the maximum clique). All the above steps (starting with the protein query) are automated in a single Perl script.

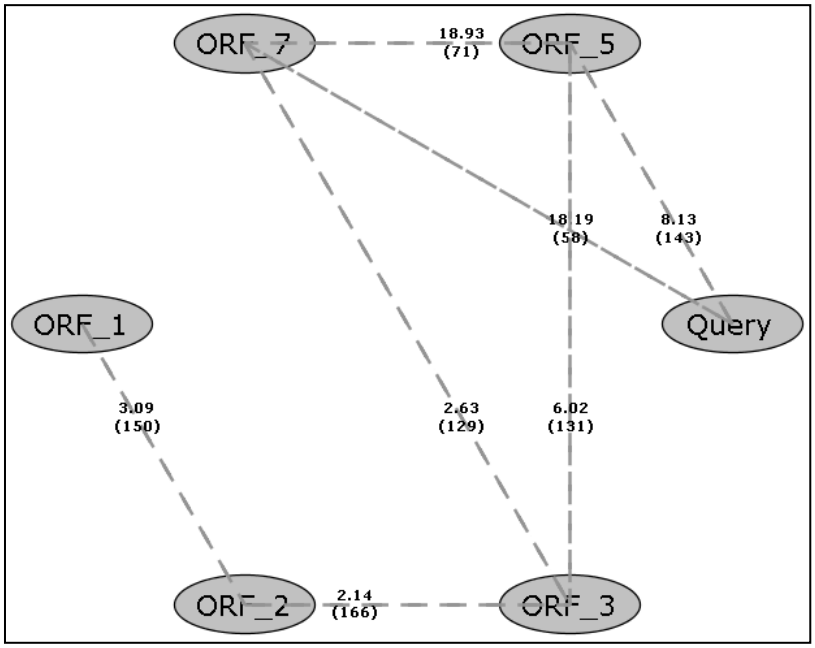

Figure 2. dS/dN values between potential paralogs in the At RPL19 family

The number of substitutions (Sd + Sn) are in parenthesis

2.2. VALIDATION OF THE NSP STRATEGY

The representative ORF for each putative gene family member identified is compared against each of the gene family member sequences in TAIR (full sequences are fetched using NCBI Entrez). This is done by aligning these ORFs against the protein sequence of each gene family member using ClustalW (with the Identity scoring matrix and gap extension penalty of 5). Any ORF that matches a member gene sequence for more than 95% of the ORF's length was considered accurate enough to be used as a marker to that gene. ORFs that were part of a clique but did not match any known gene in TAIR were run through NCBI's blastp search over *Arabidopsis thaliana* proteins (see the

"Conclusions" column in Table 1). Only exact protein matches returned by blastp were considered. If any of these ORFs failed to generate an exact match in blastp, they were either considered artifactual or possibly, a previously unidentified gene family member.

3. RESULTS

Table 1 shows the results of the automation when run on each of the 10 gene families. It shows the query protein used in each case and the CAP3 parameters that were used to assemble contigs. The constituent ORFs of all cliques in the final graph are tabularized along with the gene family members that each ORF identifies. In 9 of 10 gene families (all except the *ABC-type transport-like* family) at least one additional TAIR gene family member was identified.

Figure 1 demonstrates how the contigs assembled are usually only subsequences of the genes they might represent. This particular figure is one of the output files generated by the NSP automation when run for the *60S ribosomal L19 proteins* (AtRPL19) using RPL19A as the query protein. In this family, an ORF was found in every contig assembled by CAP3. It can be seen that two pairs of ORFs did not sufficiently overlap: (ORF2, ORF4) and (ORF3, ORF4). Such pairs are excluded from the latter part of the analysis because they belong to separate regions of the gene and may produce misleading alignments. The remaining pairs are aligned and run through SNAP.pl along with the query sequence too. SNAP.pl uses the dS/dN calculation by Nei and Gojobori (1986).

The 2.0 dS/dN threshold was established empirically by dS/dN measurements among actual members of several Arabidopsis gene families (Frank et al., 2008). Figure 2 shows all the pairs of ORFs for AtRPL19 that passed this threshold. There are two maximum cliques in this graph: (Query, ORF5, ORF7) and (ORF3, ORF5, ORF7). Since the former is the only clique that contains the query, it was chosen as the representative clique for this gene family. ORF5 and ORF7 are thus distinct putative members of the same gene family as the Query protein (RPL19A). After running the validation script,

results showed that 99.17% of ORF5 is identical to the RPL19B gene, and 100% of ORF7 is identical to the RPL19C gene. Thus, by choosing the clique (Query, ORF5, ORF7), we identified two additional members among the *60S ribosomal L19 proteins*.

4. DISCUSSION

The parameters used by CAP3 during contig assembly are crucial to the success of the NSP strategy. Even when -p, the overlap percent identity cutoff, is set to 100%, there is a chance of generating artifactual contigs. The reasons for this include relatively high error in EST sequences, chimeras generated in cDNA cloning, and regions of highly conserved domains interspersed in related genes. Furthermore, with -p set to lower values (anywhere between 80 to 98%), the probability of creating artifactual contigs increases. It is therefore necessary for the NSP strategy to either filter out the artifactual coding regions (ORFs) within these contigs, or to select the real ones. In gene families where the paralogous coding regions have been constrained by negative (purifying) selection pressure, we have found that the paralogs can be identified by the ORFs after screening for amino acid substitution patterns indicative of such (NSP, Negative Selection Patterns). If the ORFs are from artifactual contigs, then no NSP should be seen between them. However, if no NSP is seen between a pair of ORFs, we cannot tell whether they represent different genes of the same family or not. Hence, as our results demonstrate, all members of a particular gene family may not be detectable by the NSP strategy. For example, in the *Adenine Phosphoribosyltransferases*, the NSP between the Query and ORF3 is not sufficient to prove the relation between them. ORF4, on the other hand, is related to both the Query and to ORF3. For the same reason, more members may be identified by iteratively using the representative ORFs of a previous NSP automation, as the future queries for the NSP automation.

It has been observed that contigs belonging to different regions of the same gene can produce high dS/dN values between each other. To some extent, this situation was avoided by analyzing pairs of ORFs from the same region only (see Figure 1). In this regard, an interesting result in AtRPL19 is the strong dS/dN ratio of 3.09 seen between

ORF1 and ORF2. These two ORFs did not resemble any known protein in TAIR or in the NCBI protein database. However, major parts of these two sequences closely resembled the mutated EMB2386 (Embryo Defective 2386) which is identical to RPL19A, the query protein. On closer analysis of the ESTs involved in contig assembly, it was found that ORF1 and ORF2 were artifactual contigs assembled from ESTs belonging to different regions of RPL19A. The resulting assembly produced an NSP between them.

The NSP strategy has, however, correctly identified additional members in 9 out of 10 gene families using EST data alone. In all 9 cases, the automated strategy successfully identified 1 to 4 additional members distinct from the member used as initial query. In the Actin family, two maximum cliques of size 5 were found, with 4 vertices in common. This brought the total number of additional members identified to 5.

The NSP strategy is now capable of identifying as many gene families as possible without sacrificing the accuracy reported here. We have automated all the steps involved after receiving the initial protein query: (1) BLAST collection of related ESTs, (2) contig assembly, (3) ORF identification, (4) partial screening of ORFs based on relative positions, (5) NSP-based screening of ORFs, and (6) identification of cliques of related ORFs (putative paralogs). The final output consists of ORFs representing at least one additional gene family member (besides the query). Since the initial protein query can be an orthologous sequence, we are currently working on identifying, in *Glycine max*, every gene family for which at least one member has been identified in another plant species. As of now, the NSP strategy is limited to gene families for which a starting query is available, i.e. a paralog or ortholog sequence of one gene family member. However, we have also had some limited success by applying the NSP strategy to clusters of related ESTs grouped only by similarity to each other. More work in this direction will allow us to expand the strategy to include identification of yet undiscovered gene families.

5. CONCLUSION

The NSP strategy is now fully automated such that, given a list of query proteins, it can identify as many putative member genes as possible. It is not yet a global gene family identification protocol, but our tests on the Arabidopsis EST dataset indicate that it performs well in distinguishing contigs that represent real genes from contigs that are artifacts. Almost every ORF predicted by the NSP strategy to be a distinct gene family member, mapped onto the appropriate gene in TAIR. Further expansion of the strategy to clustered ESTs can eliminate the need for individual query sequences allowing the identification of a significant proportion of gene families with reliable accuracy.

