2008-08-12

Domain Duplication, Darwinian Selection, and the Origin of the Globulin Seed Storage Proteins

Nathaniel S. Cannon
Brigham Young University - Provo

DOMAIN DUPLICATION, DARWINIAN SELECTION AND THE ORIGINS OF THE SEED STORAGE GLOBULINS

by

Nathaniel Scott Cannon

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Plant and Wildlife Sciences

Brigham Young University

December 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Nathaniel Scott Cannon

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date

Craig E. Coleman, Chair

Date

David A. McClellan

Date

Mikel R. Stevens

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Nathaniel Scott Cannon in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Craig E. Coleman
Chair, Graduate Committee

Accepted for the Department

Date

Loreen A. Woolstenhulme
Graduate Coordinator
Department of Plant and Wildlife Sciences

Accepted for the College

Date

Rodney J. Brown
Dean, College of Life Sciences
Brigham Young University

ABSTRACT

Domain Duplication, Darwinian Selection and the Origins of the Seed Storage Globulins

Nathaniel Scott Cannon

Department of Plant and Wildlife Sciences

Master of Science

The seed storage globulins found among virtually all spermatophytes comprise a multi-gene family of proteins with ancient evolutionary origins. The two main groups of storage globulins include the legumins (11S) and vicilins (7S), both of which play a main role in protein deposition and storage in the seed endosperm. Composed of two cupin domains (bicupin), these proteins have been recently noted not only for their close structural relationships among the two subfamilies (7S and 11S) but also for their similarity to other proteins such as germin-like proteins (GLP's), bacterial oxalate decarboxylases, and other cupin containing proteins. Previous studies have investigated the evolutionary relationships among the legumin and vicilin groups, as well as their presumed evolutionary link to other cupin containing proteins; however these have each come short of any comprehensive resolved evolutionary history of the globulin family.

This study focuses first on resolving the relationships among the cupin super-family in relation to the storage globulins, as well as the GLP's, which have been postulated to be the single domain ancestors of the bicupin storage globulins. Nucleotide coding sequences for both N-terminus and C-terminus cupin domains of the storage globulins, including conserved non-cupin domain helical repeats and inter-domain spacers were aligned to a comparably sized set of single cupin coding sequences (CDS).

The phylogenetic relationships among the two globulin domains as well as the single cupin genes were elucidated using Bayesian inference of tree likelihoods.

Further phylogenetic analysis was performed on the complete CDS's for all storage globulin sequences in the study, using an appropriate out-group of similar overall domain architecture determined by the overall topology of the cupin super-family. This globulin muti-gene tree was used, along with an alignment corresponding to structurally resolved portions of the mature globulin peptides, to perform an analysis of patterns of selection among the various lineages of cupin-containing globulins.

The results of these analyses provide evidence for a common origin of all cupin containing genes. The GLP and storage globulin domains do not appear to be immediate ancestors of one another, but are grouped with the fungal spherulins as well, suggesting that the single cupin genes which gave rise to these groups had already diverged prior to the rise of land plants. The storage globulin gene tree provides evidence supporting the notion that true legumins and vicilins were recruited as seed storage proteins independent of one another, after their divergence. This is evidenced by the fact that they comprise two separate groups each with basal non-storage 11S/7S-like proteins.

Additional insight into the differentiating selection pressures provides a clearer picture of how similar suites of physicochemical properties came under selection after the recruitment of the 11S and 7S families as seed specific proteins. Regions under strong destabilizing selection correspond to regions known to be of importance in the overall structure of storage globulins. Strong destabilizing selection at the pore of the globulin subunit suggests that this region may have undergone more functional diversification than previously thought to have occurred among the legumins and vicilins.

TABLE OF CONTENTS

Introduction ... 1
Globulin Gene Family ... 1
Related Germins ... 2
Globulin Origins ... 3
Cupin Super-family ... 4
Hypotheses, Defining and Testing ... 5
Inference of Natural Selection ... 9
Property Specific Selection ... 13
Theoretical Methodology ... 15
Data Gathering ... 15
Domain Prediction ... 15
Taxon Sampling ... 17
Bicupin alignment ... 18
Single cupin alignment ... 19
Codon Position-Specific Model Testing ... 20
Phylogenetic Tree Searching & Support ... 21
Patterns of Selection ... 24
A Posteriori Validation of Alignment ... 25
Codon Partitioning ... 28
Results ... 29
Cupin Domain Identification ... 29
Taxon sampling ... 30
Sequence Alignment ... 31
Model Selection ... 34
Phylogenetic Reconstruction ... 36
Selection Analysis ... 41
Discussion ... 45
Multiple Sequence Alignment ... 45
Cupin Phylogenetics ... 49
Selection Analysis ... 55
Conclusion ... 59
References ... 62
Tables and Figures ... 72
Table 1. ... 72
Table 2. ... 72
Figure 1. ... 77
Figure 2. ... 78
Figure 3. ... 79
Figure 4. ... 80

Figure 5. 81
Figure 6. 82
Figure 7. 83
Figure 8. 84
Figure 9. 85
Figure 10. 86
Figure 11. 87
Figure 12. 88
Figure 13. 89
Figure 14. 90
Figure 15. 91
Figure 16. 92
Figure 17. 93

INTRODUCTION

Globulin Gene Family

Seed storage globulins comprise a family of proteins found among almost all angiosperms. Pre-proteins of the storage globulins are translocated to the endoplasmic reticulum where they are processed into mature proteins. These proteins are stored as aggregate bodies in vacuoles localized either in cotyledon or endosperm tissue of developing seeds (Casey 1999). During germination vacuolar processing enzymes (VPE's) alter the proteins' conformation, opening them to unlimited proteolysis. Degradation of the storage globulins by these VPE's provides the developing plant with a supply of elements and amino acids essential to growth (Muntz et al. 2002).

The globulin gene sequence encodes two cupin domains, and the gene product forms a radially symmetric homodimer. Each of these homodimers combines with two others to form a radially symmetric trimer (Hirano et al. 1985). This is accomplished via non-covalent bonding between hydrophobic regions in a pocket formed by a helical region at either end of the bicupin structure. In the case of the 7S globulin family, also known as vicilins, this is the final quaternary product. In the case of the larger 11S legumins the trimers stack in pairs to form hexamers.

The cupin domains found in the protein products of 7S and 11S genes share remarkable structural similarities at the tertiary level. Though the primary sequence structure seems to deviate significantly in several areas, the general cupin motif can be found in the peptide sequence of both domains. Each domain has a cupin motif followed by a helix and turn region. The cupin motif is composed of a single beta strand followed by an internal motif spacer and a second beta strand. These three elements form the main cupin beta-barrel, where the two beta-strands form an anti-parallel jelly-roll, with the inter motif spacer forming a hairpin turn at its center point. The motif spacer is of considerably different length in 11S and 7S versions, as well as across orthologous and paralogous copies of the globulins.

In both 7S and 11S globulins the bicupin subunits are held together through a series of bonding residues. In the 11S globulin a cystein disulfide bridge is formed between the two separate cupin domains, helping to stabilize the overall structure (Rodin et al. 1990). The peptide strand is cleaved leaving the two strands to be held only by these links. In the 7S the strands are held together by covalent bonds between the cupin barrels and the unit is not cleaved. It is thought that the final hexameric conformation may require additional structural stability, explaining this main difference between the 11S and 7S globulin classes (Adachi et al. 2003).

Related Germins

A closer look at the similarities among the cupin domains yields some insight into the origin of these proteins. The bicupin subunit itself has moderate radial symmetry, with the two cupin domains being able to super impose onto each other. As previously mentioned, the protein structure is composed of two cupin units, with similar domain architecture. It has been hypothesized that the bicupin globulin family arose from a domain duplication of a single cupin gene, followed by the duplication of that bicupin gene. Subsequent differentiation and radiation of the 7S and 11S sub-families followed (Shutov et al. 1995). This would have likely occurred previous to the rise of vascular plant since GLP’s are present among mosses, slime molds, vascular plants, and fungi in high copy number in some genomes (Dunwell et al. 2004; Lang et al. 2005).

A likely candidate for a single cupin origin of the bicupin globulins is found in the germin family. Germins, or germin-like proteins (GLPs) are a nearly ubiquitous protein family among plants. They were initially discovered in *Hordeum vulgare*, and named for their apparent importance in seed germination (Grzelczak and Lane 1984). Since their discovery germins have been found to be present in a variety of tissue types during stress conditions, germination, seed development, and in anti-microbial defense (Caliskan 2000). The GLP family can be broken down into sub-families that include oxalate oxidases, manganese oxide dismutases, and auxin-binding (Bernier and Berna 2001). It is still unclear whether these functional groups can be delineated along phylogenetic groupings, or whether many of them are simply multifunctional. Many of the GLP proteins' functions have only been putatively determined based on sequence homology, so to make any definitive statement in regards to the correlation of evolutionary history with function would at this time be premature.

Globulin Origins

The final protein product of the GLP genes is also composed of six cupin units , forming a protein ring of beta-barrels, as in the 7S or 11S trimers (Adachi et al. 2003; Adachi et al. 2001). On the other hand, the GLP nucleotide sequence only contains a single cupin domain. This seems to be an indication of the possibility that the 7S and 11S globulins arose from an original single-cupin gene, which had similar quaternary structure to the extant GLP group.

Alternative models propose the sequence of the domain and gene duplications that would have occurred from a single-cupin gene to two families of bicupin genes. Shutov et al. have proposed and investigated, over the course of multiple studies, the merits and weaknesses of two models to explain the origin of the globulin storage proteins (Shutov et al. 1995; Shutov et al. 2003; Shutov et al. 1998a; Shutov et al. 1998b). Both models involve the duplication of a single-cupin gene, which became a single gene. In one model this was followed by the duplication and differentiation of that gene into two families, the legumins and vicilins. 11S and 7S proteins gained their storage specific functions independent of each other. In an alternate model there was a minimum of three gene duplication of an original single-cupin gene. This small gene cluster was duplicated a minimum of two times, which ultimately led to the main 7S and 11S division. Within these groups there was independent loss of different cupin genes which ultimately resulted in two main groups (7S and 11S) whose cupin domains were not strictly homologous.

Either of these models can be tested by a phylogenetic analysis of conserved regions within the different domains. Previous studies have taken a simple distance based approach relying on average genetic distance between a sample of legumin and vicilin cupin domains as support for the latter model. Distance based approaches have incorporated neighbor-joining phylogenies of the conserved cupin motif from each of the two domains in a sample of vicilins and legumins amino acid sequences (Shutov and Baumlein 1999).

Cupin Super-family

The proposed models of domain duplication near the origin of the plant storage globulins hint at a connection with some more ancient relationships with single cupin proteins (Dunwell et al. 2001). The cupin fold, which describes the beta coil domain of the globulins, as well as the single cupin GLPs, is found in a wide array of proteins.

These range from bacterial oxalate decarboxylases, to fungal phosphatase isomerases; some are found in connection with DNA binding and retrotransposition genes, while others, though conserved, are of as yet unknown function (Khuri et al. 2001). Among these genes we find single and double cupins, as well as genes with a chimeric structure containing a cupin domain coupled with some other functional domain(s).

There is a high level of tertiary structural similarity among the cupin domains of these various proteins (Dunwell et al. 2000), and it has been postulated that, given this strong structural similarity, all cupin domain containing genes share a point of common origin. Under this assumption we can create a framework for investigating a phylogeny of the cupin clan. A reliable phylogeny of these cupin containing genes could shed light on the origins of the storage globulins. At the same time it must be recognized that previous studies of cupin evolution have been inconclusive in regards to the origin of these genes, and so work remains to be done to determine their place within this larger group (Dunwell et al. 2001; Dunwell et al. 2004; Khuri et al. 2001). Once this has been reasonably established we can then begin to test hypothesis regarding various questions about the cupin genes in higher plants.

Hypotheses, Defining and Testing

The overarching question of globulin evolution can be broken down into several smaller questions, which, once answered, can be knit together to form a more comprehensive view of their evolutionary history. This in turn might serve as one model among many of how gene families originate, persist and diversify. The origin of the globulin family, its division into groups, and its relationship to other similar proteins all cover different aspects of the history we need to uncover to better understand the origins of seed storage proteins in plants.

The questions regarding the origins and diversification of the storage globulins is first rooted in their relationship to other cupins. By performing a thorough phylogenetic analysis of the cupin containing genes we can determine at what point the globulins diverged, and we can infer their relationship to other structurally homologous proteins such as the GLPs. We can answer questions such as whether globulins are more closely related to the single cupin plant GLPs or the bacterial bicupin oxalate decarboxylases.

A significant problem immediately presents itself at this point by virtue of the fact that we are now considering the comparison of sequences of proteins with different numbers of homologous domains. This problem can present its own solution, as well as provide the method for answering another essential question; when did globulins become bicupin proteins? If we treat each cupin domain of any bicupin gene as a homologous structure which can evolve independent of the other domain then we can ascertain at what point the presupposed domain duplications took place which gave rise to the bicupin globulins (as well as other non-plant bicupin proteins). Since the two cupin domains share similar primary and tertiary structure it is reasonable to assume we can recover and compare homologous regions by sequence alignment. If bicupin genes are the result of the duplication and subsequent merging of a single cupin gene then it is not only reasonable, but necessary to compare each cupin domain independently. In doing this we can reconstruct the sequence of events that lead to the architecture of the globulin genes in their present state.

Once these questions have been resolved it becomes a theoretically straightforward task to reconstruct a reliable phylogeny of entire storage globulin sequences, rooted with an appropriate out-group as determined by single domain comparisons. A detailed phylogeny of the gene family has not yet been produced with high enough resolution to outline subfamilies and patterns of further gene duplication. In plants whose genomes have been fully sequenced several copies of globulin genes have been found (Okita et al. 1989), and so at this point it becomes necessary to begin to elucidate these more recent duplication events with more detail.

It is not enough, however, to simply point out divergence events and patterns of duplication. In order to understand the evolution of globulins we need to understand what significant changes are occurring across the gene family. Using a well supported, resolved phylogeny and a reliable alignment of globulin sequences an analysis of patterns of selection becomes possible. Historically, analyses of selection compared ratios of synonymous and non-synonymous substitution rates (Fares 2002; Kosakovsky Pond and Frost 2005; Nei 2005; Nielsen and Huelsenbeck 2002; Yang and Bielawski 2000; Yang and Nielsen 2002). These rates, adjusted for models of substitution and codon biases can be used to generate a statistic omega which can be used to test for significant directional selection. While this is a useful approach in identifying wide-spread directional changes it has its shortcomings. Another approach to identifying selection is to analyze only the non-synonymous substitutions for patterns of radical or minimal changes which exceed the normal expectations under neutral theory (McClellan and McCracken 2001). More detail on these topics is given in the section on selection.

By using these two theoretically different approaches together a more complete picture is produced of the patterns and history the selective forces acting on a gene over time. In so doing one can identify ways in which the globulin sub-families diverged from one another, as well as identify whether the original domain duplication of the proto-globulin cupin was followed by directional or stabilizing selection. Presumably, as the ancient globulins became integrated into the complex system of seed bearing reproduction this would leave a hallmark pattern of destabilization recognizable by radical changes in physicochemical properties and an abundance of non-synonymous substitutions.

Furthermore, analyses of patterns left behind by natural selection can shed light on whether the 7S and 11S subfamilies of the globulin storage proteins evolved their storage capacity independently, or whether this is a derived trait. Correlation of patterns of directional selection in 7S and 11S families would indicate that they both evolved storage capacity separately, while significant stabilizing selection would indicate that certain features common to 7S and 11S, were in place prior to their division and that their common ancestor likely functioned as some rudimentary nutrient store. One must be cautions in the interpretation of the results of these types of analyses, taking care to apply them only to structural/functional sites that have been well characterized. Inevitably there will be several sites under directional selection and several sites under stabilizing selection, and any blanket statement about the evolution of the gene family without putting all results into their proper context would be unwise. Nevertheless, it is entirely likely that certain regions of these proteins have been under different evolutionary pressures, which could ultimately lead to a mixed signal of directional and stabilizing selection. Again, placing the results into the context of the proteins' function, as well as placing significant mutational events into a phylogenetic context one is able to reconstruct putative patterns of functional maintenance and adaptation among homologs.

Inference of Natural Selection

Identifying positively selected amino acid sites is an important approach for making inference about the function of proteins; an amino acid site that is undergoing - positive selection is likely to play a key role in the function of the protein. Thus, inferring selection at amino acid sites within proteins has become an integral part of investigating protein structure and function. As models to detect and characterize the effects of selection on a molecular level improve, a clearer picture of evolution at the molecular level will emerge.

Two opposing views with regard to how selection acts on the molecular level are selectionism and neutralism. The vast majority of observed substitutions in proteins are neutral suggesting that functional change in proteins is the result of a few key amino acid substitutions (Nei 2005). The problem of natural selection on a molecular level has been reduced to a quantitative one in that these key changes can be observed in pair-wise comparisons of DNA or amino acid sequences. Estimates of selection can then be calculated based on the amount of non-synonymous substitutions necessary to account for the difference between two sequences.

Synonymous and nonsynonymous substitution rates (dS, dN) are the total numbers of silent and variable mutations at a specific codon site in a pair-wise comparison of two DNA sequences. The ratio w = dN/dS measures the relative proportion of synonymous and nonsynonymous differences between the two sequences.

If dN < dS, it can be said that non-synonymous mutations have occurred at a slower rate than synonymous mutations. This may be indicative of selective constraint (purifying or stabilizing selection) because the majority of mutations at a given site have been limited to changes which maintain a similar physico-chemical composition within a given region. If dN = dS then we assume nonsynonymous and synonymous mutations are occurring at equal rates, which may indicate that nonsynonymous substitutions are neutral, or in other words neither deleterious nor effecting protein function. This case fits best with a null model of no selection. If dN > dS then nonsynonymous mutations are occurring more frequently than synonymous mutations. This is evidence for positive selection because we assume that natural selection is acting on the amino acid sequence of the protein to retain changes in the protein introduced by nonsynonymous mutations.

Estimation of selection using dN/dS ratios is based on the following assumption: the codon is the unit of evolution, selection acts equally across an entire region of a gene, and non-synonymous substitutions are always an indication of selection. These assumptions can be problematic. For example; since codons cannot be viewed as independent of one another it may be impractical, not to mention a statistical faux pas, to treat codon mutations as independent events. In addition, a relatively few number of mutations around one site can lead to drastic functional changes, and yet dN/dS ratios are incapable of detecting selection except by the overwhelming presence of non-synonymous mutations. Weaknesses inherent in the assumptions, along with other short comings of using dN/dS ratios (such as the inability to quantify the strength of selection at a given site) have spurred the development of a variety of alternatives to detecting selection in protein sequences.

Patterns of codon bias may cause potential problems with respect to estimating dN and dS values due to the fact that the patterns of substitution likelihoods introduce noise into the data in the form of substitution bias. Highly biased codon usage can be caused by both mutational bias and selection greatly affecting synonymous substitution rates (Yang and Bielawski 2000). By rejecting the assumption that all synonymous substitutions are neutral and including parameters to estimate the likelihood of a given nucleotide substitution, these substitution rates can be corrected for.

This likelihood based approach is implemented by generating a rate matrix based on the distribution of nucleotide substitutions in the data (Wayne and Simonsen 1998). This rate matrix produces a likelihood of each substitution occurring given the estimated pattern of random substitution, and uses these probabilities to weight synonymous and nonsynonymous mutations in calculating dN/dS. The main pitfall of this method is that it still relies exclusively on average dN/dS rates over an entire sequence, making it difficult to estimate selection at a limited number of sites.

Most amino acids in a protein are under structural and functional constraints. Therefore natural selection is likely to only retain advantageous mutations a few sites at a time. It then follows that the generalized approach of averaging dN/dS rates over entire sequences has little power. Models that would allow the *w* ratio to vary among sites within a set of sequences could then detect selection acting on a limited number of sites, within specific regions of a protein (Yang and Nielsen 2002). A combination of Maximum Likelihood and Bayesian statistical models are implemented to estimate dN/dS rates (Scheffler and Seoighe 2005). These estimates are then used to test for selection at each codon site along the protein sequence (Nielsen and Huelsenbeck 2002).

This model allows for independent estimates of *w* at each site along the amino acid chain, as well as tests for statistically significant levels of selection at each site. Although this can help increase the overall power of the dN/dS test by allowing for higher resolution, the main assumption of the null hypothesis for this model is that there is a constant rate of synonymous substitutions across the entire sequence. This assumption has been shown to be flawed (Massingham and Goldman 2005).