Table 1. Validation of NSP generated cliques against the TAIR database

Clique Identified by NSP	Family members identified by ORFs ('NM': an ORF didn't match a member gene)	Conclusions			
2-oxoglutarate,malate translocator precursor-like: 3 known members (DIT1, DIT2.1, DIT2.2) BLAST query: NP_568283 (DIT1)		250 ESTs	cap3 -o 21 -y 14 -p 83	7 contigs	7 ORFs
qry, orf3	DIT1, DIT2.1	**1 additional member identified**			
qry, orf5	DIT1, DIT2.2	**1 additional member identified**			
orf4, orf3	DIT1, DIT2.1	orf4 matches the portion of the query that aligns with orf3			
3-hydroxy-3-methylglutaryl-CoA reductase family: 2 known members (HMG1, HMG2) BLAST query: NP_177775 (HMG1)		247 ESTs	cap3 -o 21 -y 6 -p 95	15 contigs	10 ORFs
qry, orf4	HMG1, HMG2	**1 additional member identified**			
qry, orf5	HMG1, HMG2	orf5 is similar to orf4			
qry, orf6	HMG1, HMG2	orf6 matches neither orf4 nor orf5; It belongs to a separate portion of HMG2			
3-hydroxyisobutyryl-coenzyme A hydrolase family: 3 known members (NP_180623, NP_191610, NP_201395) BLAST query: NP_180623		182 ESTs	cap3 -o 21 -y 24 -p 89	16 contigs	16 ORFs
qry, orf1	NP_180623, NP_201395	**1 additional member identified**			
qry, orf2	NP_180623, NP_191610	**1 additional member identified**			
qry, orf7	NP_180623, NM	orf7 matches a related protein (CAB10400, matched using NCBI blastp on A. thaliana)			
qry, orf11	NP_180623, NM	orf11 matches the related AIM1 protein (NCBI blastp)			
orf1, orf16	NM, NP_201395	orf16 matches the related MFP2 protein (NCBI blastp)			
60S ribosomal L19 proteins: 3 known members (RPL19A, RPL19B, RPL19C) BLAST query: NP_171777 (RPL19A)		250 ESTs	cap3 -o 21 -y 6 -p 94	8 contigs	8 ORFs
qry, orf5, orf7	RPL19A, RPL19B, RPL19C	**2 additional members identified**			
orf3, orf5, orf7	RPL19A, RPL19B, RPL19C	orf3 matches the portion of the query that aligns with orf5 & orf7			

Table 1. (Continued)

orf1, orf2	NM, NM	orf1 & orf2 matches no known protein (NCBI blastp)			
orf3, orf2	RPL19A, NM	orf2 is possibly related to RPL19A			
AAA-type ATPases: 3 known members (CDC48A, CDC48D, CDC48E) BLAST query: NP_187595 (CDC48A)		250 ESTs	cap3 -o 21 -y 6 -p 87	10 contigs	10 ORFs
qry, orf1, orf4	CDC48A, CDC48E, CDC48D	**2 additional members identified**			
qry, orf2, orf6	CDC48A, NM, NM	orf2 matches the related CDC48C; orf6 matches FtsH8 (NCBI blastp)			
qry, orf6, orf9	CDC48A, NM, NM	orf9 matches RPT1A (NCBI blastp)			
orf8, orf6, orf9	NM, NM, NM	orf8 matches NP_001117220 (NCBI blastp)			
orf5, orf9	CDC48A, NM	orf5 matches the portion of the query that aligns with orf9 & orf10			
orf5, orf10	CDC48A, NM	orf10 matches FtsH4 (NCBI blastp)			
ABC-type transport-like: 3 known members (ATATH1, ATATH11, ATATH15) BLAST query: NP_190357 (ATATH1)		247 ESTs	cap3 -o 21 -y 7 -p 98	6 contigs	5 ORFs
qry,orf6	ATATH1,NM	orf6 matches the related WBC7			
Actin: 7 known members (ACT2, ACT3, ACT4, ACT7, ACT8, ACT11, ACT12) BLAST query: NP_175350 (ACT8)		250 ESTs	cap3 -o 21 -y 58 -p 93	8 contigs	8 ORFs
qry,orf1,orf3,orf4,orf5	ACT8,ACT3,ACT2,ACT4,ACT7	**4 additional members identified**			
orf8,orf1,orf3,orf4,orf5	ACT8,ACT3,ACT2,ACT4,ACT7	orf8 matches the full length of the query			
qry,orf1,orf3,orf4,orf7	ACT8,ACT3,ACT2,ACT4,ACT11	**1 additional member identified**			
orf8,orf1,orf3,orf4,orf7	ACT8,ACT3,ACT2,ACT4,ACT11				
qry,orf3,orf5,orf6	ACT8,ACT2,ACT7,ACT11	orf6 & orf7 match separate portions of ACT11			
orf8,orf3,orf5,orf6	ACT8,ACT2,ACT7,ACT11				
Adenine Phosphoribosyltransferases: 5 known members (APT1, APT2, APT3, APT4, APT5) BLAST query: NP_564284 (APT1)		250 ESTs	cap3 -o 21 -y 10 -p 90	6 contigs	6 ORFs
qry,orf4	APT1,APT3	**1 additional member identified**			
orf1,orf4	APT1,APT3	orf1 matches the portion of the query that aligns with orf4			
orf3,orf4	APT2,APT3	**1 additional member identified** but the NSP between the query and orf3 was not sufficient			
ADP,ATP carrier-like Protein family: 4 known members (AAC1, AAC2, AAC3, NP_568345) BLAST query: NP_187470 (AAC1)		250 ESTs	cap3 -o 21 -y 6 -p 90	2 contigs	2 ORFs
qry,orf1	AAC1,AAC2	**1 additional member identified**			
orf2,orf1	AAC1,AAC2	orf2 matches the full length of the query			
Aldose 1-epimerase - like family: 2 known members (NP_197018, NP_190364) BLAST query: NP_197018		128 ESTs	cap3 -o 21 -y 6 -p 98	6 contigs	6 ORFs
qry,orf2	NP_197018,NP_190364	**1 additional member identified**			
orf4,orf2	NP_197018,NP_190364	orf4 matches the portion of the query that aligns with orf2			

Note: TAIR lists Actin with 8 members, one of which (NP_565867) was removed from NCBI's GenBank as a result of standard genome annotation processing. This member was not used during validation.

6. REFERENCES

Bie T., Cristianini N., Demuth J.P., and Hahn M.W., 2006, "CAFE: A Computational Tool for the Study of Gene Family Evolution," *Bioinformatics Applications Note* 22(10):1269-1271.

Brown S., Chang J.L., Sadée W., and Babbitt P.C., 2003, "A Semiautomated Approach to Gene Discovery through Expressed Sequence Tag Data Mining: Discovery of New Human Transporter Genes," *AAPS PharmSci*, 5(1).

Frank R.L., Kandoth C., and Ercal F., 2008, "Validation of an NSP-based (negative selection pattern) gene family identification strategy," *Proceedings, 5th Conference of the MidSouth Computational Biology and Bioinformatics Society, BMC Bioinformatics*, 9(9):S2.

Frank R.L., Mane A., and Ercal F., 2006, "An Automated Method for Rapid Identification of Putative Gene Family Members in Plants," *BMC Bioinformatics*, 7:S19.

Huang X. and Madan A., 1999, "CAP3: A DNA Sequence Assembly Program," *Genome Research*, 9(9):868-877.

Korber B., 2000, "HIV Signature and Sequence Variation Analysis," In *Computational Analysis of HIV Molecular Sequences*, Edited by Rodrigo AG, Learn GH Netherlands: Kluwer Academic Publishers: 55-72.

Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., and Higgins D.G., 2007, "Clustal W and Clustal X version 2.0," *Bioinformatics*, 23, 2947-2948.

Nei M., and Gojobori T., 1986, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions," *Mol Biol Evol*, 3(5):418-426.

Ota T., and Nei M., 1994, "Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site," *Mol Biol Evol*, 11(4):613-619.

Retief J.D., Lynch K.R., and Pearson W.R., 1999, "Panning for Genes: A Visual Strategy for Identifying Novel Gene Orthologs and Paralogs," *Genome Research*, 9:373-382.

Suyama M., Torrents D., and Bork P., 2006, "PAL2NAL: Robust Conversion of Protein Sequence Alignments into Corresponding Codon Alignments," *Nucleic Acids Research*, 34:W609-W612.

Taylor J.S., and Raes J., 2005, "Small-Scale Gene Duplications," In *The Evolution of the Genome*. Elsevier Academic Press: 289-327.

III. A FRAMEWORK FOR AUTOMATED ENRICHMENT OF FUNCTIONALLY SIGNIFICANT INVERTED REPEATS IN WHOLE GENOMES

Cyriac Kandoth[1], Fikret Ercal[1], Ronald L Frank[2]

[1]Department of Computer Science, Missouri University of Science and Technology, Rolla MO, 65401, USA

[2]Department of Biological Sciences, Missouri University of Science and Technology, Rolla MO, 65401, USA

ABSTRACT

Background - RNA transcripts from genomic sequences showing dyad symmetry typically adopt hairpin-like, cloverleaf, or similar structures that act as recognition sites for proteins. Such structures often are the precursors of non-coding RNA (ncRNA) sequences like microRNA (miRNA) and small-interfering RNA (siRNA) that have recently garnered more functional significance than in the past. Genomic DNA contains hundreds of thousands of such inverted repeats (IRs) with varying degrees of symmetry. But by collecting statistically significant information from a known set of ncRNA, we can sort these IRs into those that are likely to be functional.