There is a higher than expected rate of false-positive results when testing for selection at each codon site due to the dramatically inflated number of tests being performed (Massingham and Goldman 2005). The site-wise likelihood ratio test (SLR) is designed to reduce false-positives to a reasonable rate pursuant to the order of hypotheses being tested. This increase in robustness is built into the SLR model by reducing the underlying assumptions to disregard preconceived notions about the distribution of site mutations. The weaker assumptions allow it to be more applicable to data for which pre-calculated substitution rates may not be relevant. Thus, when data violate this assumption as in the variable ratio method described above, there is a significant increase in false-positive error. By bringing down this error rate, SLR is able to estimate selection at each site using more parameters for higher precision, while retaining the capability of estimating the strength of selection at each site. Although the SLR model allows for variation in rates along a given sequence, it does not account for the possibility of a variation in substitution rates at a given site among sequences.

All of the above models are based on the assumption of selection acting independently on each codon in a DNA sequence. Co-evolution between different codons within functional domains calls into question the assumption of a single codon as the unit of selection (Fares 2002). It is therefore desirable to develop a model which can reduce the set of codons used in a given hypothesis test to those within the same functional or structural region of the site being tested. One approach to this is the use of a sliding window. Given that a set of adjacent codons may be under the same selective pressure, a window of size *l* is used to restrict the total number of codons used to calculate the dN/dS ratio to test the null hypothesis of $w = 0$ at a specific codon. Window size is determined based on probabilities for dS and dN calculated from the binomial distribution. This is a significant improvement over other random window size methods (Hughes and Nei 1989; Tajima 1991). In this manner region-specific selection can be determined by comparing the expected to the observed numbers of nucleotide substitutions within a section of the sequence more closely resembling the region(s) thought to be under selection (Fares 2002). This approach shares some of the same problems as the previous methods, including inflated false-positive rates.

Property Specific Selection

The assumption that $w > 1$ if positive selection has significantly influenced the evolution of a protein has been shown to be too conservative to detect many events where selection is operating within protein regions (Woolley et al. 2003). Furthermore, dN/dS ratios, while they may identify regional selection under some models, cannot categorize the properties under selection. In order to quantify the changes in physicochemical properties within regions under selection individual amino acid properties can be measured and compared at each site (Xia and Li 1998). With these pair-wise differences in property values, a new statistic can be derived, based not on whether substitutions are synonymous, but to what degree they differ in their physicochemical properties. The main advantage of this approach is that rather than simply locating regions under selection, one can distinguish not only between stabilizing and disruptive selection, but also characterize and quantify the degree to which non-synonymous mutations have altered the physicochemical properties within a given region (McClellan et al. 2005). This approach employs a sliding window which is only a best guess of the regions under selection. This method also increases false positive rates far more than previously discussed models since there are *n* amino acid sites multiplies by *x* physico-chemical properties being tested for every pair of sequences. This can be corrected for by using the Bonferroni method (Bland and Altman 1995), where the alpha level is decreased proportional to the number of tests being done. Although not ideal, so far this does not seem to be a problem in obtaining significant results, as they have been shown to correlate with results found by dN/dS ratio methods.

The advantage is that the same sites under selection are also categorized by magnitude of change and by the property under selection. For the most part, all of the major variations on the central dN/dS model yield strikingly similar results, suggesting either that they all consistently find correct results, or that none is a real improvement on the others (Kosakovsky Pond and Frost 2005). In light of these findings this study has incorporated analysis using TreeSAAP (McClellan and McCracken 2001; McClellan et al. 2005; Woolley et al. 2003),so as to be able to identify not only sites under selection, but be able to determine between sites under stabilizing and destabilizing selection for a variety of physicochemical properties.

THEORETICAL METHODOLOGY

Data Gathering

Coding sequence (CDS) data corresponding to nucleic and/or peptide sequences is used in phylogenetic analyses in various studies of cupin domain containing genes (Balzotti et al. 2008; Beyer et al. 2002; Bharali and Chrungoo 2003; Carter et al. 1998; Cho and Nielsen 1989; Domoney and Casey 1985; Dunwell et al. 2001; Dunwell et al. 2000; Dunwell et al. 2004; Fischer et al. 1995; Galau et al. 1991; Gatehouse et al. 1988; Hager and Wind 1997; Hayashi et al. 1988; Khuri et al. 2001; Lang et al. 2005; Lycett et al. 1984; Mathieu et al. 2003; Mediana-Godoy et al. 2004; Membre et al. 1997; Merchant et al. 2007; Nakata et al. 2004; Okita et al. 1989; Rodin et al. 1990; Ryan et al. 1989; Samardzic et al. 2004; Shutov et al. 1995; Shutov et al. 2003; Shutov et al. 1998a; Shutov et al. 1998b; Tai et al. 1999; Takaiwa et al. 1987; Takaiwa et al. 1986; Weng et al. 1995; Wind and Hager 1996; Zimmermann et al. 2006). Sequence data from among these studies as well as CDS sequence data obtained from GenBank records using BLAST (Altschul et al. 1997) have been included in the present study. All annotated cupin domain containing genes from the *Arabidopsis* and *Oryza* genomes were downloaded for use in automated BLAST searches. The BLAST html output was parsed and the CDS sequence data for each record was downloaded. No additional *Arabidopsis* or *Oryza* cupin containing genes were found using this method. All other full CDS sequences corresponding to species within the *Viridiplantae* were retained.

Domain Prediction

The most common approach to gathering sequence data where it may not be readily available through laboratory methods is to perform a BLAST search using a sequence of known identity as a seed. One disadvantage is that because search parameters can affect different searches in different ways it can be difficult to ensure that results always return homologous sequences. In the case of the globulin storage proteins there is yet much to be done in the way of annotating and certifying the function, structure, and classification of putative globulins. In light of this a way to predict with some level of certainty whether a sequence is in fact a bicupin storage protein would be valuable. Hidden Markov models (HMM) provide a way to do this. A HMM is a statistical probability construct which allows sequential data to be analyzed and compared to an a priori model to assess how well that data fits the model. HMMs have a wide variety of applications, and lend themselves particularly well to measuring the similarity of a sequence to a motif in terms of probability of fit.

This data set of over 500 sequences was analyzed using HMMER (Eddy 1998). HMMER employs HMMs to determine the degree to which a given sequence matches a specified motif based on a pre-constructed PFam Markov model. The Cupin_1 domain model (PF00190) was implemented and all sequences with E-values >> 0.001 were rejected as not containing a cupin type 1 domain. Domain architecture was verified for the remaining sequences using the Conserved Domain Database (CDD) search algorithm. CDD uses reverse position-specific BLAST searching of a query sequence against a database of sequences with known and annotated domains (Marchler-Bauer et al. 2002). Searching was done using the CDD 2.13 database which contains over 24,000 position-specific score matrices. Based on the results of these analyses the sequences were divided in to two main data sets; one including all single cupin genes, the other containing all bicupin genes. Only coding sequences with complete and unambiguous domain architectures were retained for further analysis.

Taxon Sampling

Preliminary alignment and phylogenetic analysis for the purpose of taxon selection was conducted using MUSCLE (Robert 2004) and GARLI (Zwickl 2006) for both single and bicupin data sets. The nucleic acid sequences were first translated using AlignmentHelper 1.2 (http://biology.byu.edu/faculty/dam83/cdm) and then aligned using the default parameters of MUSCLE. The multiple sequence alignments (MSA) were then reverse translated using AlignmentHelper thus allowing alignment at the more highly conserved amino acid level, and maintaining the integrity of the reading frame and the information at the nucleotide level.

Maximum likelihood (ML) phylogenetic tree reconstruction carried out by GARLI was done under the general time-reversible (GTR) model of evolution allowing the program to estimate rates of substitution and nucleotide frequencies. Monophyletic groups/pairs of putative paralogs were observed with high frequency among the *Arabidopsis* and *Oryza* sequences. In these cases a single representative sequence was retained. These small groups likely represent clusters of recently and self duplicating genes.   It would therefore be unnecessary to include all of them in a study of selection on deep branches.   Furthermore, inclusion of too many highly similar sequences within one or another gene subfamily may bias the estimation of the parameters within a given model of evolution toward that particular group. Among highly divergent sequences any bias towards one group may result in an inaccurate reconstruction of the topology in other sections of the tree.

An effort was made to maintain some taxonomic parallelism within 7S, 11S and GLP data sets, as identified by preliminary neighbor-joining (NJ) phylogenetic analyses, so as to minimize the effects of speciation in analyzing patterns of selection within the gene family. The more species are represented among all clades of a paralagous gene superfamily the less the topology of that tree is based on the sampling (or lack of sampling) of given species. This approach also served to give some preliminary validation to the question of homology raised in the previous section. Any sequences remaining post HMM analysis which did not align well or clearly group within a well defined clade of storage proteins was excluded from further analysis. The resultant data set contained 87 bicupins and 88 single cupin sequences from the *Viridiplantae*, as well as several N and C-terminus cupin domains from bicupin oxalate decarboxylases and gentisates from a previous study on the evolution of the cupin super-family.

Bicupin alignment

Once a final data set had been created using the aforementioned methods a final alignment of homologous regions was needed. Different alignment methods often result in vastly different solutions. This can in turn affect the predicted topology of a phylogenetic tree. In order to investigate the extent to which this data set was affected by diffeent alignment strategies the following protocols were employed.

The bicupin data set was aligned using MUSCLE (-noanchors, -maxiters 9999, -maxtrees 9999), and MAFFT (Katoh and Toh 2005) (LINSI, GINSI, EINSI, with -maxiter 9999, and -fmodel), producing a total of four unique MSA's. The CDS for the bicupin oxalate decarboxylase gene of Bacillus *subtilus* was included in the alignment. The four alignments were visualized in MEGA 4.0 (Tamura et al. 2007) and checked for general reliability based on gap tendency in presumably conserved domains, mis-alignment of clearly non-homologous regions, and identification of conserved sites in functionally important regions.

The MAFFT-linsi and MUSCLE MSA's were imported into ClustalX 1.83 (Thompson et al. 1997) and column quality scores were generated based on a scoring parameter of 10 (most relaxed). Column scores were exported and all sites with scores lower than 25/100 were removed from the alignment. Care was taken to preserve the reading frame of the alignment. The resulting trimmed alignments were exported for use in phylogenetic and selection analyses.

Single cupin alignment

In order to address the questions of domain duplication and divergent evolution among putatively ancestral single and bicupin genes it was necessary to treat each cupin domain in the bicupin genes as an independent evolutionary unit. In order to address the problem of splitting the cupin domains of the seed globulin genes their MSA's were profile aligned to a structural alignment based on two 11S and two 7S Protein Data Bank (PDB) structural records (Berman et al. 2000). PROMALS3D (Pei et al. 2008) was used to align the peptide sequences of the following proteins structures deposited in PDB: 2EVX, 1FXZ, 2CAV and 2PHL. The alignment was based on an algorithm which minimizes the root mean square distance (RMSD) value while maximizing sequence identities. The corresponding nucleotide sequences were obtained from GenBank for use with AlignmentHelper.

The structurally based amino acid alignment was profile aligned to the translated MAFFT and MUSCLE bicupin MSAs using the profile option in MUSCLE. Forward and reverse translation was done in AlignmentHelper. The resulting MSA containing the bicupin data set and the profiled structural alignment was imported into MEGA where the region corresponding to the inter-domain spacer (IDS) was identified by visualizing the 3D alignment of the tertiary structure files aligned by VAST (Gibrat et al. 1996) in Cn3D. The MSA was edited and saved as two separate files, one containing the complete 5' chain (containing signal peptide, alpha-cupin domain and a helix domain) and the other containing the 3' chain (including the IDS, beta-cupin domain and a helix domain).

The PROMALS3D PDB sequences were then removed and the two data sets were added to the 88 other single cupin gene sequences yielding a large data set of 264 single cupin sequences. This data set was aligned using MAFFT (once each with ginsi, linsi, einsi, and -maxiters 9999) and MUSCLE (-noanchors, -maxiters 9999, -maxtrees 100). These four alignments were visually inspected for biological feasibility as described above. The MUSCLE and MAFFT-ginsi alignments were trimmed of highly ambiguous or gap-rich regions using ClustalX column scoring as outlined above.

Codon Position-Specific Model Testing

Substitution likelihood models are implemented within a Bayesian framework to obtain the posterior probability of a tree solution given the data. Previous studies have shown that when inferring tree topology from protein coding sequences independent estimation of model-parameter sets for codon position partitions can be advantageous. The major caveat of this approach is that introduction of multiple models can lead to increased error without gains in accuracy. In order to provide some grounds for an a priori assumption of independent models of evolution based on codon position the Akaike Information Criterion (AIC) was employed using ModelTest (Posada and

Crandall 1998) to evaluate the best set of model parameters for each codon position (CP) in each MSA. Under the assumption of the codon position being independent of the model of evolution we would expect ModelTest to find not only similar models for each CP but also that the parameter estimates would be similar across each CP partition. In the event that CP-based models or their parameters differed among partitions we might assume, a priori, that CP partitioning of the data would be an appropriate approach in a phylogenetic reconstruction.

An incongruence length difference (ILD) test (Farris et al. 1995), (also known as a partition homogeneity test) using 1000 replicates for all pair-wise CP partitions in PAUP* (Swofford 1999). The ILD test assesses the likelihood that two or more partitions within a MSA have evolved under the same conditions based on the proportion of parsimony trees of equal length recovered from each partition over N number of replicates. Under the assumption that two partitions share the same evolutionary history under an otherwise equal weighting scheme, the two partitions will recover parsimony trees of equal length the majority of the time. We can therefore assume that under a model of evolution independent of codon position there should be no significant difference in the topology of tree solutions derived from the three CP data partitions. Significant differences would indicate the need for CP specific weighting schemes (i.e. individually estimated model-parameter sets).

Phylogenetic Tree Searching & Support

The fundamental questions which are most easily investigated using tree topology have to do with resolving ancient as well as recent relationships among taxa. In this study the taxa in question are represented by individual cupin domains from globulin and other cupin containing proteins in one case, as well as entire globulin storage protein sequences in another case. In the prior case the question which phylogenetic analysis will resolve is whether each of the cupin domains in the globulins are monophyletic (ie descend from a common ancestral domain) and at what point these domains arose via a duplication event. In the latter case the question at hand is whether the two globulin sub-families (7S and 11S) are monophyletic.

The GTR+I+G or the GTR+G substitution rate models were implemented for each CP partition and a phylogenetic tree reconstruction was performed using MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001, 2005). Two MCMC runs with 8 chains each ran for 20M generations for a total of 320M trees searched for each MSA of both single cupin and bicupin alignments. Runs were sampled every 1000th generation and the tree with the highest -lnL among the active chains was saved along with its parameter estimates. It is assumed that as independent runs approach stationarity at -lnL distributions of equal mean and variance that they have each reached and are sampling from the global optima of the estimated tree space. The burn-in sample size was determined based on visual estimation of the convergence between each run's –lnL distribution using Tracer (Rambaut and Drummond 2006). The burn-in sample was excluded from future analysis in order to eliminate trees sampled from other than the true posterior distribution.

The log-likelihood scores of the combined trees recovered from each run (minus burn-in) are considered to be an accurate representation of the posterior probability distribution of the -lnL of the true tree. It follows by extension that a set of trees whose –lnL distribution approximates this distribution constitute a reliable estimate of the topology of the true tree. This distribution can be used to estimate branch lengths and infer pP supports for tree topology. Given a burn-in sample the SUMT option in MrBayes generates a bipartition support cladogram as well as a branch-length phylogram. Non-parametric bootstrap values for the resulting phylograms from each MSA were generated from 1000 maximum likelihood tree replicates using RAxML (Stamatakis 2006). RAxML bootstrap analysis was run under model settings similar to the MrBayes analysis, consisting of CP based partitions for which GTR model parameters were estimated independently.

RAxML allows the user to import two files, one containing an optimal tree such as a maximum likelihood tree or a Bayesian consensus tree and the other containing a set of trees such as might be derived from a bootstrap analysis or a set of trees from a Bayesian MCMC run. This tree set can then be used to generate bipartition support values (bootstrap or posterior probability) for the nodes/branches on the other tree. RAxML requires that the single tree be fully resolved and have branch lengths. A fully resolved majority rule consensus tree from the Bayesian MCMC searches of the two bicupin alignments was used to draw results from the RAxML bootstrap analyses from each of these data sets to generate bootstrap support values for each tree, respectively. The trees from the MCMC searches from the MAFFT and MUSCLE alignments of the globulin cupin domains were not fully resolved at the 0.5 level, and so the sumt analysis in MrBayes was run using the ALLCOMPAT option. This resolved tree was supplied for bootstrap analysis in the way described above. Trees are shown with both pP and bootstrap values in the results section.

Patterns of Selection

Neutral theory dictates that non-synonymous (dN) and synonymous (dS) substitution rates should be equal, relative to the number of potential substitutions in each category (Yang 1998). Using his assumption we can test whether the ratio ω = dN/dS = 1. It is assumed that for an ω > 1 there is stabilizing (also purifying) evolution occurring, and for an ω < 1 there is destabilizing (also directional) selection occurring. This is the approach used, with some slight model based variations, by most software to detect natural selection in coding sequences.

TreeSAAP is a program which employs an algorithm allowing the researcher to identify not only patterns of selection among the non-synonymous substitutions in an alignment but also the intensity of those changes among a suite of physicochemical amino acid properties. Each non-synonymous substitution can be described in terms of the quantitative shift in a given physicochemical property. All possible changes in physicochemical property due to a single nucleotide substitution can be plotted to yield a probability distribution which is then divided into magnitude categories. TreeSAAP allows the user to decide how many categories to divide the distribution into, and then perform an analysis on the multiple alignment across a provided phylogenetic tree to test to see if substitutions in any category occur more often than would be dictated by chance. In this way if a tree/alignment produces more frequent selection for increases in one property this can be taken as an indication of strong directional selection for that property. The researcher must be cautious in interpreting these results however, since weak directional selection can also be present, but may be mistaken for no strong selection whatsoever. Interpretation of results should always be in context of the phylogenetic tree; sites under selection should not be attributed to all sequences in the alignment; and overall patterns of selection should be used, rather than single event to establish patterns of natural selection in genes.

A Posteriori Validation of Alignment

Methodological assumptions are inherent in any molecular evolutionary analysis. These assumptions are often drawn from the theoretical framework of neutral theory, which provides an overarching paradigm from which researchers can draw upon and modify a fundamental set of null hypotheses to evaluate a variety of research questions. In this study assumptions are made regarding homology of gene sequences, correctness of multiple sequence alignments and the evolutionary models derived from them, as well as the phylogenetic solutions which best explain them. Validation of these assumptions is critical in any study.

The assumption of sequence homology may be viewed as tentative, despite best efforts to classify each gene sequence that was included through a variety of approaches. BLAST searches are used to infer sequence homology by local similarity, with high scoring pairs (HSPs) being significantly similar (homologous) by definition. It is a generally accepted approach to infer functional and structural similarity among nucleotide or peptide sequences based on local alignment scores. We further verified that many of the sequences used in our initial BLAST searching were well annotated and experimentally verified as containing legumin, vicilin, germin or other genes as expected. We further verified that the coding sequences of the gene records in question contained at least one cupin 1 domain by use of hidden markov models using HMMER as described above. Lastly, preliminary alignment and tree reconstruction can be used as an indicator of unintentional inclusion of non-homologous sequence in an alignment or phylogeny. The ClustalW and NJ algorithms were implemented in MEGA 4.0 in order to identify whether any sequences did not share a reasonable amount of sequence similarity.

Given a set of homologous sequences it is critical that an alignment of those sequences reveals site specific homology. As homologous sequences become more divergent site specific homology becomes more difficult to recover and different alignment algorithms can find widely differing alignment solutions. This can have cascading effects from model selection, to phylogenetic reconstruction and tests of selection. Therefore it is important to asses the differences between alignments produced through various methods. No standard test of significance has been defined for MSA's; however the degree to which they differ can be described in various ways.

The most common and long standing approach is to visually assess patterns found or missed by different algorithms. While this can be effective in initially ascertaining whether two algorithms have arrived at the same or sufficiently similar solutions there is no quantitative test, nor is there any way to set limits on sufficiency of similarity. Alignments can be described using basic statistics such as nucleotide diversity, total gaps, and total length. While these measures may be useful to characterize alignments or the genes/proteins they contain, it is difficult to decide to what extent differences in these statistics constitute a significant difference between alignments.