Results - A novel method was developed to scan genomic DNA for partially symmetric inverted repeats and the resulting set was further refined to match miRNA precursors (pre-miRNA) with respect to their density of symmetry, statistical probability of the symmetry, length of stems in the predicted hairpin secondary structure, and the GC content of the stems. This method was applied on the *Arabidopsis thaliana* genome and validated against the set of 190 known Arabidopsis pre-miRNA in the miRBase database. A preliminary scan for IRs identified 186 of the known pre-miRNA but with 714700 pre-miRNA candidates. This large number of IRs was further refined to 483908 candidates with 183 pre-miRNA identified and further still to 165371 candidates with 171 pre-miRNA identified (i.e. with 90% of the known pre-miRNA retained).

Conclusions - 165371 candidates for potentially functional miRNA is still too large a set to warrant wet lab analyses, such as northern blotting, on all of them. Hence additional filters are needed to further refine the number of candidates while still retaining most of the known miRNA. These include detection of promoters and terminators, homology analyses, location of candidate relative to coding regions, and better secondary structure prediction algorithms. The software developed is designed to easily accommodate such additional filters with a minimal experience in Perl.

1. BACKGROUND

In the last decade, non-coding RNA (ncRNA) sequences have become more essential to our understanding of gene organization. They were once considered insignificant in comparison to protein coding sequences. But since then, a variety of new types of ncRNA genes have been discovered, each of them revealing new biological roles and cellular mechanisms like gene silencing, replication, gene expression regulation, transcription, chromosome stability, and protein stability [1,2,3]. Therefore, the identification of ncRNA has significant importance to the biological and medical community. To date, the genomes of numerous organisms have been fully sequenced, making it possible to perform genome-wide computational analyses. Computational methods of ncRNA identification typically involve scanning genomic DNA or transcriptome data for candidate sequences, after which wet lab techniques like northern blotting are required to verify their cellular function [4].

The precursors of non-coding RNA sequences like transfer RNA, ribosomal RNA, microRNA, and small-interfering RNA, typically adopt hairpin-like, cloverleaf, or similar symmetric structures that act as recognition sites for proteins. These structures are the result of dyad symmetry, i.e. inverted repeats (IRs) in the RNA sequences. But the number of ncRNAs identified to date is only a fraction of the hundreds of thousands of IRs that can be found in genomic DNA. This makes it difficult to claim that any inverted repeat in a genome has functional significance, but it potentially raises the number of functional RNA sequences that have yet to be identified.

In this paper, we focus on the identification of microRNAs (miRNA) which are short, ~22 nucleotide long ncRNAs that are involved in gene regulation at the level of translation. This can occur through cleavage of the messenger RNA, or through translational repression causing regulation of a specific protein. They have been linked with the onset of cancer and other diseases based on miRNA expression levels [5]. The processing of miRNA from genomic DNA and its subsequent activation in cells is a multistep process that starts with transcription from genomic DNA into RNA transcripts called primary miRNAs (pri-miRNAs). These variable length transcripts contain the mature miRNA as a subsequence, with inverted repeats that usually form a more stable hairpin-like structure called a precursor-miRNA (pre-miRNA). This structure is generally around 70 nucleotides long, with 25 to 30 base-pair stems, and relatively small loops. A sample pre-miRNA hairpin structure is shown in Figure 1. The pre-miRNA hairpin is released from the pri-miRNA transcript with the help of ribonuclease Drosha [6]. Recently, a type of miRNA that bypasses Drosha processing has been discovered [7], but most known miRNAs are still subject to processing by Drosha. After the pre-miRNA hairpin is released, it is exported from the cell nucleus to the cytoplasm where the ribonuclease Dicer cleaves the pre-miRNA approximately 19 bp from the Drosha cut site resulting in a double-stranded RNA. One of these two strands becomes the mature miRNA sequence by associating itself with the RNA-Induced Silencing Complex (RISC). RISC uses the miRNA as a template for recognizing complementary target messenger RNA (mRNA) to regulate a specific protein coding gene. Several miRNA identification strategies take advantage of this understanding of miRNA processing and activation and they are discussed below.

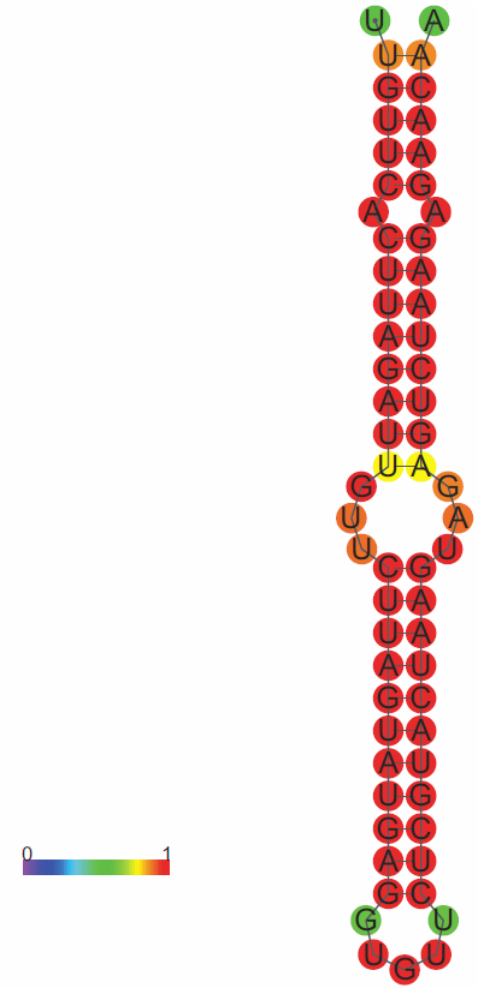

Figure 1. A typical hairpin-like secondary structure of a microRNA precursor

This secondary structure was generated using the RNAfold secondary structure predictor of the Vienna RNA WebSuite on a known Arabidopsis microRNA precursor retrieved from the miRBase database [miRBase:MI0008304]. The color-code used represents the base-pair probabilities based on a minimum free energy analysis.

Transcription from DNA to RNA is typically guided by the presence of promoter and terminator sequences in the genome that usually lie in the vicinity of non-coding or protein coding genes. However, current methods can only detect certain classes of promoters and terminators, and the degrees of accuracy of such methods are insufficient for genome-wide scans [8]. In addition to this, the starting points of the transcripts in the genome are not always known, even for commonly studied genes. It has been reported that some intergenic regions (DNA between protein coding genes) contain ncRNA that

act to regulate the genes nearby. Hence, many RNA detection methods make the assumption that ncRNA is present in the vicinity of known genes and between coding regions within genes (introns). However, most intergenic DNA still have no known function and the basis for this assumption is anecdotal. Current estimates show that approximately 60% of miRNAs are expressed independently, 15% of miRNAs are expressed in clusters, and 25% are in introns [9].

The cellular function of many non-coding RNA is directly dependent on the three-dimensional structure that it adopts. This is in turn dependent on the secondary structure, which is dependent on the transcribed RNA sequence. If the RNA is functionally significant, then the structure and sequence can be conserved over the generations. This is a commonly used evolutionary rationale for ncRNA detection using homolog analyses. In the case of miRNA, pre-miRNA sequences adopt a stable hairpin structure that is necessary for processing by the ribonuclease enzyme Dicer, after which the mature miRNA sequences associate with RISC to perform their regulatory roles. A method called miRNAminer [10] searches for such evolutionarily related miRNA sequences from different species (homologs). Given a query miRNA, candidate homologs from different species are tested for secondary structure, alignment and conservation, in order to assess their candidacy as miRNAs. By computationally identifying small sections of a genome that could form hairpin-like secondary structures, some researchers have been able to identify sets of potential miRNA sequences which include a subset of known miRNA. Two such methods, miRSeeker [11] and miRScan [12], first identify hairpin structures within highly conserved intergenic regions. To these candidates, miRSeeker applies mutation patterns that are typical of miRNA precursors, and miRScan identifies those structures having features such as symmetric bulges or a highly conserved stem near the terminal loop. miRRim [13] represents the evolutionary and secondary structural features of all known miRNA and their surrounding regions with a sequence of multidimensional vectors. It uses these to train hidden Markov models (HMM) for miRNA and non-miRNA sequences. These models are combined into a single HMM and used to search genomic sequences for miRNA. Current methods of secondary structure prediction involve a dynamic programming algorithm similar to those used for sequence alignment.

These methods are promising, but cannot predict more complex secondary structures like pseudoknots (non-nested pairing). Clote et al [14] proposed that the secondary structures of functional ncRNA are more thermodynamically stable than random RNA. The Gibbs free energy ($\Delta G°$) is a popularly used measure of this thermodynamic potential energy, and some ncRNA detection methods incorporate it as a threshold for detection of miRNA [10].

The hairpin-like secondary structure of a microRNA is a result of the inverted repeats that it contains. It is believed that IRs are the result of inverted DNA duplication events that occurred during the course of evolution of most organisms [15]. If this is the case, the asymmetries and bulges as seen in Figure 1 are formed later as a result of accumulation of mutations, insertions, and deletions. However, the inverted repeats remain highly conserved since the base-paired stem loops of the hairpin structures are relatively much longer than the asymmetries. The degree of dyad symmetry can therefore be used as a criterion for miRNA detection. We present a fast genome-wide scanning algorithm named irScan that first finds all sufficiently symmetric IRs in a given genomic DNA sequence (typically a whole chromosome). This large number of ncRNA candidates is then further reduced based on user-defined criteria for ncRNA detection. We demonstrate the capability of this algorithm using criteria for miRNA detection like the density of symmetric matches in the IR (density of base-pairs in the hairpin-like secondary structure), statistical probability of the symmetry, average length of contiguous symmetric matches in the IR (length of base-paired stems in the hairpin), and the GC content of the matches in the IRs. Detection of inverted repeats by itself is an insufficient criterion for ncRNA detection. Our preliminary scan using irScan's base thresholds on the fully sequenced Arabidopsis chromosomes revealed around 1.1 million mostly symmetric IRs. It is thus necessary to bring this number down to a small set of candidates that are most likely to be functionally significant ncRNA and hence warrant further wet lab analyses.