In the event that two alignments yield differing models of best fit, or widely differing parameter estimates, this is a good indicator that the alignments might yield potentially different interpretations of the evolutionary history of the sequences they contain. Again, no clear test of significance has been presented to address this problem.

Various alignment algorithms can be used to generate MSA's, and in turn phylogenetic trees can be generated from each alignment. Significant topological differences among the resulting trees are an indicator of possibly significant differences in the original alignments.

There are problems in defining statistical significance in this approach, because the most common tests such as the Shimodaira-Hasegawa (SH) test which test for significant difference between two tree likelihood scores, are based on the assumption that the two trees being tested are derived from the same data (e.g. MSA). Thus likelihood scores are directly comparable. However, in comparing two trees derived from different data there is no well defined statistical approach.

By presenting the results from the aforementioned approaches the reader is able to decide to what extent the results and conclusions of this study violate the assumptions inherent in its methods. A final test of significant difference between alignments is suggested herein. An a posteriori assessment of differences between alignments based on a thorough phylogenetic search such as a Bayesian MCMC is a reasonable validation of methods of model selection. An MCMC search can be used to produce mean and variance estimates for each model parameter. Once model parameter estimates are produced one can test for significant differences among the parameter estimates between two alignments in a pair wise fashion. Common adjustments for multiple comparisons in a single test (e.g. Tukey, Bonferroni, etc.) can be applied to prevent alpha inflation of type I errors.

Under the condition that an arbitrarily derived portion (as determined by the researcher) of the model parameters are found to be significantly different, it might be concluded that the approaches yielded significantly different alignments. Posterior probability (pP) distributions were generated from an MCMC search and analyzed in Tracer. A simple test of significance (α = 0.01) was performed for each parameter using the Bonferroni adjustment. The threshold of having 80% of parameters found significantly different was set in order to reject the null hypothesis that the model parameters generated from the two alignments were not different. This test was used to infer significant and practical difference between the different MSAs.

Codon Partitioning

Alignment partitioning is a common approach used in phylogenetics, especially where multiple genes or coding and non-coding regions are included. This approach can allow different regions to be aligned under different parameters, and allows independent estimation of models of substitution. Studies have shown that providing mixed models for partitioned data decreases bias in the phylogenetic search and more often recovers the true tree. It has also been suggested, with supporting evidence, that partitioning by codon position can increase the accuracy of a phylogenetic search if the model parameters can be re-estimated for each position independently. The main caveat of this approach is that a point of diminishing returns is reached in the results because as parameters are added to a model its error is also increased. There is no standard approach for testing whether alignment partitions have significantly different substitution models. In the event that a model testing algorithm returns different ideal models this can indicate that mixed models should be used, however this is not a direct test of significance. Although model parameter sets may differ significantly among partitions this is not necessarily an indication that the gains in accuracy by including mixed models outweighs the increase in error. This study’s approach is the same as outlined above; to estimate means and variances of model parameters and perform pair wise tests of significance among parameter sets. This approach necessitates that the full parameter GTR model be used, (or at least the same model among partitions) and if significant difference are found among parameter sets then mixed model partitioning is said to be justified.

RESULTS

Cupin Domain Identification

HMMER results (Table 1) were used to identify sequences from initial BLAST searches which were unreliably identified as globulin homologues. These often included incomplete CDS’s, non-globulin sequences with strong local alignments, and other putatively identified globulins with little sequence similarity to known globulins. Putative vicilins (7S), legumins (11S), GLPs, and non-storage cupin subsets of the initial data set were each subjected to HMM analysis independent of one another. In this way an E-value cut-off could be determined for each group independently, since it was not expected that each group would match the PFam motif to the same degree. In each case a clear cutoff was found, where a small number of sequences were scored with an E-value several orders of magnitude larger than the remaining sequences (see table 1). It was found that an E-value << 1e-10 separated low and high scoring sequences. Any sequence scoring higher than this was removed from the data set.

In addition to HMM analysis on the sequence data set the conserved domain database (CDD) (Marchler-Bauer et al. 2002) was used to verify domain architecture (Figure 1) of sequences as being either single or bicupin containing genes. Genes presumed to be, or annotated as, bicupins which had only one complete cupin domain based on CDD analysis were discarded. Sequences with partial or fragmentary hits were evaluated visually based on alignment with CDS's from known globulins using ClustalW in MEGA 4. Genes presumed to be single cupins (GLP's, PMI's, etc.) were also subjected to similar standards of domain identification prior to inclusion in the final dataset.

Taxon sampling

Taxon sampling is fundamentally important to building a reliable dataset for phylogenetic analysis of gene families. In the case of a tree containing orthologous, paralogous, and even non-homologous sequences spread across the kingdoms Plantae, Fungi and Monera it becomes critically important to establish a reasonable level of taxonomic parallelism among the gene sequences included. In other words, each gene subfamily ought to be given as balanced a representation as possible among the species selected. The more taxonomic balance can be had among the various sub-families, the less the overall topology will be influenced by taxonomic sampling bias.

Given the current volume of submitted gene sequences, which highly favors model organisms and those of economic importance, obtaining a sufficient parallel cross-sample of the plant kingdom for each of the three main genes included in this study (11S, 7S, GLP) is difficult. Several species were included in the study for which legumins, vicilins and germin-like genes were available. These include *Oryza sativa*, *Triticum aestivum*, *Zea mays*, *Arabidopsis thaliana*, *Gossypium hirsutum*, and *Pisum sativum*. Germin-like, vicilins, and legumins sequences were all found among the paraphyletic gymnospermata, among 11 species spanning the group from *Araucaria angustifolia* to

Zamia furfuracea. One vicilin-like gene has been found in the filicophytes, in *Matteuccia struthiopteris*, which are basal to the spermatophytes within the vascular plants. No other genes have been described which encode for bicupin globulin storage proteins below this taxonomic level.

GLP's are found among many disparate species including some fungi, mycetozoa, bryophyte, marchantiophyta, and chlorphyta, as well as among vascular plants. In this light, germins may be considered to be more ancient than the storage globulins, which are only present in the vascular plants. Based solely on sequence homology and taxonomy it is not possible to tell whether the GLP's are a true monophyletic gene family. One aim of the phylogenetic analysis is to determine whether all germins, plant and non-plant, form a single clade in a phylogenetic tree with other cupins, including the storage globulins.

In all 172 gene sequences were included representing 88 total unique species, 73 of which are from the kingdom *Viridiplantae*. There were 71 sequences for genes encoding GLP's, 54 legumin sequences, 33 vicilin sequences, 4 spherulin sequences, and 11 sequences from genes containing at least one cupin domain such as bacterial oxalate decarboxylases or phosphomannose isomerases. All of these sequences contain the full coding region of the mature peptide.

Sequence Alignment

The cupin domain sequences which were aligned using five different algorithms yielded total aligned lengths ranging from 2,856 (MUSCLE) to 3,522 (MAFFT-einsi) total positions. This length disparity was due mainly to the differences in the ways that the algorithms handled length variable, repetitive or ambiguous regions such as the inter motif spacer (IMS) of the 11S alpha, or the 7S c-terminus domains. This was noticeable in the alignment of the full bicupin gene sequences as well, with the MAFFT-linsi yielding the longest alignment (4,227 positions) and the MUSCLE alignment the shortest (4,056). The lower variability among the bicupin alignments was due to the fact that less genetic divergence was represented overall.

Due to the variability among the different alignments they appeared to have yielded quite different solutions upon visual inspection. However, in analyzing them to identify any consistency using their Q-scores from ClustalX 1.83 clear patterns emerge indicating that the alignments were not as different as first thought. The quality curve, with the Q-score plotted along the alignment position, is shown in figure 1. Highly similar blocks can be seen, and are an indication that along somewhat conserved domains even the two most different alignments shared much in common. Visual inspection of several of these common regions based on Q-score confirmed that in fact the same sites were being aligned by the different algorithms.

Homologous sites and regions of the gene sequences could be clearly defined across the various domains. Similarities among the domains can be identified by eye, such as a conserved “F-L-A-G” motif from position 159 to 162. Several highly conserved glycine sites are apparent (e.g. positions 35, 83, and 115, fig. 3), as well as sites conserved for the aromatic residues phenylalanine, tryptophan, and tyrosine (e.g. positions 79 and 182, fig. 3). General patterns of hydrophobic and hydrophilic regions can also be seen, and appear to be maintained across all aligned sequences to a greater or lesser degree.

These and other patterns remain evident in the alignment of the full gene CDS sequence of the legumins and vicilins (fig. 4). Along with being able to identify conserved sites one can also note the changes in site-wise conservation in different regions of the alignment. The IMS and IDS regions tend to be somewhat ambiguous and gap-prone. This is likely due to decreased evolutionary constraint, both in terms of conservation of amino acids and length variability. IMS regions are noted for having a general pattern of presence of an insertion in the 11S alpha-chain at IMS1 and the presence of an insertion at IMS2 in the 7S C-terminus domain. This commonality is made more interesting by virtue of there being some scant levels of sequence homology between the two, making this unusual asymmetry stand out within the alignment.

The overall symmetry of the coding region cannot be overlooked either. The clearly followed domain architecture with a beta-coil motif followed by an inter-motif spacer and another beta-coil motif is the basic cupin domain structure. The beta-coil of the cupin domain is followed by an alpha-helix-turn-helix region, which comprises the first of the joining arms of the trimer subunit. The inter-domain spacer, which is highly length- and sequence-variable is clearly distinguishable. The second cupin domain adheres strictly to the prescribed architecture, with the exception of being slightly shorter, and being flanked by a conserved C-terminus region. This overall gene architecture appears to be well conserved among all of the storage globulins in the alignments, suggesting there is very little deviation from the empirically deduced tertiary structure of the storage globulins.

Model Selection

Model selection analysis on the two MSAs, performed using ModelTest, indicated that the codon positions were evolving under different models of evolution, as did the ILD tests run in PAUP*. Although the risk of over-parameterization is always a factor to consider in model selection it was felt the following results and reasoning justified not only the use of the full GTR model, but also of applying it independently to each of three codon positions.

Models suggested by the hierarchical likelihood ration test (LRT) and by the Akaike Information Criterion (AIC) fell broadly into the category of general time reversible (GTR) models, including TrN, TIM, and GTR. The TVM model was the only other non-GTR model of best fit based on either the LRT or AIC results. Model selection using MrModelTest yielded the GTR model for each codon position. MrBayes estimates the value of each parameter in a given model, and after successive generations in an MCMC search the output can be tabulated using Tracer (Figures 5-6). The Tracer output can be used to identify which parameters within the model differ significantly from the others; a kind of model parameter landscape. The model parameter estimates can be used to evaluate whether the model selection process was reasonable, a posteriori. In a given model certain substitution rates or base frequency categories are set equal to each other. Posterior distributions of the model parameter estimates can be used to create a kind of ad hoc model based on whether estimates differed significantly from one another. This ad hoc model can be compared to the selected model to evaluate the reasonability of such a selection.

It was found that these ad hoc models not only differed from conventionally used models, but also differed among codon positions, suggesting that other models besides GTR might not only be under-parameterizing, but also mis-categorizing parameters. This would introduce significant bias into calculations of the log-likelihood of tree values. There are only a limited number of models available to perform a test of best fit using available software, and although they exist in increasing levels of parameter complexity they cannot take into account all possible combinations of substitution rates and base frequency categories. Although the hierarchical nature of the available models has greatly simplified the procedure of model selection it has limited this analysis to those models which treat transition and transversion substitution categories as axiomatically unequal (excepting the JC model in which all categories are equal). The estimates of parameter values from this analysis suggest that no pre-existing model is a good fit and, although model testing may nevertheless yield a model of best fit, it is important to remember that “best fit” does not ensure “good fit”.

Patterns of substitution within protein coding sequences are far more heavily affected by codon degeneracy and the issue of synonymy than by the differences in the biochemical likelihood of a transition or a transversion. Over time, the fixation of substitutions affected by degeneracy will overwhelm any signal left by differences in the rates of transitions and transversions. Therefore, substitution rate patterns will differ based on the organism, the genetic code and the gene in question. Clearly such custom fitted models are not at this time a reality, and it is recognized that ad hoc creation of such models would not account for the increase in power by the addition (or removal) of any of the given parameters. For that such tests and the LRT and the AIC would be necessary, although this cannot be done with the present state of the art.

Ultimately, in application this means that in cases where a model which disregards transition and transversion categories would be more appropriate, any model selected will either be inherently biased, or in the case of GTR, run the risk of over-parameterization. This study errs on the side of minimizing bias while risking increased statistical margins of error by using the full GTR model and partitioning the data by codon position.

Phylogenetic Reconstruction

MAFFT and MUSCLE alignments of the cupin data set including individual cupin domains from the storage globulin sequences, GLP sequences, and other cupin containing sequences were analyzed using MrBayes. A variety of statistical measures are used to assess the reliability of the MCMC search which is then used to generate a phylogenetic tree. These include assessing the co-linearity of the –log likelihood (-lnL) of trees found in independent MCMC runs, identifying any in the change in –lnL over time between the two runs, and testing for whether the average standard deviation of split frequencies (which is a measure of the distances between the chains of multiple runs) has reached a stable and sufficiently low value.

Tracer 1.41 was used to visually assess whether the –lnL of the two independent runs for each of the MAFFT and MUCLE MCMC searches had reached the same plateau and had remained stable for a the duration of the search. In this way a best guess of when stability (or at least linearity) has been reached, and a burn-in cutoff can be set. It was felt that a burn-in of 1 million generations was sufficient to discount the portion of the MCMC searches in which the –lnL was unstable.

Tracer can also be used to export the trace data in a readable table format. This data was then sampled using a simple random number system and imported into Excel. The trace data (post burn-in) was analyzed to determine a linear regression line of best fit in order to determine whether the –lnL had reached stationarity. This is an important assumption in estimating the posterior probability of the tree space, because a constant drift in the –lnL value would suggest the search had not yet reached a global optima. Analysis found that there was a statistically significant ($p<0.01$) difference in the regression slope from a null hypothesis of zero (Figure 7), and that the runs had not yet reached stationarity.

The question at this point becomes one of whether rerunning the analysis using the last tree from the MCMC search or whether narrowing the range of trees for sampling from the posterior distribution would be the best course of action. Due to time limitations for this study the latter option was more reasonable, however, given a longer initial run-time, or continuing with a new run from the last tree might eventually allow the MCMC search to reach a global peak in the tree space. The caveat with not doing that the search may be sampling from a less than optimal region of the tree space in order to estimate the posterior. On the other hand, in sampling from the last five million generations of the two runs the closest approximation of that posterior distribution can be had.

Although the estimation of the posterior may have been slightly biased, and could have been improved by increased run-time, it was shown that the two runs had at least converged and were sampling from the same region of the tree space. Though the increase in –lnL over time was significant, the slopes between the two runs was not, which suggests that each run was sampling from the same tree space to which they had independently converged. Furthermore, as a measure of convergence among runs the average standard deviation of split frequencies (σ_{SF}) was analyzed to assess whether the difference among chain likelihoods in the runs had become sufficiently small.

A multi-chain MCMC analysis should be run until the σ_{SF} reaches a predetermined minimum value. This was set at 0.01 for the purposes of this study. The MAFFT alignments for both the cupin and globulin domain data sets came close to but did not quite reach that cut-off (Figure 8). The σ_{SF} did however, appear to be reaching a stable plateau, suggesting that given the unique tree-space created by the combination of model parameters, alignment, and sampled tree topologies some runs may not reach this suggested cut-off within a reasonable amount of time, if ever.

This approach was used for the full cupin dataset and the globulin domain dataset aligned with MUSCLE and MAFFT. The alignments of the bicupin globulin coding sequences reached stationarity much faster based on the described measures, and the high posterior probability branch supports on the final tree suggested that including the entire MCMC sample (post burn-in of one million generations) was a good conservative estimate of the posterior distribution.

The SUMT algorithm in MrBayes was used to produce, from those samples of the posterior distribution, phylogenetic trees with branch supports based on the percent agreement among the trees sampled (Figures 9-12). The ALLCOMPAT setting was used to produce fully resolved trees with pP values assigned to each node. The advantage to this over a majority rule consensus is that all node can be displayed in their most well supported arrangement, while giving an indication of how well each portion of the tree is. This allows the viewer to assess where support is poor, marginal or strong.

The phylogenies of cupin domains (Figures 9-10) is color coded to reflect the different groups included in the alignment (N and C-terminus domains of the globulins, cupin domains of the GLPs, spherulins, and various genes of the cupin super-family). In the tree based on the MAFFT alignment (Figure 9) each of these groups is monophyletic. The GLP group forms a single clade with the spherulins basal to it. The N-terminus domains (group "a") of the globulin storage proteins form on large, deeply divided clade, while the C-terminus domain (group "b") forms another. Within each of those clades the 11S and 7S sequences are monophyletic, with the exception of the basal non-storage C-terminus domains, which are basal to the 11S C-terminus domains in the tree based on the MUSCLE alignment (Figure 10). Orange arrows indicate the main incongruence between the two trees within the "b" group. The other cupin-containing sequences from the cupin family form a clade with deeply connected branches. While the pP branch supports are higher for the MAFFT tree than for the MUSCLE tree, neither has strong ($pP < 0.95$) support for the arrangement of the interior nodes of the globulin domains. Division of the GLPs from other groups is also ambiguously supported with a pP of 1.0 and 0.66 on the MAFFT and MUSCLE trees, respectively.

The positions of some sequences within the tree deviate from expectations. The GLP of *Barbicula unguiculata* (bBaUnGo03) groups with the hypothetical GLP from *Ostreococcus lucimarinus* (cOsLuG2h01) within the clade containing the bacterial and fungal oxalate decarboxylases, and phosphomannose isomerases. One unknown protein of *Marchantia polymorpha* from the GenBank EST database groups with a small basal clade of the GLP group along with sequences from *Oryza, Medicago, Physarum, and Barbula*. While taxonomic boundaries are not strictly expected to be adhered to in a multigene tree due to effects of paralogy, gene loss, and duplication, some of these discrepancies are noteworthy, and may merit further exploration. It is difficult to ascertain whether positioning of these genes is based on phylogenetic signal or simply the random nature of the heuristic tree search, since the pP supports within these clades are mixed, ranging from 0.57 to 1.0.

Overall branch support for the phylogeny of the globulin cupin domains (Figure 11) is significantly higher, with the internal relationships among the clades all being well supported (pP > 0.95) in both MAFFT and MUSCLE based trees. The two trees agree in general topology, separating each of the 11S and 7S N-terminus and C-terminus domains into separate clades. Given the agreement between the two trees in this aspect, regardless of the alignment used this topology seems strongly supported. It may be that there is increased phylogenetic signal due to estimating the model parameters specifically for these sequence groups, as opposed to in the larger cupin domain alignment where a single model may be less applicable which may obscure the signal.

Results form the phylogenetic reconstruction based on the MSA of the globulin CDSs also resulted in a well supported tree. Again, the tree is rooted using an outgroup based on results from the full cupin tree; the oxalate oxidase gene from *Bacillus subtilis* was aligned along with other globulin storage protein gene sequences and used as the root. The Bayesian majority rule consensus resulted in a tree with pP > 0.95 for all intenal nodes which separate each respective group; the 11S and 7S globulins each form a monophyly, as well as do small groups of sequences basal to each of those main groups.

These small groups have been predicted in previous studies to be of non-storage function, at least in seeds. As expected these form what appear to be small and ancient clades basal to the main storage protein groups.

Sequences from species across the spermatophytes are found within these main and basal groups. Some general patterns are observed: the grouping of a paraphyletic gymnosperm-type storage protein group in each of the 7S and 11S clades. Dicots and monocots are both represented in each of the main storage proteins caldes as well, and generally separate these two groups further into two sub-groups.

Selection Analysis

Results are shown for both individual chains of the globulin proteins (Figures 13-14) as well as in the complete trimer structure of the proteins (Figures 15-16). TreeSAAP results and output allow the researcher to investigate a seemingly endless number of very specific hypotheses. The results shown in these figures are designed to identify overall patterns of selection and divergence among the storage globulins.

Differential patterns of selection were observed between the two cupin domains in both the 7S and 11S globulins. Differential selection in the 11S globulin chain is apparent in the form of a higher number of sites under either stabilizing or destabilizing selection on the N-terminus domain than on the C-terminus domain (Figure 13). The N-terminal helical region is under complete stabilizing selection, whereas the patterns of non-synonymous substitution at the C-terminal helical region reveal little to no stabilizing selection. The beta-coil of the N-terminus domain shows a mix of stabilizing and destabilizing selection at a much higher level than the C-terminus beta-coil, which is also under a mix of selective pressures.

Differential selection between the two chains of the 7S globulin is also apparent, though not in such surprising amounts as in the 11S globulin. There appears to be more balanced levels of selection between the two chains, with a number of sites in each being under stabilizing or destabilizing selection. The majority of sites under destabilizing selection are on the external surface of the beta coil of either cupin domain. The helix region of the N-terminal domain has several sites under stabilizing selection, while the C-terminal helix has no apparent pattern to the non-synonymous substitutions occurring there. The C-terminus is highly conserved in both 7S and 11S storage proteins.

Patterns of sites under selection in the 11S and 7S groups are shown in figure 15. The face of the legumin trimer which bonds to an opposing trimer to form the mature hexamer structure has a different pattern of selection, with sites under stabilizing and destabilizing selection being located predominantly on the protruding base of the N-terminal beta-barrel. The corresponding face of the 7S trimer has a more centralized pattern, with sites under directional selection more or less forming a ring surrounding sites under stabilizing selection. Three sites (one per subunit) are under destabilizing selection at the edge of the pore. Similarly placed sites on the 11S legumin are also under directional selection.

The external surface of the legumin trimer (that which forms the surface of the legumin hexamer) shares a high degree of similarity with the corresponding surface of the vicilin trimer. Both exhibit a high level of stabilizing selection along similar portions of the protein. Residues that line the outer edge of the central pore are under directional selection.

The patterns of selection in the non-storage 11S/7S-like proteins (figure 16) differ considerably from one another; more so than do the 11S/7S storage proteins. Patterns of selection in the 11S-like group generally mirror that of the 11S storage protein on the internal face of the primer, although with fewer sites under directional selection. The opposite, external side of the trimer exhibits directional selection in a number of sites located at the protein's core, along with the most central residues inside the pore under directional selection. Very little similarity is seen between the 11S and 7S patterns.

The 7S-like group displays an outer ring of sites under selection on the side corresponding to the inner face of the legumin trimer. Several residues surrounding the inner pore are also under selection. the opposite face of the 7S-like protein has a more ambiguous pattern of sites under selection with sites spread around the outer edge and central portion of the trimer. the same sites which comprise the outer edge of the central pore and are under directional selection in the 7S storage group are also under directional selection in the non-storage protein.

Suites of physicochemical properties under selective pressure differed among groups of 11S and 7S seed-storage proteins and their respective non-storage protein groups (Table 3). 28 of 31 properties were found to be under selection among these four main groups. The two storage groups had a more similar profile of selection to one another, as did the two non-storage groups. There were nine properties under the same level of selection between the 11S and 7S storage proteins, out of 20 between them. An additional three properties were in common between the seed storage groups, albeit in different magnitude categories. There were nine properties under the same level of selection between the two non-storage groups, out of 21 between them. Five properties were under similar selection intensity between the storage and non-storage groups (7S and11S-like groups).

It is difficult to asses the statistical significance of these results since the likelihood of a property being found to be under selection is difficult to estimate a priori. We could assume this probability to be 0.5 to maximize the probability-mass function and use the binomial distribution to estimate the likelihood of a given number of co-occurrences among selection categories. Under these assumptions the probability of one co-occurrence out of one trial would be 2 x 0.5^2, or 0.5, as in a coin toss with two simultaneous coins, where either two heads or two tails represents a successful co-occurrence of outcomes.

In the case of multiple selection categories the likelihood that the same property is found in each of three selection categories (none, stabilizing, destabilizing) in two independent analyses becomes $1/3^2$ (one in three possible outcomes occurring simultaneously) assuming there is an equal probability of any outcome. This case becomes more like rolling two three sided dice, which are likely to come up with one of three co-occurrences one-ninth of the time. Since all three co-occurrences will be counted as a success the expectation is that two independent analyses of two different aligned data sets will have physicochemical properties found in the same selection categories about one-third of the time. Significant deviation from this indicates some correlation between the data.

Between the 7S and11S seed storage groups there were 20 property-category matches out of 31 properties analyzed. Between the 7S-like and 11S-like non-seed storage groups there were also 20 property-category matches out of 31 properties analyzed. The likelihood of this many or more matches occurring by chance, under the assumptions outlined above is < 0.0001. These statistics can be used to test the *ad hoc* null hypothesis that under random or unrelated evolutionary pressures the expected mean $H_o = 1/3$ versus the alternative $H_a \neq 1/3$. In both cases we reject the null hypothesis that the similarities between the two seed-storage and the two non-storage groups are due to random chance.

The most similarity between seed storage and non-seed storage groups was found between the 7S storage and the 11S non-storage groups. There were a total of 13 matching property-categories out of 31. the likelihood of this occurring by chance is 0.1152, which is not significant at a reasonable alpha level, and so the null hypothesis cannot be rejected. Their minimal similarity cannot be said to be statistically due to anything but random chance.

DISCUSSION

Multiple Sequence Alignment

Alignment of the cupin domains of the globulin storage proteins has not been previously attempted or reported on a scale similar to that of this study. Furthermore, the alignment of putatively homologous domains from the same gene is a somewhat novel, though not unheard of approach to sequence analysis. In the context of the storage globulins this allows for an in depth analysis of the history of this gene family not previously undertaken. Most phylogenetic studies investigating the storage globulins have used only portions of the coding sequence, a small number of sequences, or have only investigated one or the other of the two main groups. Phylogenetic resolution, presumably due to ambiguity in the alignments has often caused results to be mixed or inconclusive in detailing the evolutionary history of these proteins.

The optimal multiple sequence alignment is able to identify homologous sites and regions among a set of nucleotide or peptide sequences. Due to the highly divergent nature of the globulin storage proteins, as well as their putative relative the GLPs and non-plant cupins, nucleotide alignment yields little resolution. It is also important to consider preservation of the reading frame when aligning protein coding sequences. With these details in mind the alignments were performed on translated AA sequences, using a variety of alignment algorithms.

The results of these alignment strategies revealed that even the most different alignments recovered many of the same conserved sites (Figure 2). This was true both for the alignment of the full bicupin globulin sequences, as well as the alignment of individual cupin domains. In this way the alignments validated one another to a great extent. Sufficient sequence homology exists across both cupin domains of the storage globulins, as well as across the GLP and other cupin domain containing genes to identify sequence homology.

The alignment of the individual cupin domains yielded higher than expected levels of sequence conservation at the amino-acid level. Previous studies which have produced crystallized structures for diverse cupin domain containing proteins have hinted at a common evolutionary origin. The alignment (figure 3) of these sequences reveals that the overall domain architecture is well conserved among all of these groups. Each has a typical cupin motif, with two beta-coils separated by a length-variable spacer. This is followed by an alpha helical region at the C-terminal end. Several sites are highly conserved among the cupin domains of the different taxonomic kingdoms represented. an example is seen in the proline at position 121 in figure 3. This proline is generally present in the N-terminus and C-terminus domains of the storage globulins, the GLPs, the spherulins and the two bacterial oxalate oxidase cupin domains. The glycine at position 82 is also nearly completely conserved among all the sequences analyzed, and is completely conserved in the sequences shown in figure 3. Many other conserved sites, both within and between individual domain groups, are present and indicative of stabilizing synonymous mutation

The level of sequence homology that exists not only between the GLP, globulin, and other cupin domains, but also between the two domains of the globulin proteins, is in line with evidence from other studies which have suggested a possible single cupin origin for the globulin storage proteins. As expected, if there were a single cupin domain gene as a common ancestral link among these genes there would be some level of preserved homology among them. Convergent evolution from unrelated proteins to form similar structures would result in a more random and ambiguous alignment. While sequence similarity does not always prove shared ancestry, at the levels observed among such divergent organisms as bacteria, fungi, slime molds, and vascular plants, and across such a multi functional set of genes, including storage, anti-microbial, enzymatic, and metal-ion binding genes it leaves little doubt that in this case these similarities are due to homology by common ancestry.

The similarities among the globulin storage proteins are more pronounced when aligned as full coding sequences and without ambiguous regions resulting from inclusion of more divergent sequences. Again, the overall domain architecture consisting of anti-parallel beta-coils a helical region separated by a length-variable spacer can be seen (figure 4). Notably the IMS in the 11S sub-family is more pronounced in the N-terminus domain, and more pronounced in the C-terminus domain in the 7S sub-family. This spacer plays an important role in the stabilization of the mature protein in the legumins, since it contains one of four cysteines (position 83, figure 4) that form disulfide bridges to secure the protein chains to one another.

It is of particular interest that such a functionally important and highly conserved region would arise in one lineage and be absent from its nearest relative. This IMS region of the gene is likely to have been due to an insertion event after the divergence of the 11S-like non-storage group from what would become the 11S seed storage proteins, since it is present in the gymnosperm as well as angiosperm sequences, but is notably absent from those shown in previous studies to be basal to the 11S storage proteins. Otherwise, this region may have been present in the ancestral protein, but would have to have been lost in the 7S lineage, as well as in the 11S-like lineage. The C-terminus IMS is present in a much more pronounced state in all of the 7S sequences which seems equally likely to be the result of an insertion in the 7S lineage or a deletion from the 11S lineage.

The one other major difference between the domain architecture of the globulins is the presence of an extended coil region between the second beta-coil and the alpha-helix region in the N-terminus domain. This coil spacer is predominantly hydrophilic and is likely to extend out from the main body of the pro-peptide. Due to its placement it may play a role in maintaining the positional conformation of the helical region, thus facilitating contact and bonding with its corresponding helical region in the formation of the trimer structure. No functional analysis of this specific region has been carried out, and so discussion of why it is present in the 11S storage group is purely speculative. It should be noted that this region is also absent in the non-storage 11S sequence (mOsSa11S4), as in the case of the N-terminus IMS, and so one might infer that it is of some structural or functional import to the activity of legumins as seed storage proteins.

Beyond overall domain architecture the sequence alignment of the globulin proteins reveals patterns of conservation of physicochemical properties. Patterns of hydrophobic and hydrophilic residues emerge when the sequence is color coded using a scale based on the hydrophobicity of each amino acid. Aromatic rings, which are often found at conserved sites in this alignment, as well as the structurally important cysteines are colored green. Using this basic approach to compare patterns among the protein sequences it becomes apparent that even at non-conserved sites property specific substitutions have become fixed in the different orthologs. Although anecdotal, this evidence points to a level of conservation in favor of preserving amino acid properties across a fairly divergent gene family. A more thorough analysis using TreeSAAP reveals a high level of stabilizing selection among these protein sub-families.

Cupin Phylogenetics

In order to conduct an analysis of patterns of natural selection on a multiple alignment it is fist necessary to provide a reliable phylogenetic tree of the inter-relationships of the gene or protein sequences. Furthermore, a phylogenetic tree, or more specifically in this case a gene tree, can directly answer questions about the evolutionary history of related sequences. The main question regarding the evolutionary history of the storage globulins surround the conversion of a hypothetical single cupin ancestor to a bicupin gene, and various hypotheses have been submitted for consideration.

The approach in the study was to take each cupin domain from a wide sample of storage globulins and align them in one single MSA. The resultant tree(s) could then lend support to one hypothesis of domain and gene duplication. It becomes important, then, to first establish the phylogenetic implications of those different hypotheses. Under the simplest model a duplication event occurred, placing two single-cupin genes adjacent to each other in the genome. These two genes evolved into a single reading frame, and subsequently were duplicated again. The two resultant bicupin globulins became the prototypic ancestors of the two main globulin lineages; the 11S legumins and 7S vicilins.

Other models have also been put forth, mainly to explain the apparently rapid divergence of the N-terminus domain of the 11S globulins. In one such model the hypothetical single-cupin gene underwent several duplication events (for a minimum of three cupin domains), leading to a small cluster of cupin genes which were likely transcribed together. This cluster was duplicated, and subsequently different mutational event lead to loss of function of different of these cupin genes between the two clusters. The remaining functional copies of each cluster eventually merged, as in the previous model, to form two different bicupin genes.

In both of these scenarios one or more gene/domain duplications occurs to give rise to the bicupin globulins. The main difference between the two theories is the sequence of events leading to the final state of two separate bicupin gene families. Interestingly in each of these scenarios the 11S and 7S families must be independently recruited and evolve in parallel as seed storage proteins, since the events leading to their creation as separate groups must have occurred prior to the rise of the spermatophytes.

Each of these proposed models can be described as a different phylogenetic arrangement. The first is the simplest and would simply appear as a symmetric bifurcating tree where the 11S and 7S N-terminus domains formed one monophyletic group and the 11S and 7S C-terminus domains formed another. These would be centrally rooted with other single cupin sequences such as GLPs, spherulins and other non-plant cupins which are all hypothetical candidate models of the ancestral single cupin gene. In this way the most basal division between the N-terminus and C-terminus domains would be representative of the initial domain duplication leading to a bicupin ancestor. Subsequent divergence of the 11S and 7S domains would represent the gene duplication and divergence of the ancestral legumin and vicilin sequences.

Deviations from this model would result in the tree deviating from a symmetrical bifurcating phylogram. Entire clades might swap places, changing the interpretation of the order of event, with gene duplication followed by independent domain duplication. In this case 11S domains would group together, and the 7S domains would group together, since the divergence of these two would have occurred before the divergence of their respective domains.

In the hypothetical model with a minimum of three proposed cupin domains the phylogenetic tree would have to be asymmetrical. Either the rooted out-group would separate the domain groups into clades of one and three domains, or the out-group could remain centrally placed, and one domain could be found to be nested within another. A number of other deviations could be imagined, each with slightly differing interpretations. Interestingly, the phylogenies produced using the MAFFT and MUSCLE alignments of the cupin domains were inconclusive in selecting one of the above scenarios as the most likely. This was due do the placement of the non-storage 7S C-terminus domains as either basal to the 11S C-terminus domain (figure 9), making the 7S C-terminus domain a paraphyletic group, or as basal to the rest of the 7S C-terminus domains, making the 7S C-terminus domains monophyletic. This basic incongruence yields different interpretations on the sequence of events leading to the bicupin storage globulins.

In neither case was the topology well supported, and the fact that the two trees disagreed in this regard indicated that the data and models used were not accurate enough to resolve these deep relationships. In order to correct this problem the alignment of the globulin cupin domains was attempted, adding only a few sequences from the monophyletic GLPs, spherulins and prokaryotic cupins. The placement of these groups was generally agreed on between the two different trees, so using these as an out-group for rooting the tree could be done with confidence.

It is thought that the initial inclusion of so many of these other sequences may have affected the estimation of model parameters to the extent that additional error prevented the likelihood and MCMC searches from being able to adequately distinguish between the two general solutions. In using alignments with predominately globulin genes it is thought that model parameter estimations more accurately reflect the patterns of substitution among the globulin genes, thus lending additional power to resolve the internal topology of their evolutionary history.

In the resultant trees (both MAFFT and MUSCLE alignments were used) the symmetrical divergence of N/C-terminus domains followed by the divergence of duplicated bicupin genes is strongly supported (figure 11). Given the agreement between the topologies of the two trees, in addition to the strong support values for those internal branches, the simpler explanation of domain duplication followed by gene duplication seems the most likely.

Not only does this analysis help solidify a picture of the evolutionary history of these genes, but it also confirms the practicality of aligning the entire sequences in order to obtain better resolution for estimating the gene family tree. Previous uncertainty as to the strict homology of the 11S and 7S domains as presently arranged in their genomic sequences has lead some to focus on alignments of only a short region in the conserved C-terminal domain. This is a less than optimal approach which is no longer a necessary precaution, since evidence points to the fact that the respective cupin domains of the globulin proteins are in fact strictly homologous.

Alignment and phylogenetic reconstruction of the globulin gene family can be facilitated by the inclusion of the bicupin oxalate oxidase of *Bacillus subtilis*, which the previous phylogenetic analysis showed to be an out-group to the storage globulins. Several interesting observations can be made based on this multi-gene tree. Firstly, the division between 11S and 7S sub-families becomes more obvious, and is well supported. the non-storage 11S/7S-like sequences are placed as basal groups of their respective storage globulin clades. This supports a previous analysis which placed non-storage groups as early offshoots of the main globulins (Borroto and Dure 1987). The inclusion of a fern spore-specific globulin in the 7S-like non-storage clade, as done previously, suggests these groups broke off prior to the 11S and 7S globulins being recruited as seed storage proteins. This further suggests that the 11S and 7S storage globulins acquired their storage functionality independently. This is thought to have happened at some point during or after the divergence of the non-storage groups, since the next most closely related extant sequences are from the gymnospermata, which are known storage proteins, and bear all the sequence hallmarks of such.

The gymnosperm sequences form a paraphyletic group basal to the angiosperm storage proteins in both 11S and 7S clades. Nested within the angoisperm clades are gene sequences from both both monocot and dicot species, and these are generally divided into their own groups. From this one can infer that both vicilin and legumin storage proteins are present in almost all, if not all genera of seed bearing plants, and that they are the original storage reservoir that was developed at the time that land plants began to evolve seed bearing capabilities.

Any inference in terms of the topology within any of the aforementioned clades beyond this is somewhat unreliable; not due to poor support values, since these are generally high, but because taxon sampling becomes increasingly important as topologically based inferences become more nuanced. Availability of complete storage globulin sequences does not give a balanced and deep enough sampling of the plant kingdom to adequately depict the evolutionary tree of the entire gene family. Not only taxonomic sampling, but also genomic sampling has an impact on the gene tree topology. As can be seen in the globulin gene tree several species of plants have multiple copies of one or the other or both of the main storge proteins. These gene copies are often not recent duplications, and are even located in distant clades. As more sequenced genomes become available it will be important to scan these for potential cupin-containing genes to add greater detail and depth of understanding regarding the patterns of gene duplication that have occurred since the early divergence of the seed storage globulins.

Selection Analysis

Although we can observe patterns of divergence within a gene tree by using phylogenetic tools, this alone does not tell us how genes or proteins have differentiated over time. Presumably, in a case such as this, the divergence of 11S and 7S storage proteins and their subsequent retention in all major lineages of seed bearing plants was due to selective pressures that lead to fitness advantages for species that had both, at least in the early stages of the radiation of spermatophytes. The question remains; if 11S and 7S seed storage gene are paraphyletic, and their common ancestor was not a seed storage protein, what selection pressures led to their acquiring storage capacity independent of each other? Using a property based selection analysis tool such as TreeSAAP, the different groups within the gene family can be analyzed for patterns of selection and then compared amongst one another to identify example of convergent, divergent, stabilizing and destabilizing selection. These patterns can shed light on the evolutionary pressures that gave rise to these two classes of storage proteins.

In comparing patterns of selection among related gene families one can identify which physicochemical properties are under selection in each of those groups and identify commonalities and differences that might help explain observable patterns in biological processes. The TreeSAAP results (table 3) indicate which properties are under stabilizing, destabilizing or no selection. These profiles were compared and tested for significant deviation from the null hypothesis that coincidences of properties under selection were due to chance. There were significant co-occurrences of properties under selection between the two storage groups and between the two non-storage groups. Interestingly, since each of these groups (11S/7S storage versus 11S/7S-like non-storage) is paraphyletic the similarity of their selection profiles is likely due to convergent evolution under similar selection pressures.

It can be assumed that a certain proportion of properties will be found under similar selection pressures due to random chance, however, the disparity between the two functional classes, and the similarities within them seem to go beyond random chance. This provides evidence that after the main 11S and 7S sub-families diverged, certain lineages were recruited for storage functions and were thus submitted to similar evolutionary pressures. Under these circumstances we would expect the same suite of properties to be under selection in the two clades, despite being paraphyletic. This is in essence a situation of convergent or parallel evolution among the two groups of storage proteins.

Although similar selection pressures may have resulted in convergent evolution in the 7S and 11S lineages thus explaining their simultaneous adaptation of seed storage functionality, selective pressures were apparently not identical on the two duplicated cupin domains. The results from the selection analysis at the level of the bicupin subunit reveal that once the cupin domain was duplicated each came under different pressures and responded accordingly by adapting in different ways (figures 13-14). The 11S and 7S trimers are composed of three bicupin subunits, while their GLP relatives are composed of six identical single cupin subunits.

We would expect that since any mutation in a GLP gene is essentially repeated in each of the six subunits there would be a great deal more selective pressure on that one cupin domain. On the other hand, the duplication and subsequent asymmetrical adaptation of the globulin cupin domain would likely allow for more flexibility in the evolution and adaptation of those proteins. Evidence of this is seen in the differential patterns of selection between the cupin domains in both the 11S and 7S lineages. This seems to be a novel adaptation which may have played a role in the diversification of the storage proteins.