2. METHODS

2.1. DETECTION OF INVERTED REPEATS

irScan starts by scanning for IRs in the given genomic sequence using a variation of the Smith-Waterman (SW) local alignment algorithm [16]. The original SW algorithm is a dynamic programming technique that generates an optimal gapped local alignment between two given sequences based on a predefined scoring matrix for matches, mismatches, and gaps. An implementation of this algorithm was written in C++ that takes only one sequence as input, translates the DNA character set (ACGT) to the RNA character set (ACGU), generates a reverse complement of it, and then aligns it against the original sequence. The resulting local alignment would then reveal an optimal inverted repeat in the original sequence based on the match and mismatch penalties shown in Table 1, and a gap penalty of -6. These penalties appeared to work best at predicting the secondary structures of the known pre-miRNA sequences. Since this algorithm returns only one IR per input sequence, longer input sequences need to be subdivided further to detect shorter clustered IRs. So irScan used scanning windows of sizes 600, 300, and 150 base pairs to reflect the various sizes of known pre-miRNA. Each scanning window skips through the given genomic sequence by half the number of base pairs i.e. a skip size of 300bp is used for the 600bp scanning windows, 150bp skip size for 300bp windows, and 75bp for 150bp windows. The shorter of the matching IRs (duplicates) generated by adjacent overlapping windows of the same size are removed. But the duplicates generated by overlapping windows of different sizes are retained because the duplicate removal process cannot distinguish between nearby identical IRs and duplicate IRs generated by overlapping scanning windows. However, the benefit of substantial reduction in the number of IRs using this simple duplicate remover outweighed the computational cost of implementing a more accurate but complex duplicate remover. For all runs of irScan, the simpler duplicate removal process was implemented to considerably reduce the number of pre-miRNA IR candidates, but only after all the miRNA specific filters were applied. These filters are explained below.

Table 1. Scoring matrix used by irScan's IR detector

	A	C	G	U	N
A	5	-4	-4	-4	0
C	-4	5	-4	-4	0
G	-2	-4	5	-4	0
U	-4	-2	-4	5	0
N	0	0	0	0	3

The local alignment scoring matrix used when aligning a nucleotide sequence against its reverse complement. Matches score 5 points, while mismatches are penalized by 4. All loci with unresolved ambiguity in the assigned base are treated as N. This matrix appears to work best at predicting the secondary structures of the known 190 pre-miRNA. G-A and U-C mismatches are not penalized as much to accommodate the occurrence of G-U and U-G base pairs respectively.

2.2. MICRORNA PRECURSOR ANALYSIS

Since our initial target genome for functional IR identification will be a plant, criteria for distinguishing potentially functional from nonfunctional IRs were obtained from an analysis of 190 known miRNA precursors from *Arabidopsis thaliana*. The nucleotide sequences of these 190 pre-miRNA were retrieved from the miRBase database [17] and aligned against each of their reverse complements using irScan's variant of the SW algorithm. This generates the inverted repeat portion of the pre-miRNA that can be represented in the dot-bracket notation as shown in Figure 2 for the same secondary structure shown in Figure 1. It shows 52 matches among 63 nucleotide bases producing a relatively high 82.54% density of matches. This became our first criterion for miRNA detection. The density of matches in an IR, denoted *D*, generated from genomic DNA has to pass a predefined threshold, denoted D_{min}, to be considered a sufficiently symmetric IR. The values of *D* seen among the 190 known pre-miRNA precursors ranged between 57% and 89%. To sufficiently reduce the number of IRs generated in the preliminary scan, we selected D_{min} = 59% that excludes only 3 of the known 190 miRNA. It is important to note that the value of *D* for a pre-miRNA could be slightly different from what the IR detector finds for the same sequence in genomic DNA because the equivalent IR seen in genomic DNA could be a subset or a superset of the known pre-miRNA. Also, *D* can become 100% if the sequence contains a perfectly symmetric IR. This is never the

case in pre-miRNA because loops in the hairpin are necessary for miRNA processing, but it is seen in low complexity regions of genomic DNA. We therefore also apply a D_{max} of 95% to exclude such low complexity regions. A frequency distribution of D on the 190 known pre-miRNA is shown in Figure 3.

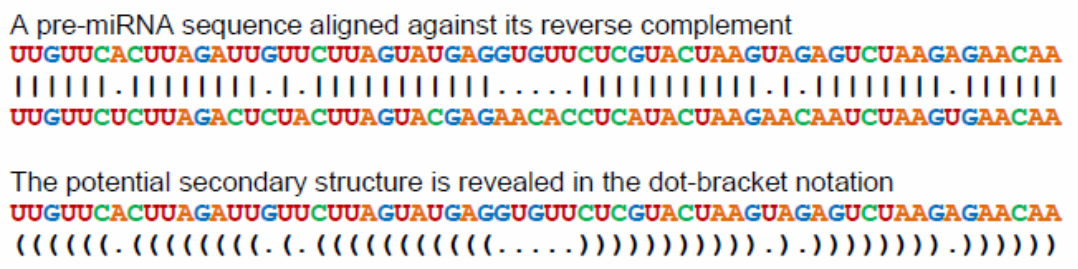

Figure 2. irScan's IR detector emulates a secondary structure predictor

The variant of the Smith-Waterman algorithm used by irScan's IR detector generates this dot-bracket notation of the secondary structure when used on a known Arabidopsis precursor [miRBase:MI0008304]. The inverted repeats form 25 base pairs of which two are weaker G-U matches surrounded by matches.

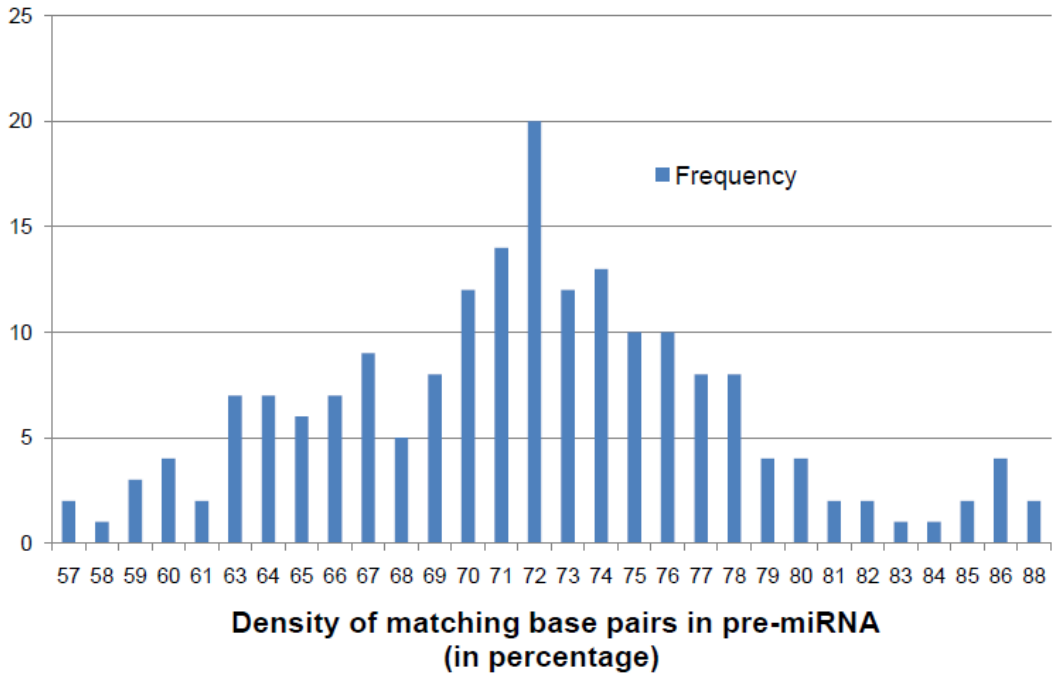

Figure 3. Frequency distribution of parameter D on the 190 known pre-miRNA

Shows the frequency distribution of densities of base-pair matches (*D*) between IRs detected on the 190 known pre-miRNA sequences of *Arabidopsis thaliana*. The values are rounded to the closest integer.

Our second criterion of detection is based on the probability of occurrence of an IR in a randomly generated RNA sequence. Let us denote this as *P*. Small values of *P* most likely indicate highly conserved dyad symmetries and hence potential functionality, while large values of *P* most likely indicate a random RNA sequence. But they could also indicate a potentially functional RNA that has lost most of its symmetry but retained its functionality. Using *P* and *D* values as filters excludes such ncRNAs, but from our understanding of pre-miRNA processing, sufficient symmetry between inverted repeats is a necessary condition for the formation of stable hairpins that can be processed by the ribonuclease Drosha. The calculation of *P*, like *D*, depends on the ratio of matches to mismatches in the IR generated. This is described below.