Although little is still known about whether most of these globulins have secondary functions beyond their role as storage proteins, differential selection and asymmetrical adaptation between the domains may have led to a wider variety of functional adaptation than previously thought. Several storage globulin genes have been identified as having substrate binding properties, including sugar (Kummer and Rüdiger 1988), membrane and antigen-binding, and more diverse functionality (Dunwell et al. 2004) is likely to be discovered as experimental studies seek to gather empirical evidence for the functional nuances of this gene family.

While more directed and gene specific studies would be best in determining the correlation of selection pressures with functional properties of these proteins, an overview analysis such as was conducted in this study can identify likely regions for functional and structural adaptation based on patterns of sites under directional and stabilizing selection. The results of this analysis can be divided into three general conclusions regarding the evolutionary pressures.

First, 11S and 7S storage globulins diverged functionally mainly in the region of the bicupin subunit which composes the internal face of the legumin hexamer while stabilizing selection preserved structural and functional features of the opposite surface (figure 15). This does not come as a surprise since the region of the legumin trimer where the most destabilizing selection occurs is that which is responsible for covalent bonding to a second, identical trimer. Since a mutation in one region of this surface must be countered by another mutation elsewhere on this surface it is reasonable to expect a higher level of adaptive selection occurring at these residues. Interestingly, whereas the 11S globulins appear to be selected on to preserve the self-interactive nature of this region of the protein, the 7S globulins also have undergone intense directional selection in this region. Although they are not known to form hexameric structures at any time, it is possible that this surface of the trimer is responsible for some other interaction with surrounding proteins, substrates, or cellular structures.

Second, as the 11S and 7S globulins diverged from one another, so too did 11S- and 7S-like non-storage proteins diverge from the newly formed legumin and vicilin clades, not only phylogenetically, but in clear distinctions in patterns of selection (Figure 16). These likely gave rise to new functional adaptations in what are relatively uncharacterized globulin sub-families. This is a little explored, though important chapter in the evolution of the globulin storage proteins. Since these two sub-families are, in theory, placed just before the evolution of seed storage the divergence and functions of extant genes is of utmost importance in determining just how the ancestral globulins functioned. This would shed more light on how and when legumins and vicilins acquired storage capabilities. Since the predominate patterns among the non-synonymous substitutions of the non-storage sequences revealed destabilizing selection very different from their storage cousins we assume that they diverged in function as they split from the main legumin and vicilin groups. This, coupled with the patterns of stabilizing selection on one surface of the legumin and vicilin proteins, suggests that the original function of their common ancestor may have been some type of vegetative storage. At the very least legumin and vicilin storage proteins retained a great deal more of their ancestral characteristics than did the off-shoot non-storage clades.

Third, although patterns of adaptive selection differ among the different sub-families, one commonality is the presence of destabilizing selection around the central pore of the trimer. The sites surrounding the edge of this pore, as well as those forming its center are under radical destabilizing selection in each of the four main groups (Figures 15-16).   This suggests that this region is critical to the acquisition and diversification of function within the globulin family. A number of physicochemical properties are under both stabilizing and destabilizing selection in this regions, and a closer look at these sites reveals how adjacent residues under selection may interact to form a functional region at the center of the globulin trimer (Figure 17). This realization may serve to better understand those secondary functions of the storage globulins which have been noted by various studies.

CONCLUSION

The cupin super-family is an excellent example of the ancient origin of functionally divergent, though structurally conserved gene families. By aligning individual cupin domains across bicupin and single cupin gene families the mechanisms of domain and gene duplication which gave rise to the seed storage globulins becomes evident. They provide a textbook example of *de novo* addition of domains, followed by structural and functional diversification. A preliminary gene tree of the storage globulins reveals two independent lineages evolving parallel to each other as storage proteins.

The legumin and vicilin gene families have undergone diversifying selection that has lead to a wide array of structural and functional adaptations, allowing for a greater diversity among seed bearing plants. Although the extent of functional diversity among extant globulins remains unclear, evidence suggests that there is more than previously suspected. Evidence suggests that a critical aspect of functional divergence among storage globulins is centered on the pore created by the joining of three bicupin subunits, forming the basic trimer structure of both legumins and vicilins.

Additional work is needed to ascertain to what degree this region plays a role in seed storage. Possibilities include tissue desiccation, metal ion binding, anti-microbial activity, etc. Empirical study of the function of these genes in a wide variety of species is lacking, although much has been discovered in economically important crop species. Future study from the molecular evolution stand point might take several different approaches. Studies of allelic variations at the species or population level might reveal more finely detailed patterns of selection, and comparisons between these genes in closely related species might reveal how they function differently in different species. Studies looking exclusively at patterns of gene duplication at the genome level might be well suited to identifying how recently duplicated genes might be diverging from one another based on differences in selection patterns among duplicate copies. Lastly, as increased sequence data becomes available for a wider taxonomic variety of plants, studies at the gene family level will be able to achieve better resolution and more accurate reconstructions of gene duplication events that lead to the diverse group of proteins found within the globulin storage family. These types of studies combined with increased empirical data on secondary protein functions will allow for more reliable prediction of gene function in non-model species, will create a richer model of how seed bearing plants acquired this evolved trait, and how domain and gene duplication leads to structural and functional divergence through natural selection.

REFERENCES

Adachi M, Kanamori J, Masudo T, Yagasaki K, Kitamura K, Mikami B, Utsumi S (2003) Crystal structure of soybean 11S globulin: Glycinin A3B4 homohexamer. PNAS 100:7395-7400

Adachi M, Takenaka Y, Gidamis AB, Mikami B, Utsumi S (2001) Crystal structure of soybean proglycinin. J Mol Biol 305:291-305

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402

Balzotti MRB, Thornton JN, Maughan PJ, McClemman DA, Stevens MR, Jellen EN, Fairbanks DJ, Coleman CE (2008) Expression and evolutionary relationships of the *Chenopodium quinoa* 11s seed storage protein gene. Int J Plant Sci 169:281-291

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242

Bernier F, Berna A (2001) Germins and germin-like proteins: Plant do-all proteins. But what do they do exactly? Plant Physiol Biochem (Paris) 39:545-554

Beyer K, Grishina G, Bardina L, Grishin A, Sampson HA (2002) Identification of 11S globulin as a major hazelnut food allergen in hazelnut-induced systemic reactions. J Allergy Clin Immunol 110:517-523

Bharali S, Chrungoo NK (2003) Amino acid sequience of the 26 kDa subunit of legumin-type seed storage protein of common buckwheat (*Fagopyrum esculentum* Moench): molecular characterization and phylogenentic analysis. Phytochemistry (Oxf) 63:1-5

Bland J, Altman D (1995) Multiple significance tests: the Bonferroni method. British Medical Journal 310:170

Borroto K, Dure L (1987) The globulin seed storage proteins of flowering plants are derived from two ancestral genes. Plant Mol Biol 8:113-131

Caliskan M (2000) Germin, an oxalate oxidase, has a function in many aspects of plant life. Turk J Biol 24:717-724

Carter C, Graham RA, Thornburg RW (1998) *Arabidopsis thaliana* contains a large family of germin-like proteins: characterization of cDNA and genomic sequences encoding 12 unique family members. Plant Mol Biol 38:929-943

Casey R (1999) Distribution and some properties of seed globulins. In: Shewry PR, Casey R (eds) Seed Proteins. Kluwer Academic Publishers, Dordrecht, Netherlands, pp 159-169

Cho T-J, Nielsen NC (1989) The glycinin Cy3 gene from soybean. Nucleic Acids Res 17:4388

Domoney C, Casey R (1985) Measurement of gene number for seed storage proteins in *Pisum*. Nucleic Acids Res 13:687-699

Dunwell JM, Culham A, Carter CE, Sosa-Aguirre CR, Goodenough PW (2001) Evolution of functional diversity in the cupin superfamily. Trends Biochem Sci 26:740-746

Dunwell JM, Khuri S, Gane PJ (2000) Microbial relatives of the seed storage proteins of higher plants: conservation of structure and diversification of function during evolution of the cupin superfamily. Microbiol Mol Biol Rev 64:153-179

Dunwell JM, Purvis A, Khuri S (2004) Cupins: the most functionally diverse protein superfamily? Phytochemistry (Oxf) 65:7-17

Eddy S (1998) Profile Hidden Markov Models. Bioinformatics 14:755-763

Fares MA, *et al.* (2002) A sliding-window method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol 55:509-521

Farris JS, Kallersjo M, Kluge AG, Bult C (1995) Testing significance of incongruence. Cladistics 20:315-319

Fischer H, Haake V, Horstmann C, Jensen U (1995) Characterization and evolutionary relationships of *Magnolia* legumin-encoding cDNAs representing two divergent gene subfamilies. Eur J Biochem 229:645-650

Galau GA, Wang HYC, Hughes WD (1991) Sequence of the *Gossypium hirsutum* D-genome alloallele of *Legumin A* and its mRNA. Plant Physiol (Rockv) 97:1268-1270

Gatehouse JA, Brown D, Gilroy J, Levasseur M, Castleton J, Ellis THN (1988) Two genes encoding 'minor' legumin polypeptides in pea (*Pisum sativum* L.). Biochem J 250:15-24

Gibrat J, Madej T, Bryant S (1996) Surprising similarities in structure comparison. Current Opinions in Structural Biology 6:377-385

Grzelczak Z, Lane B (1984) Signal resistance of a soluble protein to enzymic proteolysis, An unorthodox approach to the isolation and purification of germin, a rare growth-related protein. Canadian Journal of Biochemistry and Cell Biology 62:1351-1353

Hager K-P, Wind C (1997) Two ways of legumin-precursor processing in conifers: Characterization and evolutionary relationships of *Metasequoia* cDNAs representing two divergetn legumin gene subfamilies. Eur J Biochem 246:763-771

Hayashi M, Mori H, Nishimura M, Akazawa T, Hara-Nishimura I (1988) Nucleotide sequence of cloned cDNA coding for pumpkin 11S globulin β submit. Eur J Biochem 172:627-632

Hirano H, Fukazawa C, Harada L (1985) The primary structures of the A4 and A5 subunits are highly homologous to that of the A3 subunit is the glycinin seed storage protein of soybean. FEBS 181:124-128

Huelsenbeck JP, Ronquist F (2001) MrBayes: Bayesian inference of phylogeny, version 3.1.2. Bioinformatics 17:754-755

Huelsenbeck JP, Ronquist F (2005) Mr Bayes v 3.1.2, Bayesian analysis of phylogeny. University of California

Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A 88:958-962

Katoh K, Toh M (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment Nucleic Acids Res 33:511-518

Khuri S, Bakker FT, Dunwell JM (2001) Phylogeny, function, and evolution of the cupins, a structurally conserved, functionally diverse superfamily of proteins. Mol Biol Evol 18:593-605

Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-1222

Kummer H, Rüdiger H (1988) Characterization of a lectin-binding storage protein from pea (*Pisum sativum*). Biological chemistry Hoppe-Seyler 369:639-646

Lang D, Eisinger J, Reski R, Rensing S (2005) Representation and high-quality annotation of the *Physcomitrella patens* transcriptome demonstrates a high proportion of proteins involved in metabolism in mosses. Plant Biol 7:238-250

Lycett GW, Croy RRD, Shirsat AH, Boulter D (1984) The complete nucleotide sequence of a legumin gene from pea. Nucleic Acids Res 12:4493-4506

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30:281-283

Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genet 169:1753-1762

Mathieu M, Neutelings G, Hawkins S, Grenier E, David H (2003) Cloning of a pine germin-like protein (GLP) gene promoter and analysis of its activity in transgenic tobacco Bright Yellow 2 cells. Physiol Plant 117:425-434

McClellan DA, McCracken KG (2001) Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. Mol Biol Evol 18:917-925

McClellan DA, Palfreyman EJ, Smith MJ, Moss JL, Christensen RG, Sailsbery JK (2005) Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. Mol Biol Evol 22:437-455

Mediana-Godoy S, Nielsen NC, Paredes-Lopez O (2004) Expression and characterization of a his-tagged 11S Seed Globulin from *Amaranthus hypochondriacus* in *Escherichia coli*. Biotechnol Prog 30:1749-1756

Membre N, Berna A, Neutelings G, David A, David H, Staiger D, Vasquez JS, Raynal M, Delseny M, Bernier F (1997) cDNA sequence, genomic organization and differential expression of three *Arabidopsis* genes for germin/oxalate oxidase-like proteins. Plant Mol Biol 35:459-469

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, Marshall WF, Qu L-H, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Chlamydomonas Annotation Team, JGI Annotation Team, Grigoriev IV, Rokhsar DS, Grossman AR (2007) The *Chlamydomonas* Genome reveals the evolution of key animal and plant functions. Science 318:245-251

Muntz K, Blattner FR, Shutov AD (2002) Legumains - a family of asparagine-specific cysteine endopeptidases involved in propolypeptide processing and protein breakdown in plants. J Plant Physiol 159:1281-1293

Nakata M, Watanabe Y, Sakurai Y, Hashimoto Y, Matsuzaki M, Takahashi Y, Satoh T (2004) *Germin-like protein* gene family of moss, *Physcomitrella patens*, phylogenetically falls into two characteristic new clades. Plant Mol Biol 56:381-395

Nei M (2005) Selectionism and neutralism in molecular evolution. Mol Biol Evol 22:2318-2342

Nielsen R, Huelsenbeck JP (2002) Detecting positively selected amino acid sites using posterior predictive p-values. Pac Symp Biocomput:576-588

Okita TW, S HY, Hnilo J, Kim WT, Aryan AP, Larson R, Krishnan HB (1989) Structure and expression of the rice glutelin multigene family. J Biol Chem 264:12573-11781

Pei J, Kim B-H, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295-2300

Posada D, Crandall K (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818

Rambaut A, Drummond AJ (2006) Tracer. Evolutionary Biology Group, University of Oxford, Oxford

Robert EC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797

Rodin J, Ericson ML, Josefsson L-G, Rask L (1990) Characterization of a cDNA clone encoding a *Brassica napus* 12S protein (Cruciferin) subunit: Disulfide bonding between subunits. J Biol Chem 5:2720-2723

Ryan AJ, Royal CL, Hutchinson J, Shaw CH (1989) Genomic sequence of a 12S seed storage protein from oilseed rape (*Brassica napus* c.v. jet neuf). Nucleic Acids Res 17:3584

Samardzic JT, Milisavljevic MD, Brkljacic JM, Konstantinovic MM, Maksimovic VR (2004) Characterization and evolutionary relationship of methionine-rich legumin-like protein from buckwheat. Plant Physiol Biochem (Paris) 42:157-163

Scheffler K, Seoighe C (2005) A bayesian model comparison approach to inferring positive selection. Mol Biol Evol 22:2531-2540

Shutov A, Kakhovskaya I, Braun H, Baumlein H, Muntz K (1995) Legumin-like and vicilin-like seed storage proteins: evidence for a common single-domain ancestral gene. J Mol Evol 41:1057-1069

Shutov AD, Baumlein H (1999) Origin and Evolution of Seed Storage Globulins. In: P.R. S, Casey R (eds) Seed Proteins. Kluwer Academic Publishers, Dordrecht, Netherlands, pp 543-561

Shutov AD, Baumlein H, Blattner FR, Muntz K (2003) Storage and mobilization as antagonistic functional constraints on seed storage globulin evolution. J Exp Bot 54:1645-1654

Shutov AD, Braun H, Chesnokov YV, Baumlein H (1998a) A gene encoding a vicilin-like protein is specifically expressed in fern spores: Evolutionary pathway of seed storage globulins. Eur J Biochem 252:79-89

Shutov AD, Braun H, Chesnokov YV, Horstmann C, Kakhovskaya IA, Baumlein H (1998b) Sequence peculiarity of gnetalean legumin-like seed storage proteins. J Mol Evol 47:486-492

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690

Swofford D (1999) PAUP—phylogenetic analysis using parsimony (and other methods), 4.0 beta version. Sinauer Associates, Inc., Sunderland, Mass.

Tai SSK, Wu LSH, Chen ECF, Tzen JTC (1999) Molecular cloning of 11S globulin and 2S albumin, the two major seed storage proteins in sesame. J Agric Food Chem 47:4932-4938

Tajima F (1991) Determination of window size for analysing DNA sequences. J Mol Evol 33:470-473

Takaiwa F, Ebinuma H, Kikuchi S, Oono K (1987) Nucleotide sequence of a rice glutelin gene. FEBS 221:43-47

Takaiwa F, Kikuchi S, Oono K (1986) The structure of rice storage protein glutelin precursor deduced from cDNA. FEBS 206:33-35

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. . Nucleic Acids Research 24:4876-4882

Wayne ML, Simonsen KL (1998) Statistical tests of neutrality in the age of weak selection. Trends Ecol Evol 13:236-240

Weng W-M, Gao X-S, Zhuang N-L, Xu M-L, Xue Z-T (1995) The glycinin A3B4 mRNA from wild soybean *Glycine soja* Seib. et ZUCC. Plant Physiol (Rockv) 107:665-666

Wind C, Hager K-P (1996) Legumin encoding sequences from the Redwood family (Taxodiaceae) reveal precursors lacking the conserved Asn-Gly processing site. FEBS Lett 383:46-50

Woolley SJ, Johnson J, Smith MJ, Crandall KA, McClellan DA (2003) TreeSAAP: Selection on amino acid properties using phylogenetic trees. Bioinformatics (Oxf) 19:671-672

Xia X, Li WH (1998) What amino acid properties affect protein evolution? J Mol Evol 47:557-564

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573

Yang Z, Bielawski JP (2000) Statistical methods for detecting moledular adaptation. Trends Ecol Evol 15:496-503

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at indiviual sites along specific lineages. Mol Biol Evol 19:908-917

Zimmermann G, Baumlein H, Mock H-P, Himmerlback A, Schweizer P (2006) The multigene family encoding germin-like proteins of barley: regulation and function in basal host resistance. Plant Physiol (Rockv) 142:181-192

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Integrative Biology. University of Texas, Austin

TABLES AND FIGURES

Table 1. Sample of results from HMMER. Species names correspond to putative vicilin genes identified using BLAST search. Raw score and E-value are reported. Bolded taxa are examples of sequences in the initial data set that were dropped post HMM analysis.

Species name	Score	E-value
Anacardium occidentale	187.5	1.30E-55
Arabidopsis thaliana	214.4	1.00E-63
Araucaria angustifolia	88.8	6.60E-26
Canavalia ensiformis	226.5	2.30E-67
Corylus avellana	234.2	1.10E-69
Elaeis guineensis	262	4.60E-78
Fagopyrum esculentum	-4.1	4.80E-06
Glycine max	193.2	2.40E-57
Gossypium hirsutum	401.2	6.00E-120
Guazuma ulmifolia	-58.4	0.33
Juglans nigra	274.9	6.40E-82
Lens culinaris	248.1	7.20E-74
Lycopersicon esculentum	-32.8	0.0017
Matteuccia struthiopteris	231.9	5.30E-69
Phaseolus vulgaris	128.2	8.60E-38
Picea glauca	268.8	4.30E-80
Pisum sativum	234.9	6.70E-70
Sesamum indicum	249.6	2.50E-74
Triticum aestivum	230.9	1.10E-68
Vicia narbonensis	254	1.20E-75
Zamia furfuracea	211.4	8.20E-63
Zea mays	254.8	6.90E-76

Table 2. Taxon code, accession number, species name, protein type, and function for each sequence used in this study. First letter designations of taxon code are as follows: b – bryophyte, c – chlorophyte, d – dicot, f – filicophyte, g – gymnosperm, i – fungi, m – monocot, n – marchantiophyte, p – prokaryote, y – myxomycete.