Consider an RNA sequence with *2k* nucleotide bases. The left hand side (LHS) of length *k* bases is mostly inversely symmetric with the right hand side (RHS) of equal length resulting in the hairpin-like secondary structures shown in Figures 1 and 2. Let *n* be the number of bases that are inverted repeats (part of the base-pairing stem loop). The probability that *n* is exactly 1 is represented as $P(1, k) = 0.25 \times (0.75)^{k-1} \times k$, where *0.25* is the probability that one base in the LHS is the reverse complement of the corresponding base in the RHS, out of 4 possible bases A, C, G, or U. $(0.75)^{k-1}$ is the probability that all other *k-1* bases are mismatches. And *k* is the number of combinations in which this can occur. Similarly, if *n* is exactly 2, then $P(2, k) = (0.25)^2 \times (0.75)^{k-2} \times {}^kC_2$, where ${}^kC_2$ is the number of combinations in which 2 matches can occur among *k-2* mismatches. In general, we can use Equation 1 below.

$$P(n, k) = (0.25)^n \times (0.75)^{k-n} \times {}^kC_n \quad (1)$$

The values of *P* among the known pre-miRNA ranged from 10^{-7} to 10^{-62}. We selected 9.99×10^{-9} as P_{max}, an upper bound threshold for *P*, which excludes two of the known pre-miRNA. A frequency distribution of *P* on the 190 known pre-miRNA is shown in Figure 4. Additionally, the values of P for various combinations of *n* and *k* were plotted and it was noted that *P* also reduced when *n* was much smaller than *k*, i.e. when there are many more mismatches than matching base-pairs in the IR. This was because

the statistically most probable ratio of *n*:*k* is1:4 i.e. 25% of base-pairs in an IR are statistically most likely to be matches than mismatches in a randomly generated sequence. This follows from the fact that the nucleotide bases have an alphabet size of 4 (A, C, G, U). Hence, *P* tends to get smaller as this ratio becomes larger (or smaller) than 25%. With a combination of filters D_{min} and P_{max}, only the more symmetric IRs are detected. It can be argued that the two filtering criteria can be replaced with just D_{min}, but the value of *P* is much more indicative of the statistical significance of an IR.

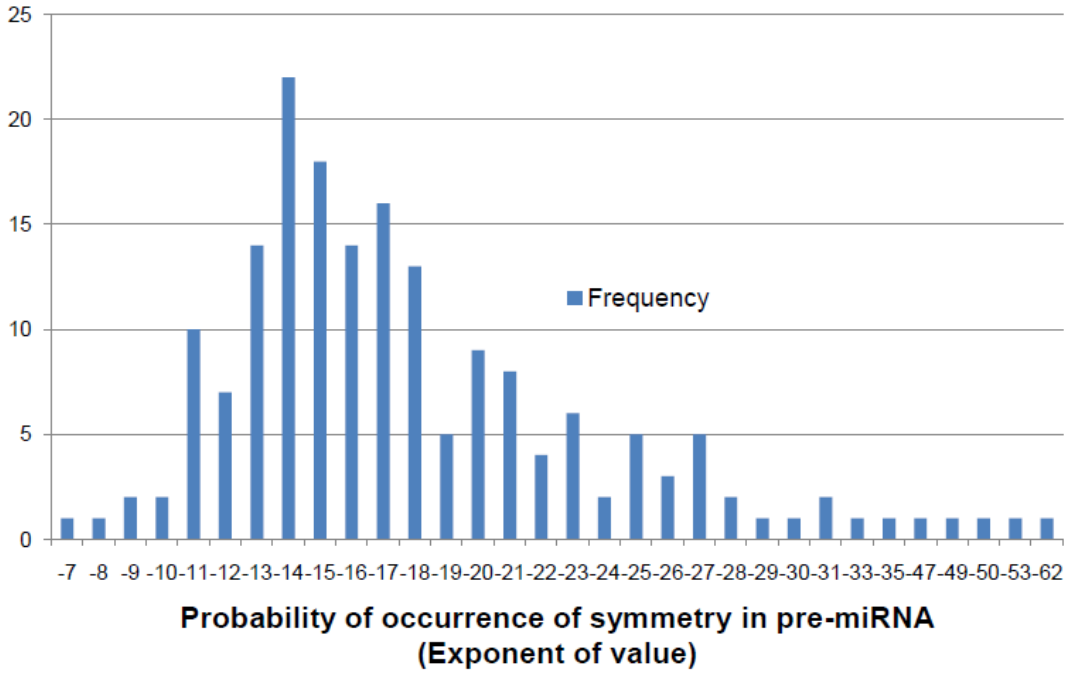

Figure 4. Frequency distribution of parameter *P* on the 190 known pre-miRNA

Shows the frequency distribution of the exponents of probability of occurrence of base-pair matches (*P*) between IRs detected on the 190 known pre-miRNA sequences of *Arabidopsis thaliana*. The values shown are the exponents of the probability value i.e. the exponent -11 indicates *P* values from 1.00×10^{-11} to 9.99×10^{-11}.

2.3. ADDITIONAL FILTERS

Using thresholds of D_{min} = 59%, D_{max} = 95%, $P_{max} = 9.99\times10^{-9}$, and a minimum IR length of 50 bp, irScan returns around 1.5 million IRs which include 186 of the 190

known pre-miRNAs. If duplicates are removed, this number goes down to 1.1 million with 183 known pre-miRNAs identified. To reduce this number further, additional filters are required. The third criterion for pre-miRNA detection was based on the observation that pre-miRNA secondary structures have relatively long stems. In the dot-bracket notation of Figure 2, these stems would be represented as contiguous matches. We calculated the average of contiguous match lengths in the IRs of the known pre-miRNA, denoted as *A*, and they ranged between 2.1 and 10.6 base pairs. For the IR in Figure 2, this is calculated as the average of lengths 6, 8, 1, and 11 making $A = 6.5$ bp. A base threshold of $A_{min} = 2.2$ was selected which excluded 3 of the known 190 pre-miRNA. Any IR detected in genomic DNA with a value of *A* lower than 2.2 bp was filtered out. The frequency distribution of values of *A* seen on the 190 known pre-miRNA is shown in Figure 5.

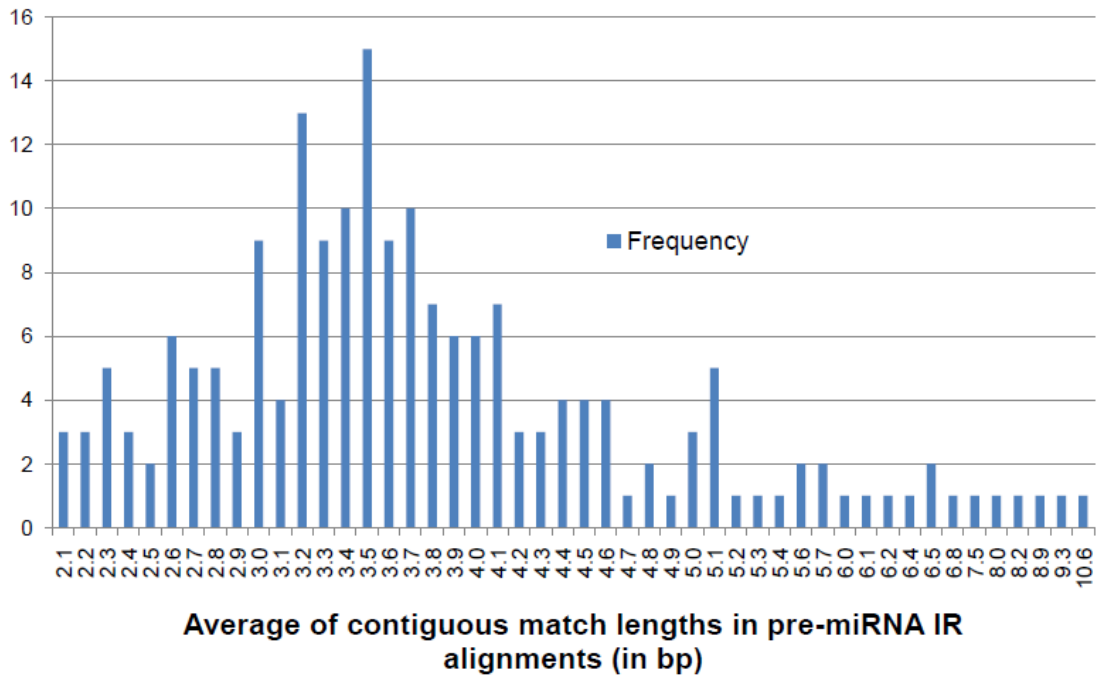

Figure 5. Frequency distribution of parameter *A* on the 190 known pre-miRNA

Shows the frequency distribution of average contiguous match lengths (*A*) for the 190 known pre-miRNA sequences of *Arabidopsis thaliana*. The values of *A* for each pre-miRNA IR has been floored to the closest number with 1 decimal place.

G-C base pairs in RNA sequences have three hydrogen bonds, making them more thermodynamically stable than A-U base pairs with two hydrogen bonds. Additionally, there is evidence that pre-miRNA hairpins are more thermodynamically stable than random sequences [18]. We therefore use GC content of the hairpin stems as the fourth criterion for pre-miRNA detection. The percentage of GC pairs in contiguous matches longer than 3 bp was calculated for each of the 190 known pre-miRNA. Contiguous matches that were 3 bp or shorter were more likely to belong to a loop than a stem in the hairpin-like secondary structure. So the GC content of these sufficiently long contiguous regions were calculated and denoted as *G*. For the IR in Figure 2, this is the percentage of GC pairs within the contiguous matching base-pairs of lengths 6 bps, 8 bps, and 11 bps. Among the 190 known pre-miRNA, *G* ranged from 18% to 62%. The base threshold G_{min} was set to 18% that did not exclude any of the known pre-miRNA. The frequency distribution of values of *G* seen on the 190 known pre-miRNA is shown in Figure 6.

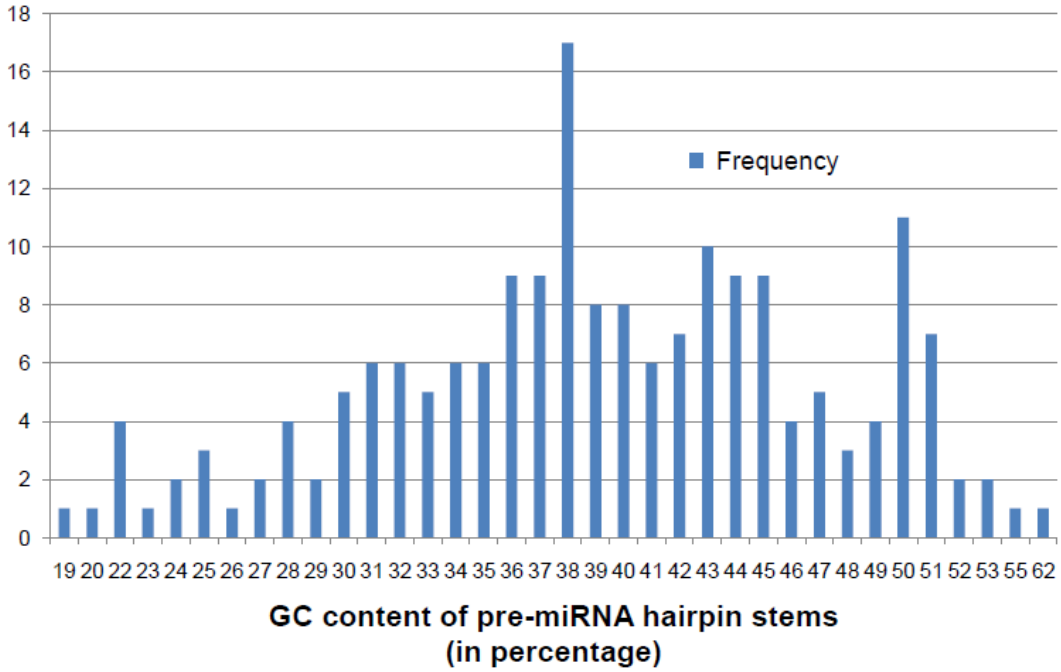

Figure 6. Frequency distribution of parameter *G* on the 190 known pre-miRNA

Shows the frequency distribution of GC content density (*G*) for the 190 known pre-miRNA sequences of *Arabidopsis thaliana*. The values of G for each pre-miRNA IR has been floored to the closest whole number lesser than it.

2.4. THE IRSCAN FRAMEWORK

Figure 7 shows how irScan's software framework was organized so as to allow the identification of any type of ncRNA with the addition of new filters and requiring only Perl programming knowledge.

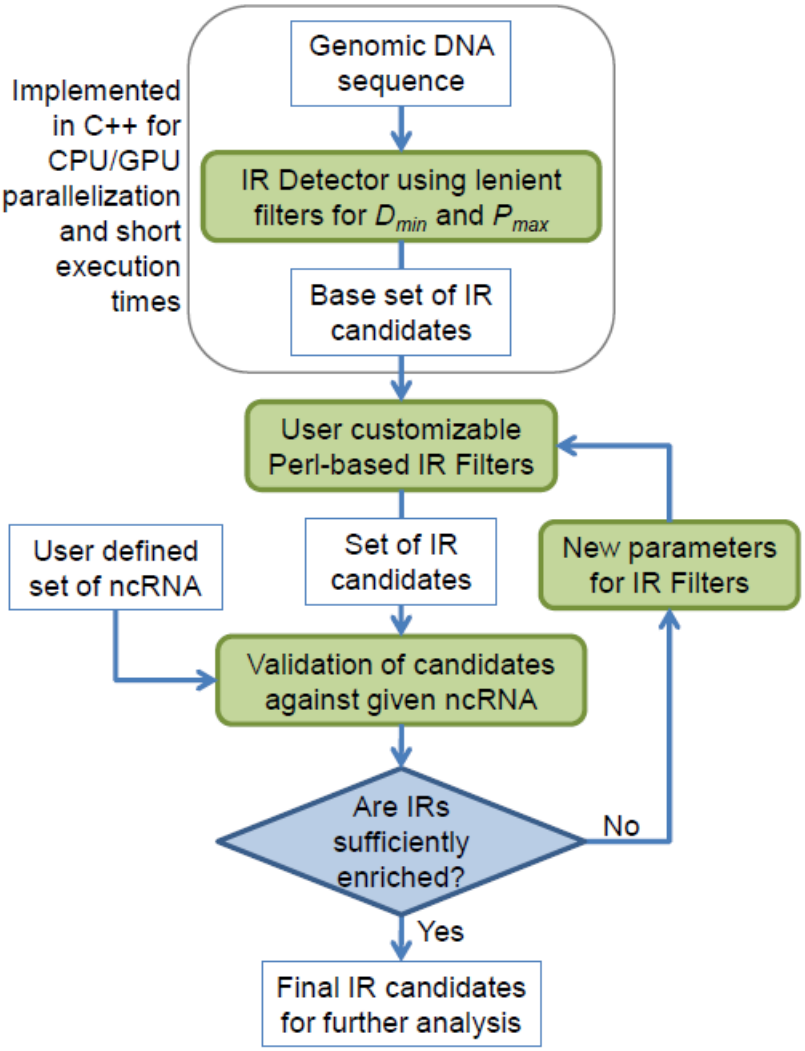

Figure 7. The irScan framework for ncRNA identification

The irScan software was designed and organized such that anyone with Perl programming knowledge could design their own filtering criteria for the base IRs detected by the irScan's C++ based IR detector that uses only the D_{min} and P_{max} (these are not specific to any particular type of ncRNA). Additional IR Filters are then user-defined in Perl and automated to find optimal parameter sets.

The IR detector represents the most computationally demanding portion of the framework and was implemented in C++ to quickly produce a base set of IRs filtered using the preliminary base threshold values for D_{min} and P_{max}. The resulting large set of preliminary IRs detected could then be further enriched using customized filtering criteria coded in Perl. It was decided to use Perl to implement these additional filters because of its popularity among biologists and bioinformaticians. The parameters for these custom designed IR filters could be tested in a validation loop that continually tweak the parameters and rerun the validation until the IRs are sufficiently small in number for further analysis, while still retaining at least a predefined number of known ncRNA in the validation set.

3. RESULTS

3.1. IRSCAN USING BASE PARAMETERS

The base parameters for irScan were selected to identify as many of the known 190 *At* pre-miRNA as possible, while keeping the total number of IRs detected less than 1 million. These parameters were D_{min} = 59%, $P_{max} = 9.99\times10^{-9}$, A_{min} = 2.2 bp, and G_{min} = 18 bp. In all runs of irScan, D_{max} was set to 95% to exclude low complexity regions, and IRs had to be at least 50 bp long to qualify as potential miRNA precursors. The irScan program returned 714700 IR candidates with these base parameters which included 186 of the known pre-miRNA sequences. If duplicate IRs generated by overlapping windows of different sizes were removed, then 483908 IR candidates remained with 183 of the known pre-miRNA sequences. Three of the initial 186 pre-miRNA were skipped because the simpler duplicate removal process cannot always distinguish between nearby identical IRs and duplicate IRs generated by overlapping scanning windows.

3.2. FINDING OPTIMAL PARAMETERS FOR IRSCAN

Optimal parameters for irScan was defined as those that generate the fewest IR candidates but still retained at least 90% of the 190 known pre-miRNA or 171 of them. IRs that were either a substring of a known pre-miRNA or that contained a known pre-miRNA sequence, were called Identified IRs (IIRs). A Perl script was written to

repeatedly run irScan on a user-defined starting parameter set, find the number of IIRs identified, then increase or decrease the irScan parameters to identify closer to 171 IIRs, while minimizing the total number of IRs detected. This repetition was terminated if it found a set of parameters that identified exactly 171 IIRs, or if the user terminated the script when it was close enough to 171. Table 2 shows the IR and IIR counts for various combinations of parameters A_{min} and G_{min}. The values of D_{min} and P_{max} were fixed at 60% and 9.99×10^{-11} respectively, which by themselves retain around 94% of the known pre-miRNA (178 IIRs). Figures 8 thru 11 show the frequency distributions of all 4 parameters on the IRs detected on genomic DNA.