Taxon Code	Accession #	Species	Protein type	Function
bBaUnGo01	AB036797	*Barbula unguiculata*	Germin-like	mn superoxide dismutase
bBaUnGo03	AB028460	*Barbula unguiculata*	Germin-like	mn superoxide dismutase
bPyPaGu02	AB185322	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu04	AB185323	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu05	AB185324	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu06	AB185492	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu07	AB185325	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu1a	AB177347	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu1b	AB177646	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu3a	AB177349	*Physcomitrella patens*	Germin-like	unknown
bPyPaGu3b	AB177645	*Physcomitrella patens*	Germin-like	unknown
cOsLuG2h01	ABO94770	*Ostreococcus lucimarinus*	Germin-like	unknown
dAmHy11S	X82121	*Amaranthus hypochondriacus*	Legumin	seed storage
dAnOc11S	AF453947	*Anacardium occidentale*	Legumin	seed storage
dAnOc7S	AF395893	*Anacardium occidentale*	Vicilin	seed storage
dArTh11S	DQ056550	*Arabidopsis thaliana*	Legumin	seed storage
dArTh12S1	BT009682	*Arabidopsis thaliana*	Legumin	seed storage
dArTh12S2	AY117228	*Arabidopsis thaliana*	Legumin	seed storage
dArTh7S1	BT008623	*Arabidopsis thaliana*	Vicilin	seed storage
dArTh7S2	AY090307	*Arabidopsis thaliana*	Vicilin	seed storage
dArThG1m01	X91921	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG1m02	BT024837	*Arabidopsis thaliana*	Germin-like	manganese binding

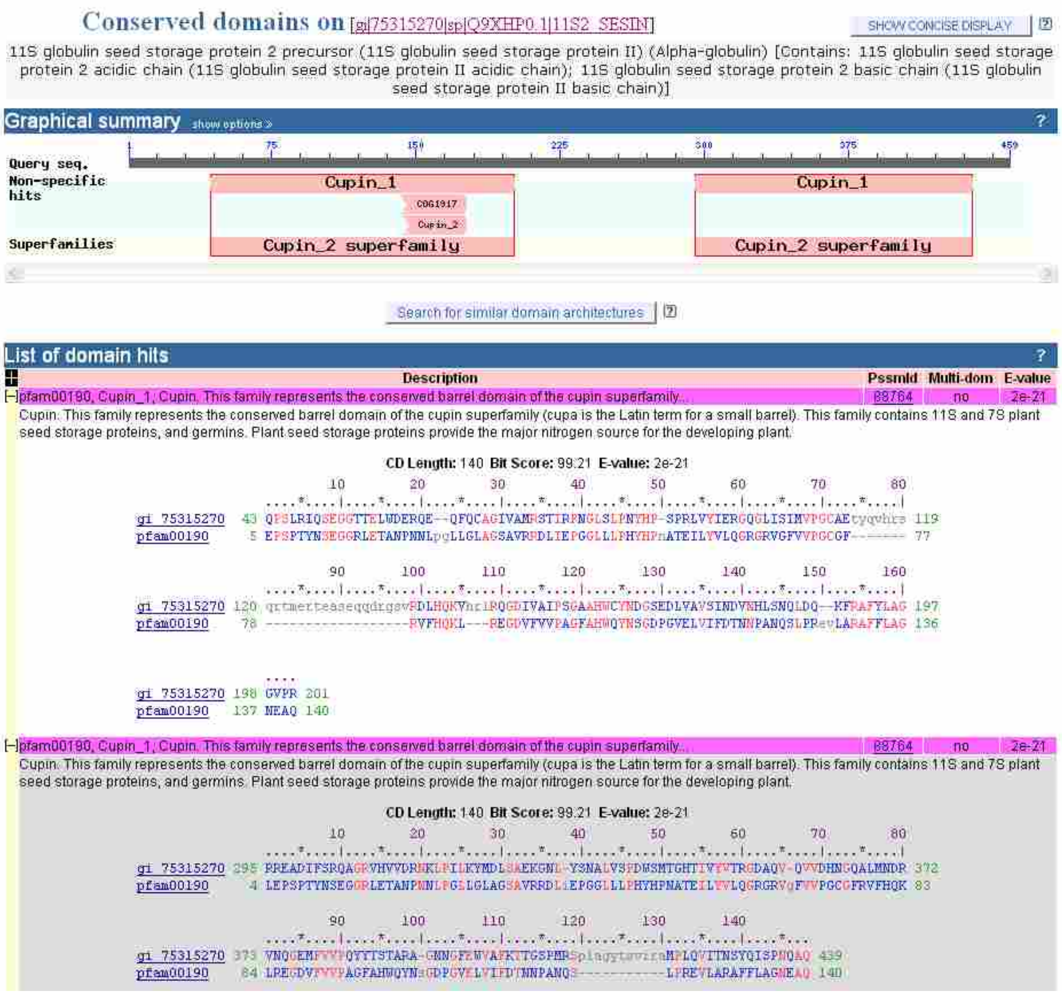
dArThG1m04	BT029449	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG1m05	DQ446430	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG3m01	BT030312	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG3m02	BT028993	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG3m03	AK229153	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG5m01	U75194	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG5m02	U75203	*Arabidopsis thaliana*	Germin-like	manganese binding
dArThG5m11	DQ447102	*Arabidopsis thaliana*	Germin-like	manganese binding
dBeEx11S	AY221641	*Bertholletia excelsa*	Legumin	seed storage
dBeVuGo01	AF310016	*Beta vulgaris*	Germin-like	oxalate oxidase
dBrNa11Sa	M16860	*Brassica napus*	Legumin	seed storage
dBrNaGo01	U21743	*Brassica napus*	Germin-like	oxalate oxidase
dCaEn7S	X59467	*Canavalia ensiformis*	Vicilin	seed storage
dChQu11S	AY562549	*Chenopodium quinoa*	Legumin	seed storage
dCiAr11S	Y15527	*Cicer arietinum*	Legumin	seed storage
dCoAr11S	AF054895	*Coffea arabica*	Legumin	seed storage
dCoAv11S	AF449424	*Corylus avellana*	Legumin	seed storage
dCoAv7S	AF441864	*Corylus avellana*	Vicilin	seed storage
dCuKu7S	D29803	*Cucurbita maxima*	Vicilin	seed storage
dCuMa7S	AB019195	*Cucurbita maxima*	Vicilin	seed storage/anti-microbial
dCuPe11S	M36407	*Cucurbita pepo*	Legumin	seed storage
dDaCa7S	U47078	*Daucus carota*	Vicilin	seed storage
dGoHi11S	M16905	*Gossypium hirsutum*	Legumin	seed storage
dGoHi7S1	M16936	*Gossypium hirsutum*	Vicilin	seed storage
dGoHi7S2	M16891	*Gossypium hirsutum*	Vicilin	seed storage
dGoHiGEe01	AI728954	*Gossypium hirsutum*	Germin-like	unknown
dGoHiGu02	AF116537	*Gossypium hirsutum*	Germin-like	auxin binding
dGoKiGu01	AY116171	*Gossypium hirsutum*	Germin-like	unkown
dHeAn11S	M28832	*Helianthus annuus*	Legumin	seed storage
dIpNiGu01	D45425	*Ipomoea nil*	Germin-Like	unknown
dJuNi7S	AY102931	*Juglans nigra*	Vicilin	seed storage
dJuRe11S	AY692446	*Juglans regia*	Legumin	seed storage
dLeCu7S	AJ551424	*Lens culinaris*	Vicilin	seed storage
dLiUsGu01	AF310960	*Linum usitatissimum*	Gemin-like	unknown
dLuAl11S	AJ938034	*Lupinus albus*	Legumin	seed storage
dMaIn7S	AF161884	*Macadamia integrifolia*	Vicilin	seed storage/anti-microbial
dMeCrGu01	M93041	*Mesembryanthemum crystallinum*	Germin-like	unknown
dMeTr7S	AC148289	*Medicago truncatula*	Vicilin	seed storage
dMeTrGy02	AY184807	*Medicago truncatula*	Germin-like	Mycorrhiza response
dNiAtGu01	AY436749	*Nicotiana attenuata*	Germin-like	oxalate oxidase
dNiLaGo01	AF411917	*Nicotiana langsdorffii*	Germin-like	superoxide dismutase
dPeFr11S	AF180392	*Perilla frutescens*	Legumin	seed storage
dPhVu7S	X03004	*Phaseolus vulgaris*	Vicilin	seed storage
dPiSa11S1	AJ132614	*Pisum sativum*	Legumin	seed storage
dPiSa11S2	X67424	*Pisum sativum*	Legumin	seed storage
dPiSa7S1	X67429	*Pisum sativum*	Vicilin	seed storage
dPiSa7S2	AJ276875	*Pisum sativum*	Vicilin	seed storage
dPiSaGo01	AJ250832	*Pisum sativum*	Germin-like	oxalate oxidase
dPiSaGo02	AJ250834	*Pisum sativum*	Germin-like	oxalate oxidase
dPiSaGu03	AJ311624	*Pisum sativum*	Germin-like	oxalate oxidase

dPlVuGu01	AJ276491	*Phaseolus vulgaris*	Germin-like	unknown
dPoTrG1u01	CU226469	*Populus tremula*	Germin-like	unknown
dPrAm11S	X78119	*Prunus dulcis*	Legumin	seed storage
dPrPeGa01	PPU79114	*Prunus persica*	Germin-like	auxin-binding
dQuRo11S	X99539	*Quercus robur*	Legumin	seed storage
dRaSa11S	X59808	*Raphanus sativus*	Legumin	seed storage
dSeIn11S1	AF091842	*Sesamum indicum*	Legumin	seed storage
dSeIn11S2	AF240004	*Sesamum indicum*	Legumin	seed storage
dSeIn7S	AF240006	*Sesamum indicum*	Vicilin	seed storage
dSiAl11S	AY846388	*Sinapis alba*	Legumin	seed storage
dSoLy7S	AM932874	*Solanum lycopersicum*	Vicilin	seed storage
dThCa7S	X62625	*Theobroma cacao*	Vicilin	seed storage
dViFa11S	X55014	*Vicia faba*	Legumin	seed storage
dViFa7S	Y00462	*Vicia faba*	Vicilin	seed storage
dViNa11S	Z46803	*Vicia narbonensis*	Legumin	seed storage globulin
dViNa7S	Z71987	*Vicia narbonensis*	Vicilin	seed storage
dViSa11S1	Z32835	*Vicia sativa*	Legumin	seed storage
dViSa11S2	Z32796	*Vicia sativa*	Legumin	seed storage
dVtVi7S	AM463475	*Vitis vinifera*	Vicilin	seed storage
dVtViGu01	EF064171	*Vitis vinifera*	Germin-like	unknown
dVtViGu02	DQ673106	*Vitis vinifera*	Germin-like	unknown
dVtViGu06	EF064174	*Vitis vinifera*	Germin-like	unknown
fMaSt7S	Z54364	*Matteuccia struthiopteris*	Vicilin	seed storage
gArAn7S	AF513725	*Araucaria angustifolia*	Vicilin	seed storage
gCaDe11S	X95540	*Calocedrus decurrens*	Legumin	seed storage
gCrJa11S1	X95542	*Cryptomeria japonica*	Legumin	seed storage
gCrJa11S2	X95543	*Cryptomeria japonica*	Legumin	seed storage
gEpGe11S	Z50777	*Ephedra gerardiana*	Legumin	seed storage
gGiBi11S	Z50778	*Ginkgo biloba*	Legumin	seed storage
gGnGn11S	Z50779	*Gnetum gnemon*	Legumin	seed storage
gMeGl11S	X95544	*Matatteuchia glyptostroboides*	Legumin	seed storage
gPiGl11S	X63192	*Picea glauca*	Legumin	seed storage
gPiGl7S	X63191	*Picea glauca*	Vicilin	seed storage
gPiSt11S1	Z11486	*Pinus strobus*	Legumin	seed storage
gPnRaGu01	AF049065	*Pinus radiata*	Germin-like	unknown
gPnSyGu01	AY077705	*Pinus sylvestris*	Germin-like	unknown
gPsMe11S	L07484	*Pseudotsuga menziesii*	Legumin	seed storage
gWeMi11S	Z50780	*Welwischia mirabilis*	Legumin	seed storage
gZaFu7S	Z50791	*Zamia furfuracea*	Vicilin	seed storage
iAsFuS3d01	XM_743615	*Aspergillus fusarium*	Spherulin-like	dessication
iAsNiG1h01	XM_658449	*Aspergillus nidulans*	Germin-like	unknown
iBoFuGu01	XM_001549865	*Botryotinia fuckeliana*	Germin-like	unknown
iFlVeCc1	AY238332	*Flammulina sp.*	Cupin	oxalate decarboxylase
iPhNoGh01	XM_001805933	*Phaeosphaeria nodorum*	Germin-like	unknown
mDiCa11S	X95510	*Dioscorea caucasica*	Legumin	seed storage
mElGu11S	AF261691	*Elaeis guineensis*	Legumin	seed storage
mElGu7S	AF250228	*Elaeis guineensis*	Vicilin	seed storage
mGlMa11S1a	AB195712	*Glycine max*	Legumin	seed storage
mGlMa11S2	D00216	*Glycine max*	Legumin	seed storage
mGlMa7S	AY234869	*Glycine max*	Vicilin	seed storage/sucrose binding

mHoVuGn10	DQ647620	*Hordeum vulgare*	Germin-like	nucleotide pyrophosphatase
mHoVuGo08	AF250936	*Hordeum vulgare*	Germin-like	oxalate oxidase
mHoVuGo13	DQ647619	*Hordeum vulgare*	Germin-like	oxalate oxidase
mHoVuGo16	DQ647622	*Hordeum vulgare*	Germin-like	oxalate oxidase
mHoVuGo18	DQ647624	*Hordeum vulgare*	Germin-like	oxalate oxidase
mLoPeGo01	AJ291825	*Lolium perenne*	Germin-like	oxalate oxidase
mMaSa11S1	X82464	*Magnolia salicifolia*	Legumin	seed storage
mMaSa11S2	X82465	*Magnolia salicifolia*	Legumin	seed storage
mMuAcGu01	AF417204	*Musa acuminata*	Germin-like	unknown
mOrSaC2h03	NM_001053410	*Oryza sativa*	Germin-like	unknown
mOrSaG12u1	NM_001072723	*Oryza sativa*	Germin-like	unknown
mOrSaG1o03	NM_001049124	*Oryza sativa*	Germin-like	oxalate oxidase
mOrSaG2h01	NM_001053409	*Oryza sativa*	Germin-like	unknown
mOrSaG2h02	NM_001053411	*Oryza sativa*	Germin-like	unknown
mOrSaG2o04	NM_001058155	*Oryza sativa*	Germin-like	oxalate oxidase
mOrSaG3o01	NM_001057500	*Oryza sativa*	Germin-like	oxalate oxidase
mOrSaG4h01	NM_001060423	*Oryza sativa*	Germin-like	unknown
mOrSaG8h02	NM_001067690	*Oryza sativa*	Germin-like	unknown
mOrSaG8o01	NM_001067698	*Oryza sativa*	Germin-like	oxalate oxidase
mOrSaG8o08	NM_001067693	*Oryza sativa*	Germin-like	oxalate oxidase
mOrSaG8o12	NM_001068511	*Oryza sativa*	Germin-like	oxalate oxidase
mOrSaG9u01	NM_001070506	*Oryza sativa*	Germin-like	unknown
mOsSa11S1	NM_001053047	*Oryza sativa*	Legumin	seed storage
mOsSa11S2	NM_001053304	*Oryza sativa*	Legumin	seed storage
mOsSa11S3	NM_001067442	*Oryza sativa*	Legumin	seed storage
mOsSa11S4	NM_001061004	*Oryza sativa*	Legumin	seed storage
mOsSa11S5	NM_001070411	*Oryza sativa*	Legumin	seed storage
mOsSa7S1	NM_001055806	*Oryza sativa*	Vicilin	seed storage
mOsSa7S2	NM_001058068	*Oryza sativa*	Vicilin	seed storage
mSaSa11S	Y09116	*Sagittaria sagittifolia*	Legumin	seed storage
mTrAe11S	EU482412	*Triticum aestivum*	Legumin	seed storage
mTrAe7S	M81719	*Triticum aestivum*	Vicilin	seed storage
mTrAeGu01	M21962	*Triticum aestivum*	Germin-like	unknown
mTrAeGu06	AJ237943	*Triticum aestivum*	Germin-like	unknown
mZeMa11S	NM_001111395	*Zea mays*	Legumin	seed storage
mZeMa7S	X59083	*Zea mays*	Vicilin	seed storage
mZeMaGn01	AY394010	*Zea Mays*	Germin-like	nucleotide pyrophosphatase
nMaPoGE01	C95673	*Marchantia polymorpha*	Germin-like	unknown
pAqAeCp01	AE000657	*Aquifex aeolicus*	Cupin	phosphomannose isomerase
pArFuCp01	AE000782	*Archaeoglobus fulgidus*	Cupin	phosphomannose isomerase
pBaSuCc1	Z99120	*Bacillus subtilis*	Cupin	oxalate decarboxylase
pPsAeCp01	M14037	*Pseudomonas aeruginosa*	Cupin	phosphomannose isomerase
pRaSpCg1	AF036940	*Ralstonia sp.*	Cupin	gentisate 1,2-dioxygenase
pRhSpCp01	U00090	*Rhizobium sp.*	Cupin	phosphomannose isomerase
pSpSpCg1	AJ224977	*Sphingomonas sp.*	Cupin	gentisate 1,2-dioxygenase
pSrMeCh01	CP000739	*Sinorhizobium medicae*	Cupin	unknown
pSySpCc1	BAA17550	*Synechocystis sp.*	Cupin	oxalate decarboxlase
pXaCaCp01	AM920689	*Xanthomonas campestris*	Cupin	phosphomannose isomerase
yPsPoSd01	M18428	*Physarum polycephalum*	Spherulin	dessication
yPsPoSd02	M18429	*Physarum polycephalum*	Spherulin	dessication

Protein Type	Physicochemical property																											
11S	**AHT**	*ANSR*	*BLK*	*BRD*	*CHI*	**COMP**		*CT*	**EQC**		*HYD*					**PCT**	*PMAH*	*PNT*	*Pol*				*RMSFD*		*SH*	*TnbE*	*TT*	*TTH*
7S	**AHT**	*ANSR*	**BLK**		*CHI*				**EQC**			**IP**			**MVOL**		**PMAH**	**PNT**	*Pol*				*RMSFD*		SH		*TT*	*TTH*
11S-like	**AHT**		**BLK**			**COMP**		**CT**	**EQC**	**HCA**		**IP**	**LnbE**	**MnbE**	**MVOL**					**PSV**	**PR**	**RefI**	**RMSFD**	**SARR**			**TT**	
7S-like	**AHT**						**COPR**		**EQC**	**HCA**		**IP**		**MnbE**	**MVOL**	**PCT**		**PNT**		**PSV**	**PR**	**RefI**		**SARR**		**TnbE**		**TTH**

Table 3. Physicochemical properties under radical or minimal selection for each of the four main protein groups in this study. Properties in bold letters indicate those undergoing significant directional selection, while italicized properties are under significant stabilizing selection. Results are based on independent analysis of selection patterns in different clades of the globulin gene tree (Figure 12). 31 total properties were analyzed for significant deviations from expectations under neutral conditions. Abbreviations for physicochemical properties are as follows: alpha-helical tendencies (AHT), average number of surrounding residues (ANSR), bulkiness (BLK), buriedness (BRD), chromatographic index (CHI), coil tendencies (CT), composition (COMP), compressibility (COPR), equilibrium constant (ionization of COOH) (EQC), helical contact area (HCA), hydropathy (HYD), isoelectric point (IP), long-range non-bonded energy (LnbE), Mean r.m.s. fluctuation displacement (RMSFD), molecular volume (MVOL), partial specific volume (PSV), polar requirement (PR), polarity (Pol), power to be at the C-terminal (PCT), power to be at the middle of alpha-helix (PMAH), power to be at the N-terminal (PNT), short and medium range non-bonded energy (MnbE), solvent accessible reduction ratio (SARR), surrounding hydrophobicity (SH), total non-bonded energy (TnbE), turn tendencies (TT).



Conserved domains on [gi|75315270|sp|Q9XHP0.1|11S2_SESIN] SHOW CONCISE DISPLAY

11S globulin seed storage protein 2 precursor (11S globulin seed storage protein II) (Alpha-globulin) [Contains: 11S globulin seed storage protein 2 acidic chain (11S globulin seed storage protein II acidic chain); 11S globulin seed storage protein 2 basic chain (11S globulin seed storage protein II basic chain)]

Graphical summary show options »

Search for similar domain architectures

List of domain hits

Description	PssmId	Multi-dom	E-value
pfam00190, Cupin_1, Cupin. This family represents the conserved barrel domain of the cupin superfamily...	88764	no	2e-21

Cupin. This family represents the conserved barrel domain of the cupin superfamily (cupa is the Latin term for a small barrel). This family contains 11S and 7S plant seed storage proteins, and germins. Plant seed storage proteins provide the major nitrogen source for the developing plant.