From Table 2, we can see that the optimal parameters were A_{min} = 2.3 and G_{min} = 24, with D_{min} = 60% and P_{max} = 9.99×10^{-11}. This set of irScan parameters finds 165371 IR pre-miRNA candidates which include exactly 171 IIRs. This is still too large a number of candidates to warrant wet lab analyses on each, but it is a considerable reduction from the 1.5 million found by preliminary scans.

Table 2. Number of IRs and IIRs found using different irScan filters

	A_{min}=2.2		A_{min}=2.3		A_{min}=2.4	
	IRs	IIRs	IRs	IIRs	IRs	IIRs
G_{min}=18	260568	175	218041	175	169017	171
G_{min}=19	251543	174	209739	174	161903	170
G_{min}=20	243973	174	202877	174	156063	170
G_{min}=21	232543	173	192640	173	147592	169
G_{min}=22	222404	173	183721	173	140333	169
G_{min}=23	211825	171	174453	171	132789	168
G_{min}=24	201301	171	**165371**	**171**	125521	168
G_{min}=25	193352	170	158493	170	119957	167
G_{min}=26	180251	168	147397	168	111281	165

Shows the number of IR candidates generated by irScan on the 5 chromosomes of the Arabidopsis genome using various parameters. Parameters D_{min} and P_{max} were fixed at 60% and 9.99×10^{-11} respectively while A_{min} and G_{min} were varied as shown. IIRs are the number of Identified IRs that uniquely match one of the known 190 pre-miRNA.

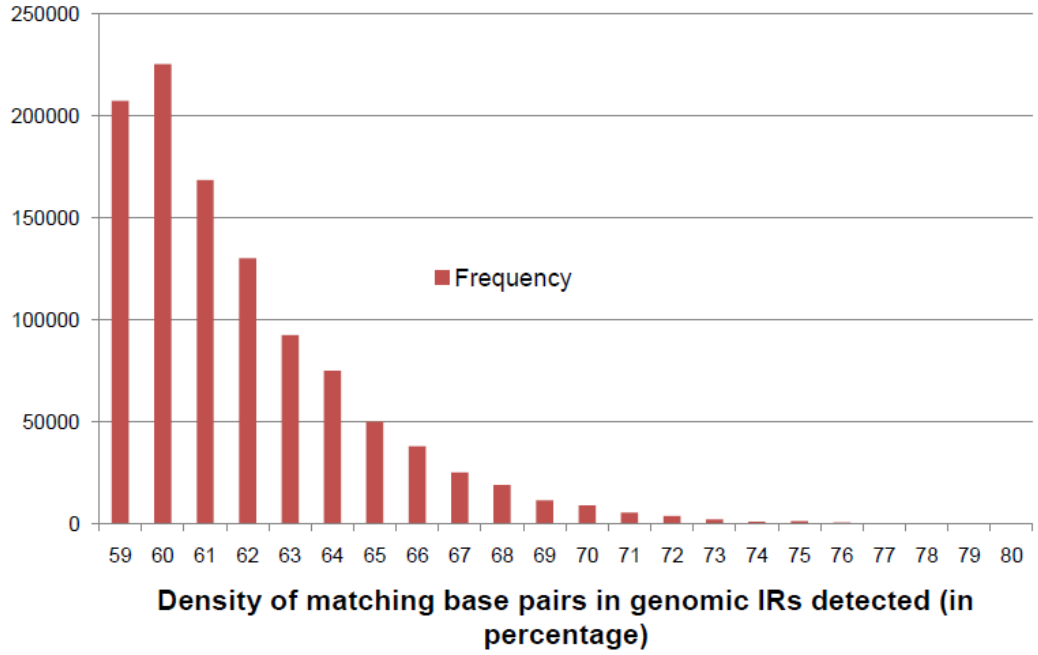

Figure 8. Frequency distribution of parameter *D* on genomic *At* IRs

Shows the frequency distribution of densities of base-pair matches (*D*) in IRs detected from genomic DNA of *Arabidopsis thaliana* with thresholds of D_{min} = 59% and P_{max} = 9.99×10^{-9}. Rounded to the closest integer.

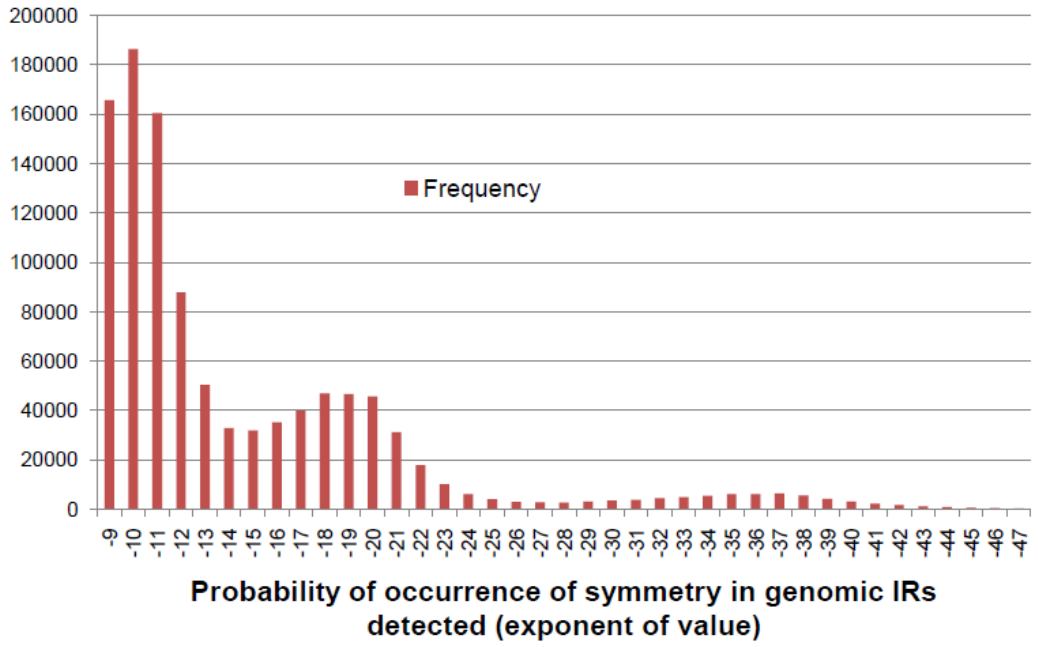

Figure 9. Frequency distribution of parameter *P* on genomic *At* IRs

Shows the frequency distribution of the exponents of probability of occurrence of base-pair matches (*P*) in IRs detected from genomic DNA of *Arabidopsis thaliana* with base thresholds of D_{min} = 59% and P_{max} = 9.99×10^{-9}. The values shown are the exponents of the probability value i.e. the exponent -11 indicates *P* values from 1.00×10^{-11} to 9.99×10^{-11}.

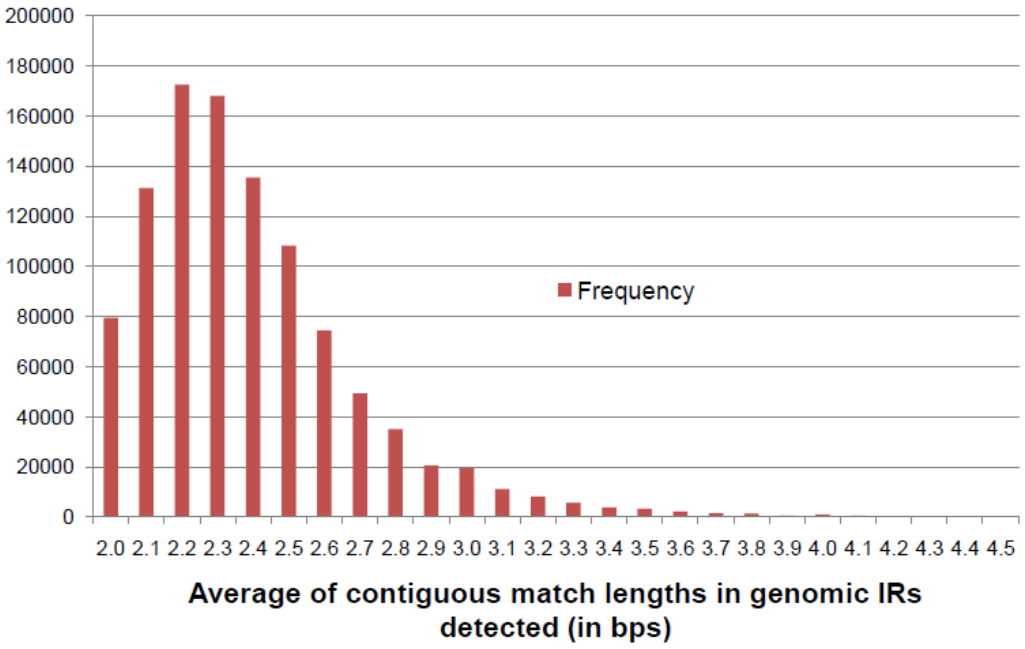

Figure 10. Frequency distribution of parameter *A* on genomic *At* IRs

Shows the frequency distribution of average contiguous match lengths (*A*) in IRs detected from genomic DNA of *Arabidopsis thaliana* with base thresholds of D_{min} = 59% and P_{max} = 9.99×10^{-9}, A_{min} = 2.0, and G_{min} = 19. The values of *A* for each pre-miRNA IR has been floored to the closest number with 1 decimal place.

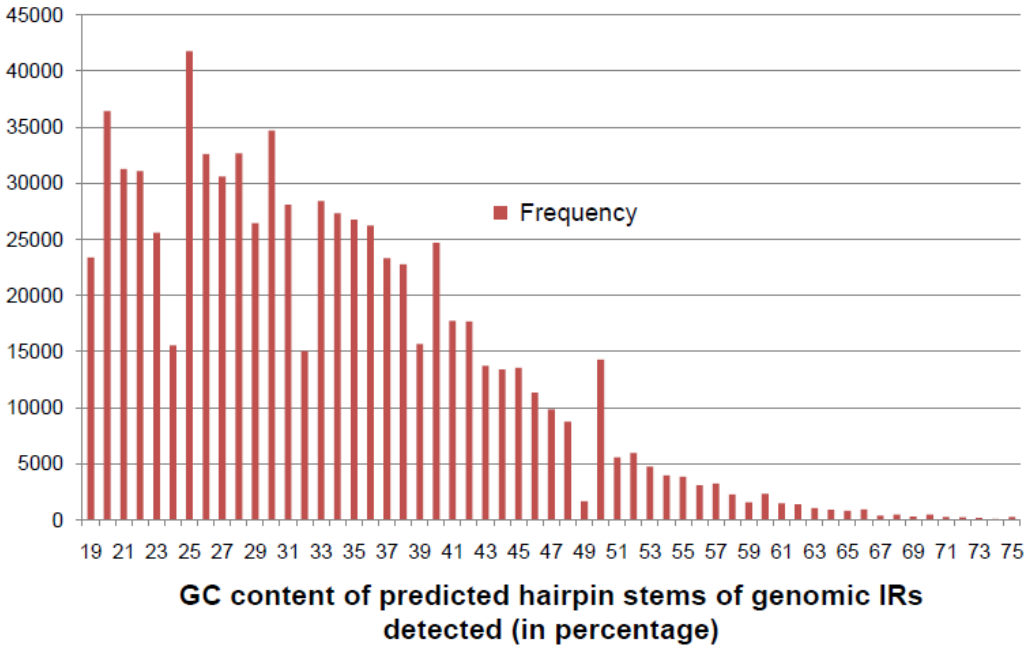

Figure 11. Frequency distribution of parameter *G* on genomic *At* IRs

Shows the frequency distribution of GC content density (*G*) in IRs detected from genomic DNA of *Arabidopsis thaliana* with base thresholds of D_{min} = 59% and P_{max} = 9.99×10^{-9}, A_{min} = 2.0, and G_{min} = 19. The values of G for each pre-miRNA IR has been floored to the closest whole number lesser than it.

4. CONCLUSION

Initially, our study revealed that partially symmetric inverted repeats are abundant in genomic DNA. However, we showed that most of these IRs are easily distinguishable from the IRs of known pre-miRNA and can be filtered out using generic criteria like density of symmetry, statistical probability of symmetry, average length of symmetric regions, and the GC content of sufficiently symmetric regions. It is then reasonable to assume that more accurate filters that are highly specific to certain kinds of ncRNA will retain a smaller final list of IRs that can then be further analysed using wet lab techniques such as northern blotting to identify novel ncRNA genes. The irScan software framework was designed to be easily expandable with such additional filtering criteria, by anyone with experience in the Perl programming language. The more computationally demanding IR detector algorithm was implemented in C++ and parallelized to be able to scan the whole Arabidopsis genome for IRs in less than a minute using a base set of filters. A user could then filter these IRs further by running various combinations of filters using Perl to find an optimal set of filters and parameters, that minimizes the number of IR candidates while maximizing the number of known ncRNAs identified.

Additional filters are required to further enrich the final set of IRs with those that are more likely to be functional ncRNA, while still retaining most of the known ncRNA. Some such filters include the detection of promoters and terminators, homology analyses, location of candidate relative to coding regions, and better secondary structure prediction algorithms. The software developed is designed to easily accommodate such additional filters by someone with minimal experience in Perl, while the computationally expensive underlying genome-wide scanning algorithms have been implemented in the more efficient C++ programming language.

5. COMPETING INTERESTS

The authors declare that they have no competing interests.

6. AUTHORS' CONTRIBUTIONS

CK participated in the conception and design of the study, developed all the software used for preliminary analyses, and designed and developed the Perl/C++ based irScan framework for ncRNA identification. RLF participated in the conception and design of the study and the biological aspects of miRNA analysis. FE participated in the conception, design, and the computational aspects of the irScan framework. All authors read and approved the final manuscript.

7. REFERENCES

1. Huttenhofer A, Schattner P, Polacek N: **Non-coding RNAs: hope or hype**. *Trends Genet* 2005, **21**(Suppl 5): 289-297
2. Machado-Lima A, del Portillo HA, Durham AM: **Computational methods in noncoding RNA research.** *J Math Biol* 2008, **56**: 15-49
3. Liu C, Bai B, Skogerbo G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R: **NONCODE: an integrated knowledge database of non-coding RNAs**. *Nucleic Acids Res* 2005, **33**: D112-D115
4. Meng Y, Huang F, Shi Q, Cao J, Chen D, Zhang J, Ni J, Wu P, Chen M: **Genome-wide survey of rice microRNAs and microRNA-target pairs in the root of a novel auxin-resistant mutant**, *Planta* 2009, **230**: 883-898
5. Lee Y, Jeon K, Lee JT, Kim S, Kim VN: **MiRNA maturation: stepwise processing and subcellular localization.** *EMBO J* 2002, **21**: 4663-4670
6. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN: **The nuclear RNase III Drosha initiates miRNA processing.** *Nature* 2003, **425**: 415-419
7. Ruby JG, Jan CH, Bartel DP: **Intronic microRNA precursors that bypass Drosha processing.** *Nature* 2007, **448**: 83-86
8. Zhou X, Ruan J, Wang G, Zhang W: **Characterization and Identification of MicroRNA Core Promoters in Four Model Species.** *PLoS Comput Biol* 2007, **3**(Suppl 3): e37

9. Artzi S, Kiezun A, Shomron N: **miRNAminer: a tool for homologous microRNA gene search.** *BMC Bioinformatics* 2008, **9**: 39
10. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP: **Vertebrate microRNA genes.** *Science* 2003, **299**: 1540
11. Lai EC, Tomancak P, Williams RW, Rubin GM: **Computational identification of Drosophila microRNA genes.** *Genome Biol* 2003, **4**: R42
12. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The micro-RNAs of Caenorhabditis elegans.** *Genes & Dev* 2003, **17**: 991-1008
13. Terai G, Komori T, Asai K, Kin T: **miRRim: A novel system to find conserved miRNAs with high sensitivity and specificity.** *RNA* 2007, **13**: 2081-2090
14. Clote P, Ferré F, Kranakis E, Krizanc D: **Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency.** *RNA* 2005, **11**(Suppl 5): 578-91
15. Lin CT, Lin WH, Lyu YL, Whang-Peng J: **Inverted repeats as genetic elements for promoting DNA inverted duplication: implications in gene amplification.** *Nucleic Acids Res* 2001, **29:** 3529-3538
16. Smith TF, Waterman MS: **Identification of Common Molecular Subsequences.** *J Mol Biol* 1981, **147:** 195-197
17. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, **34:** D140-D144
18. Bonnet E, Wuyts J, Rouze P, Van de Peer Y: **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.** *Bioinformatics* 2004, **20:** 2911–2917
19. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL: **The Vienna RNA Websuite.** *Nucleic Acids Res* 2008, **36:** W70-W74
20. **The Arabidopsis Information Resource** [http://www.arabidopsis.org/]

VITA

Cyriac Kandoth was born in Kerala, India on March 23rd, 1984. His primary education was spread out at various schools in Glasgow (Scotland, UK), New Delhi (India), and Trivandrum (Kerala, India). His secondary education in the city of Cochin, India, placed a focus on Physics and Computer Science. In 2005, he completed a Bachelor's degree in Computer Science and Engineering from the Model Engineering College under the Cochin University of Science and Technology. After 6 months working in the Software industry, Cyriac joined the University of Missouri - Rolla, USA (now known as Missouri S&T) and on December 2007, received his Master of Science in Computer Science, with emphasis on Bioinformatics (specifically gene identification).

Cyriac then pursued a Doctor of Philosophy with a continued focus on gene identification, in particular analyzing inverted symmetry in DNA sequences indicative of cellular functionality. He received his PhD from the Department of Computer Science at Missouri University of Science and Technology (formerly UMR) in Summer 2010. His other research interests include Automated Surveillance Systems, Quantum Computing, Artificial Intelligence, Graphics Processing technologies, Parallel Processing, and pretty much anything that ends with -logy, -ysics, or -ience.