CD Length: 140 Bit Score: 99.21 E-value: 2e-21

```
                       10        20        30        40        50        60        70        80
                ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 75315270  43 QPSLRIQSEGGTTELWDERQE--QFQCAGIVAMRSTIRPNGLSLPNYHP-SPRLVYIERGQGLISIMVPGCAEtyqvhrs 119
pfam00190     5 EPSPTYNSEGGRLETANPNNLpgLLGLAGSAVRRDLIEPGGLLLPHYHPnATEILYVLQGRGRVGFVVPGCGF------- 77

                       90       100       110       120       130       140       150       160
                ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 75315270 120 qrtmerteaseqqdrgsvPDLHQKVhrlRQGDIVAIPSGAAHWCYNDGSEDLVAVSINDVNHLSNQLDQ--KFRAFYLAG 197
pfam00190    78 ------------------RVFHQKL---REGDVFVVPAGFAHWQYNSGDPGVELVIFDTNNPANQSLPRevLARAFFLAG 136

                ....
gi 75315270 198 GVPR 201
pfam00190   137 NEAQ 140
```

Description	PssmId	Multi-dom	E-value
pfam00190, Cupin_1, Cupin. This family represents the conserved barrel domain of the cupin superfamily...	88764	no	2e-21

Cupin. This family represents the conserved barrel domain of the cupin superfamily (cupa is the Latin term for a small barrel). This family contains 11S and 7S plant seed storage proteins, and germins. Plant seed storage proteins provide the major nitrogen source for the developing plant.

CD Length: 140 Bit Score: 99.21 E-value: 2e-21

```
                       10        20        30        40        50        60        70        80
                ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 75315270 295 RREADIFSRQAGRVHVVDRNKLPILKYMDLSAEKGNL-YSNALVSPDWSMTGHTIVYVTRGDAQV-QVVDHNGQALMNDR 372
pfam00190     4 LEPSPTYNSEGGRLETANPNNLPGLLGLAGSAVRRDLIEPGGLLLPHYHPNATEILYVLQGRGRVGFVVPGCGFRVFHQK 83

                       90       100       110       120       130       140
                ....*....|....*....|....*....|....*....|....*....|....*....|....*...
gi 75315270 373 VNQGEMFVVPQYYTSTARA-GNNGFEWVAFKTTGSPMRsplagytsviraMPLQVITNSYQISPNQAQ 439
pfam00190    84 LPEGDVFVVPAGFAHWQYNsGDPGVELVIFDTNNPANQS-----------LPREVLARAFFLAGNEAQ 140
```

Figure 1. Typical results page from a conserved domain search against the CDD at NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd). Results are shown for accession Q9XHP0, an 11S storage globulin of Sesamum indicum. PFam and COG domain hits with corresponding E-values indicate strength of match. Note the typical bicupin structure and the large inter-domain region, common in storage globulins, which shows no conserved domain.

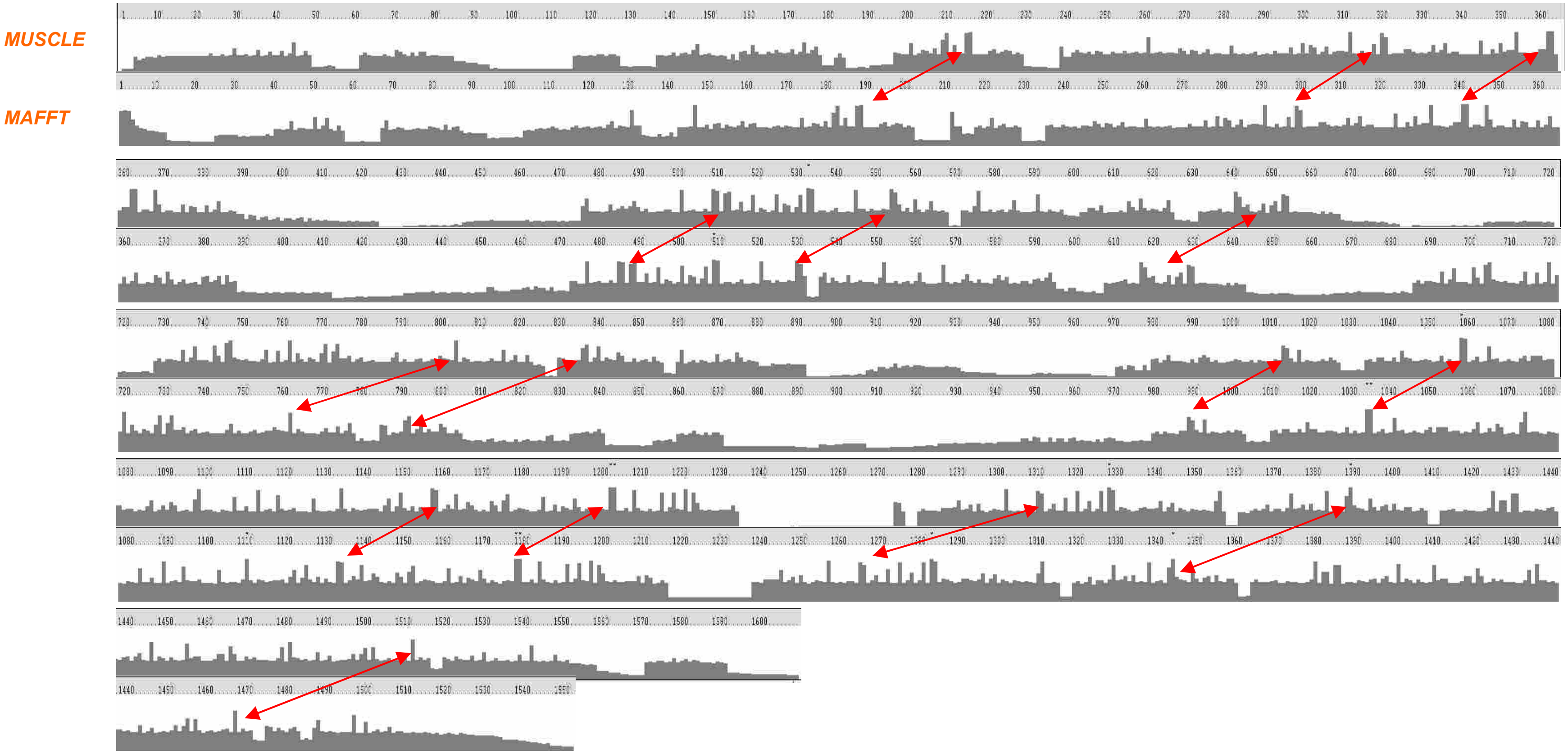


Figure 2. Q-score profiles for MAFFT-linsi and MUSCLE alignments of bicupin nucleotide sequences generated using ClustalX 1.83. Sites with Q-scores < 10 were removed from the alignment. Red arrows indicate examples of highly conserved sites identified in both alignments. Similar observations of commonly aligned blocks were seen in the cupin domain alignment, and across alignment algorithms. Low scoring regions corresponded to large indels between major groups. The lowest (<10) were removed to maximize efficiency of phylogenetic analysis for resolving internal nodes.

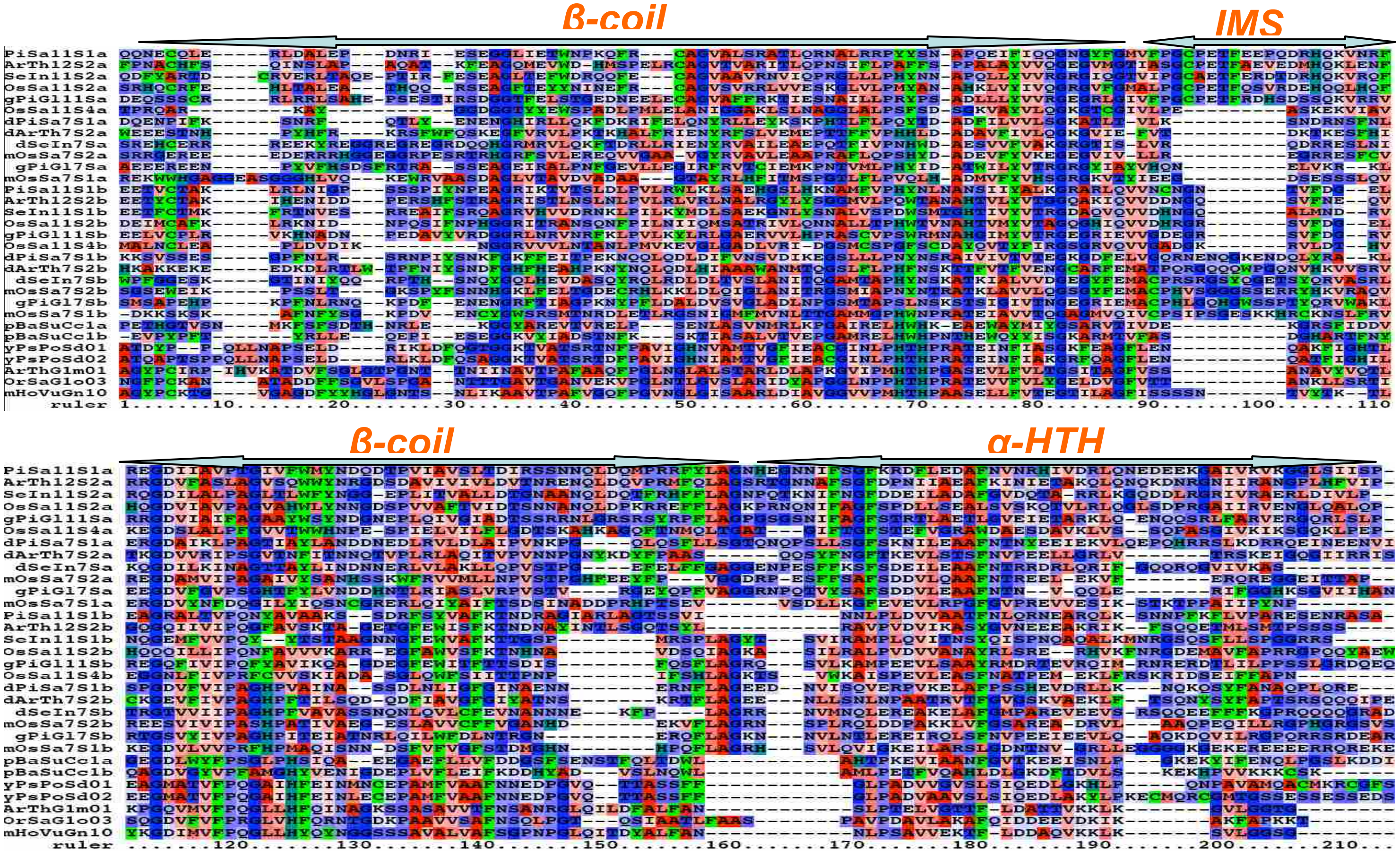


Figure 3. Selected sequences from the MUSCLE alignment of the cupin domain dataset. Coloration of amino acids based on hydrophobicity (red is hydrophobic, blue hydrophilic), with aromatic rings and cysteines colored green. Annotations above sequences show regions in the alignment corresponding to structural features of the protein including the anti-parallel beta-coils, the inter-motif spacer (IMS), and alpha-helical region. Conserved sites within and between groups of cupin domains are clearly visible (e.g. glycine at 34 and 82, F-L-A-G at 159, aromatic at 182, leucine/isoleucine at 192, etc.).

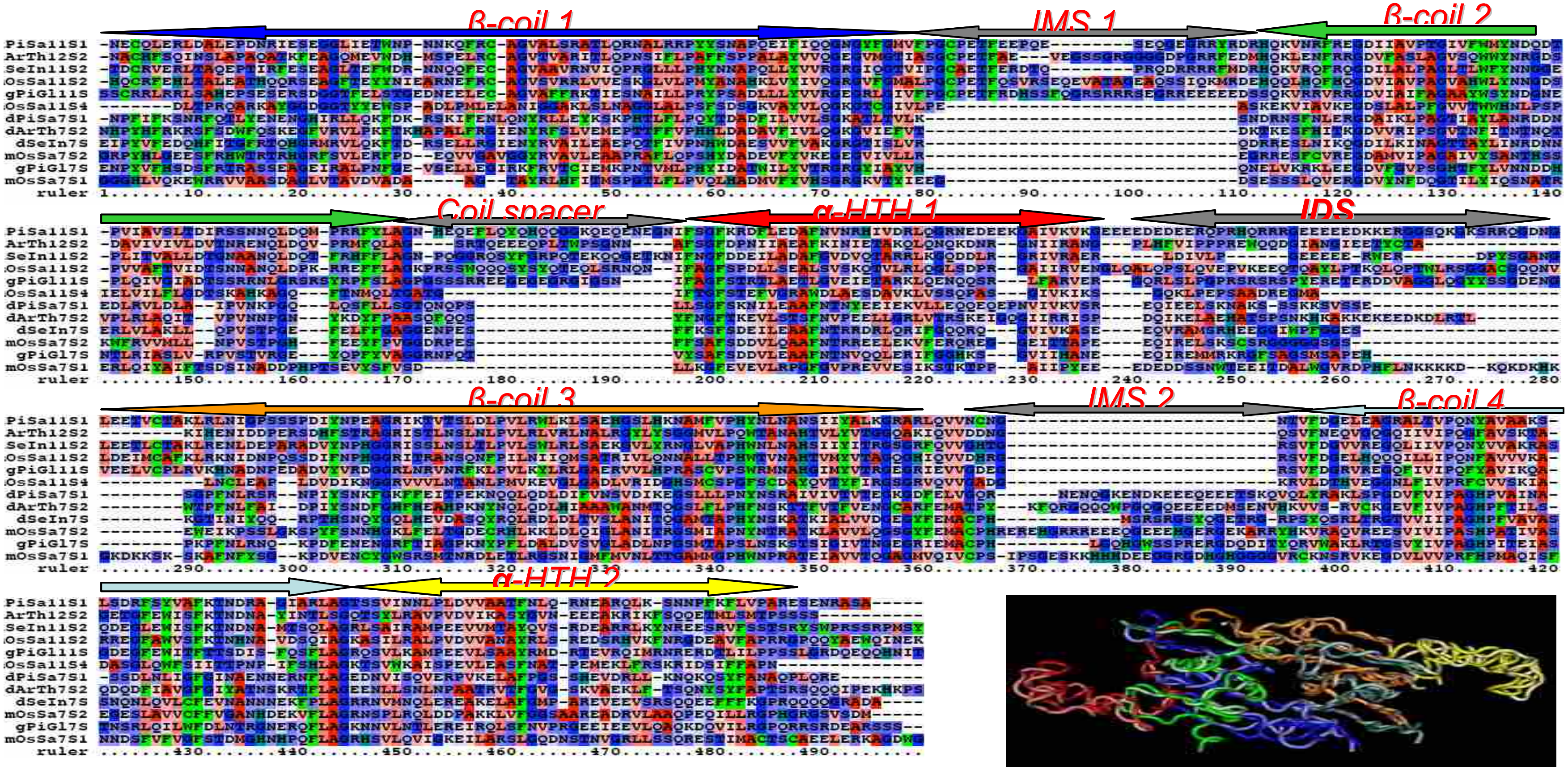


Figure 4. Selected sequences from the MUSCLE alignment of 11S and 7S nucleotide sequences. Coloration of amino acids based on hydrophobicity (red is hydrophobic, blue is hydrophilic), with aromatic rings and cysteines colored green. Annotations above sequences show regions in the alignment corresponding to structural features of the protein, including regions as mentioned in Figure 3, as well as the inter-domain spacer (IDS). Three-dimensional alignment of structural data, inset (IMS and IDS regions are not shown). Colors correspond to the annotations on the sequence alignment. Rich-colored ribbon corresponds to the *Glycine max* legumin proglycinin (1fxz). Light-colored ribbon corresponds to the vicilins of *Phaseolus vulgaris*, phaseolin.

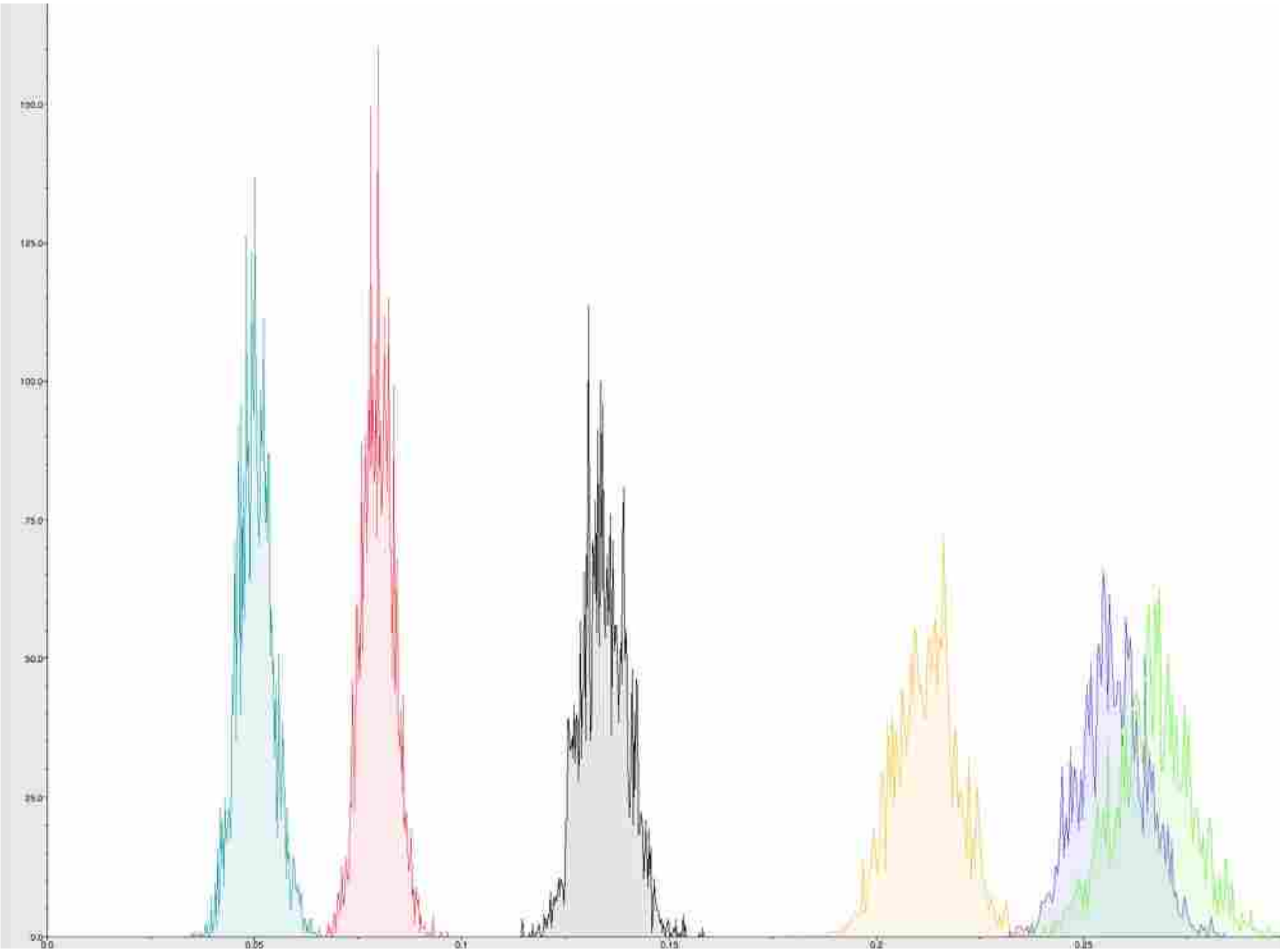


Figure 5. Tracer output shows the estimated posterior probability distribution for each of six substitution rate categories. Data are from the MUSCLE alignment MCMC run in MrBayes. Parameters are estimated for the second codon position. Five total rate categories are likely, since the *b* and *d* rate categories are not significantly different; suggested ad hoc model: *a*, *b* = *d*, *c*, *e*, *f*.

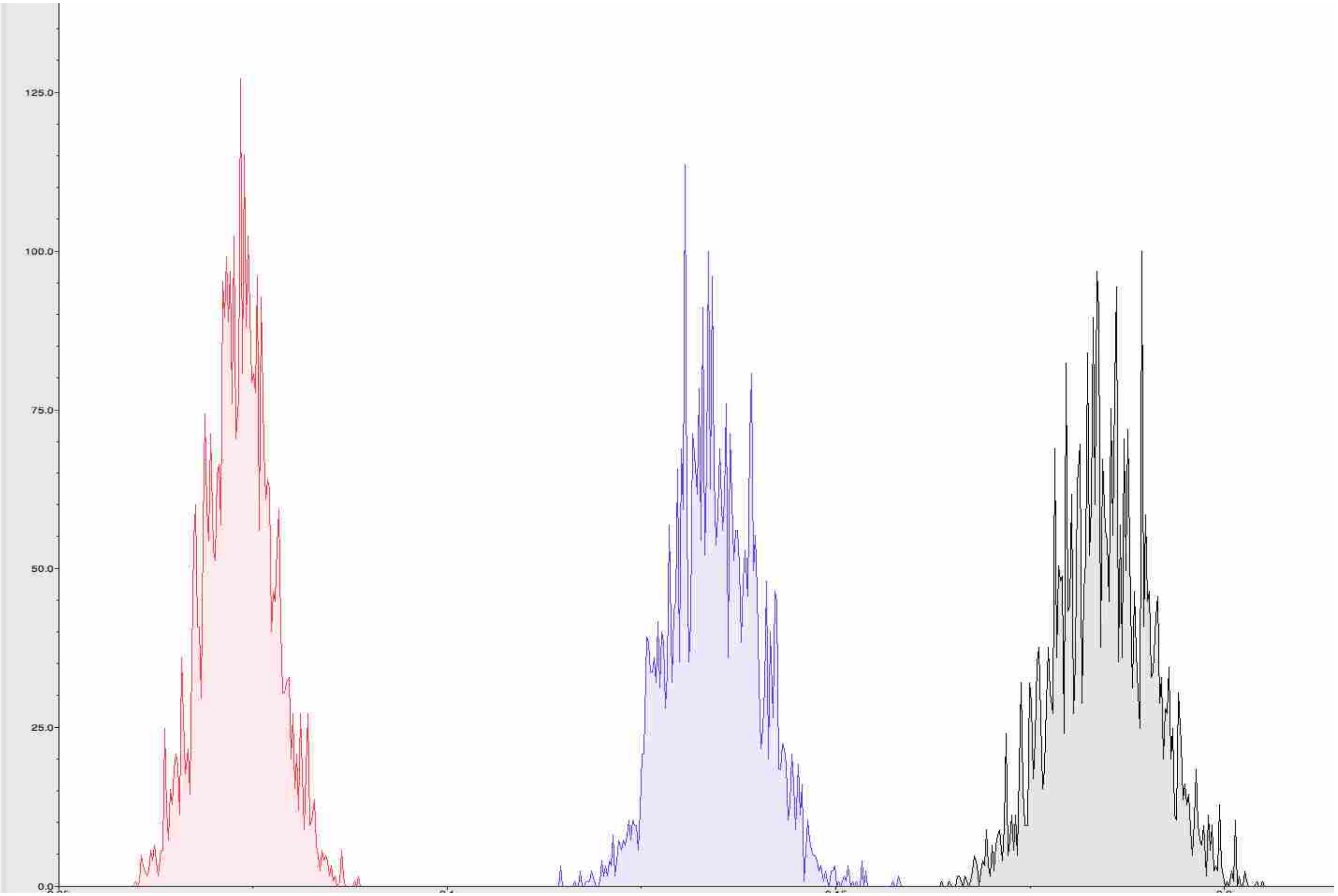


Figure 6. Tracer output shows the estimated posterior probability distributions from substitution category *a* (A<>C) for each of three codon positions. Estimates are generated over successive runs during the MCMC search in MrBayes using the MUSCLE alignment of the globulin cupin domains. Estimated mean values for the model parameter *a* can easily bee seen to be significantly different among the three codon positions.

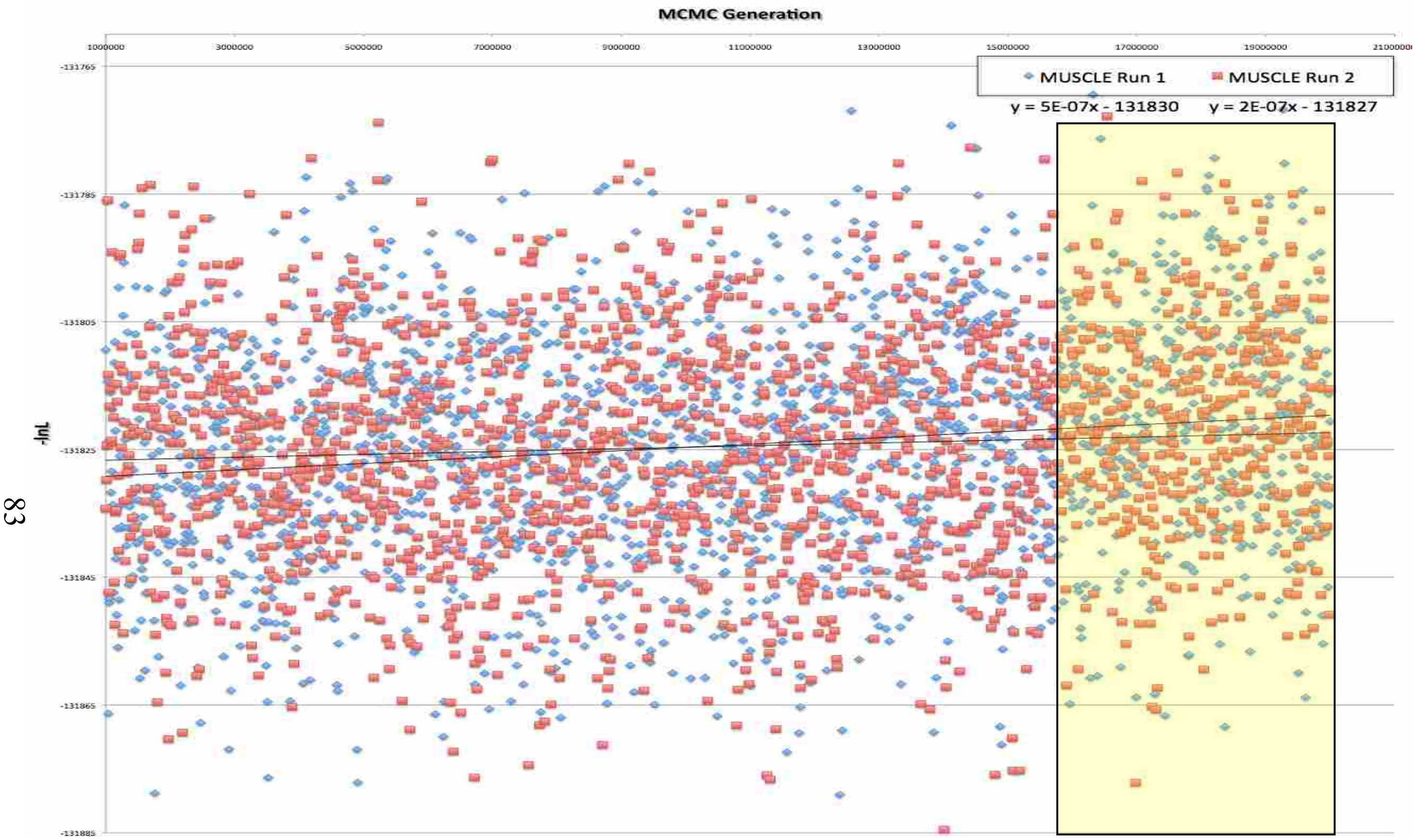


Figure 7. Linear regression of the trace data for the MCMC search performed in MrBayes on the MUSCLE alignment of globulin cupin domains. Blue and red tick marks represent run 1 and run 2 of the MCMC search, respectively. The change in –lnL as the MCMC generation increased is slight, but statistically significant. The yellow box highlights the region from 15 to 20 million generations which was sampled to get a best estimate of the posterior probability distribution. A similar approach was used in estimating the posterior distribution

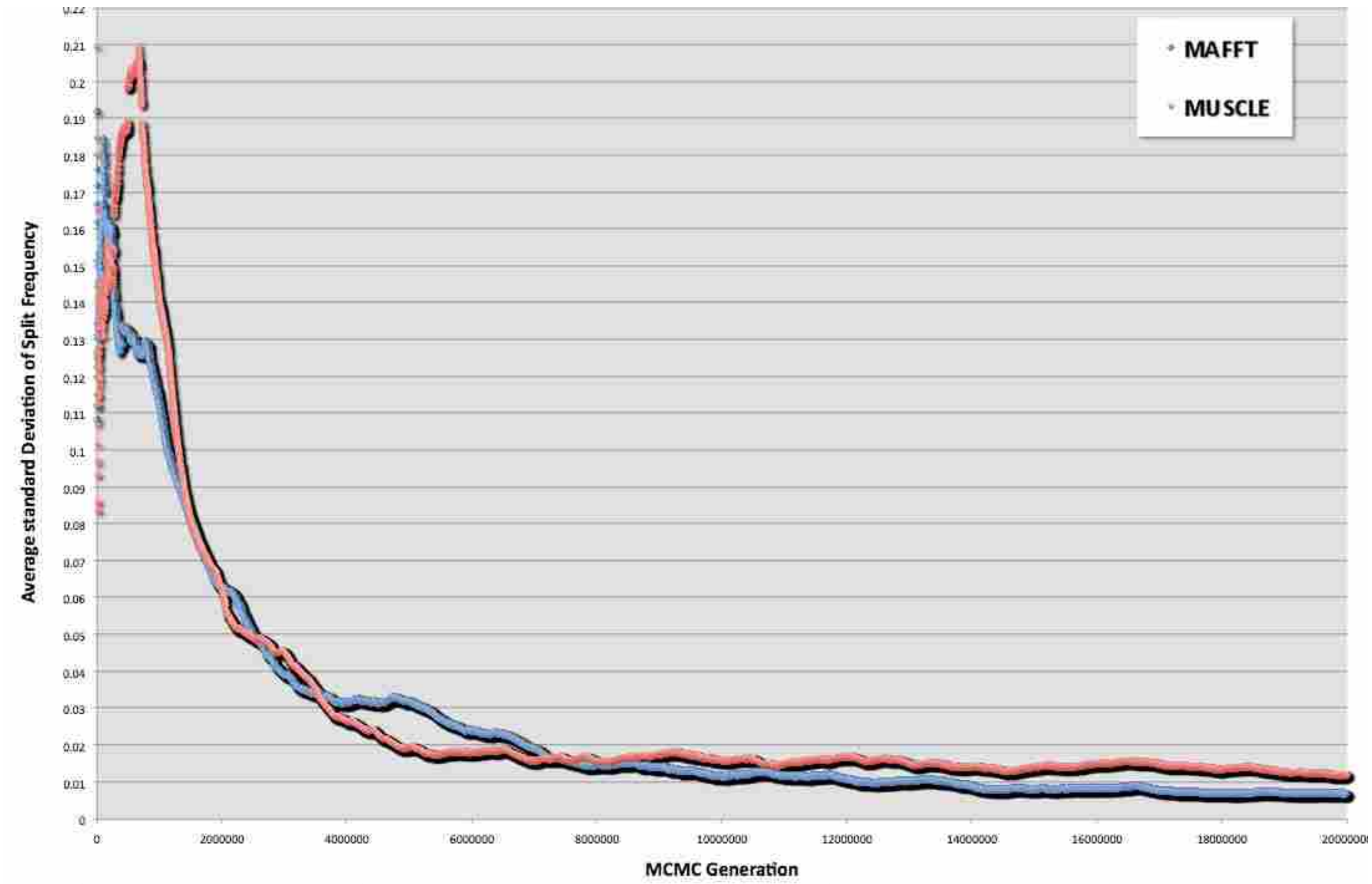


Figure 8. Linear plot of the average standard deviation of split frequencies (σ_{SF}) for the MAFFT (blue) and MUSCLE (red) based MCMC searches performed using MrBayes 3.12. Final σ_{SF} values at the end of the MCMC runs were <0.006 and <0.012 for the MUSCLE and MAFFT data sets, respectively. A clear plateau and stable decrease of the σ_{SF} value can be seen in both searches from about the eight millionth generation, suggesting the independent chains have begun to sample the same space.

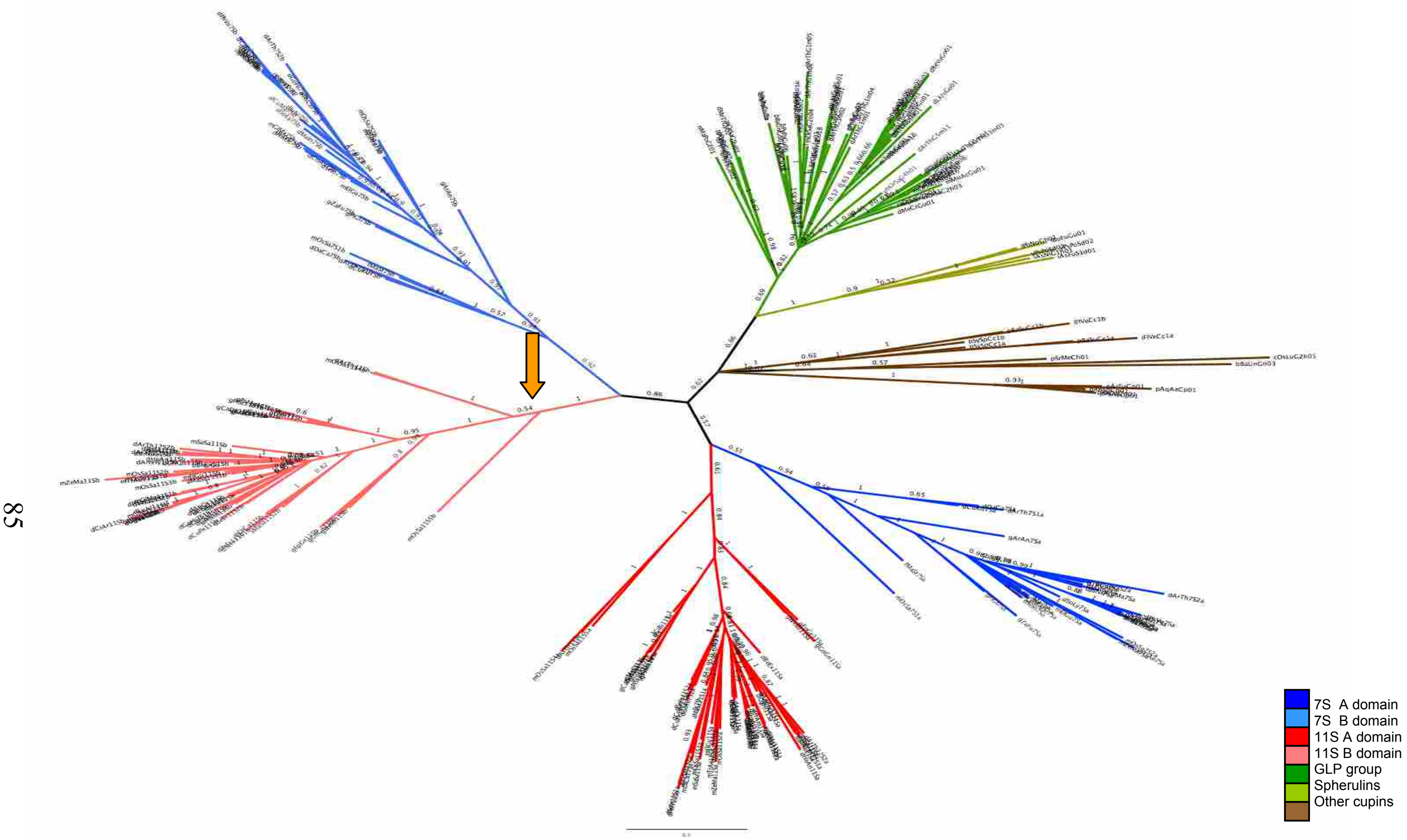


Figure 9. Bayesian consensus tree generated from 20 million generation MCMC search based on MUSCLE alignment of individual cupin domains using MrBayes 3.12. Major cupin and globulin domain groups are color coded according to the legend. Basal nodes of the globulin domain groups have only moderate to low support (pP<0.95). Orange arrow highlights main incongruence between the resulting MAFFT and MUSCLE (Figure 9) trees. This difference affects the interpretation of the mechanisms giving rise to the bicupin globulins.

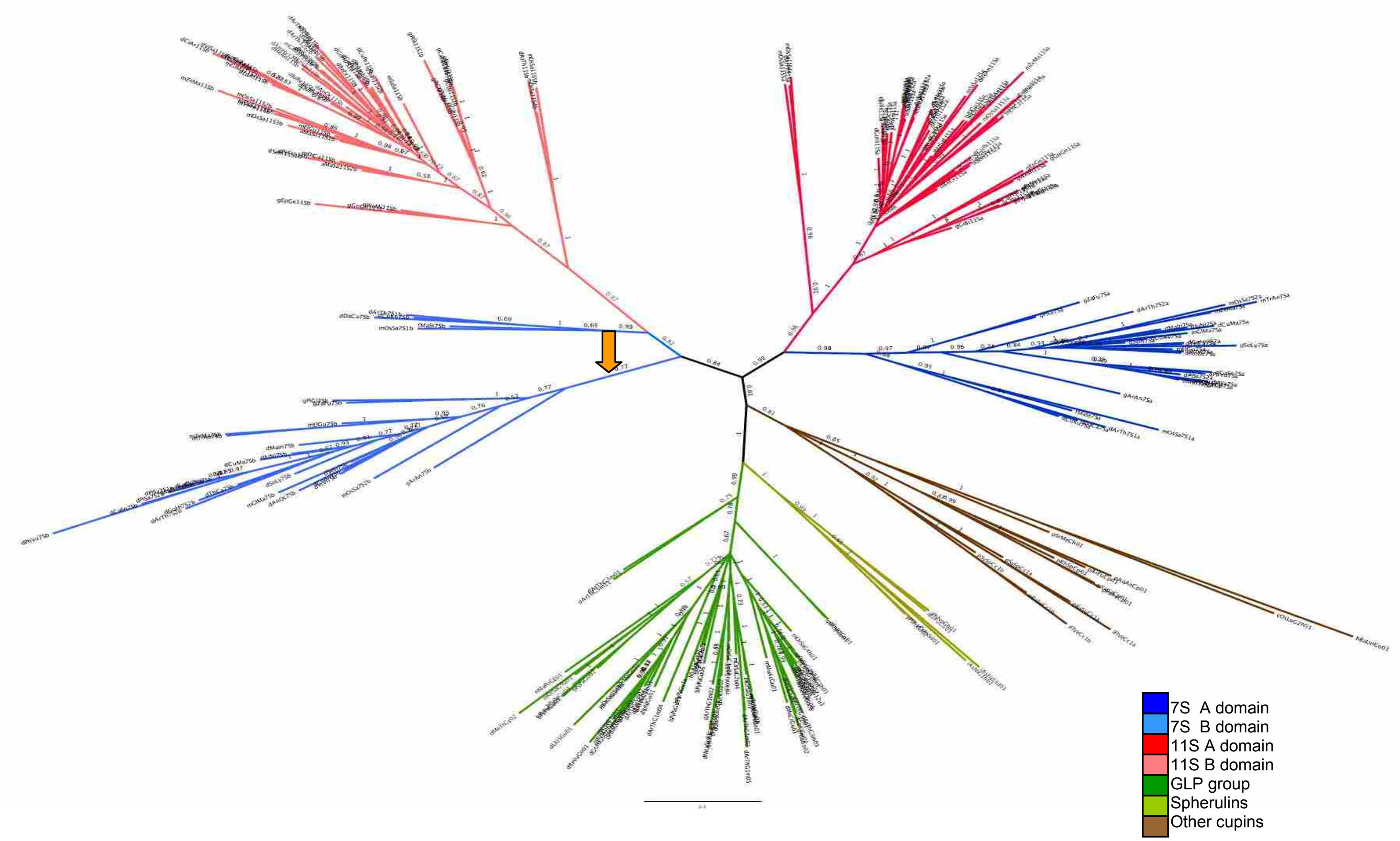


Figure 10. Bayesian consensus tree generated from 20 million generation MCMC search based on MUSCLE alignment of individual cupin domains using MrBayes 3.12. Major cupin and globulin domain groups are color coded according to the legend. Basal nodes of the globulin domain groups have only moderate to low support (pP<0.95). Orange arrow highlights main incongruence between the resulting MAFFT and MUSCLE trees (Figure 9). This difference affects the interpretation of the mechanisms giving rise to the bicupin globulins.

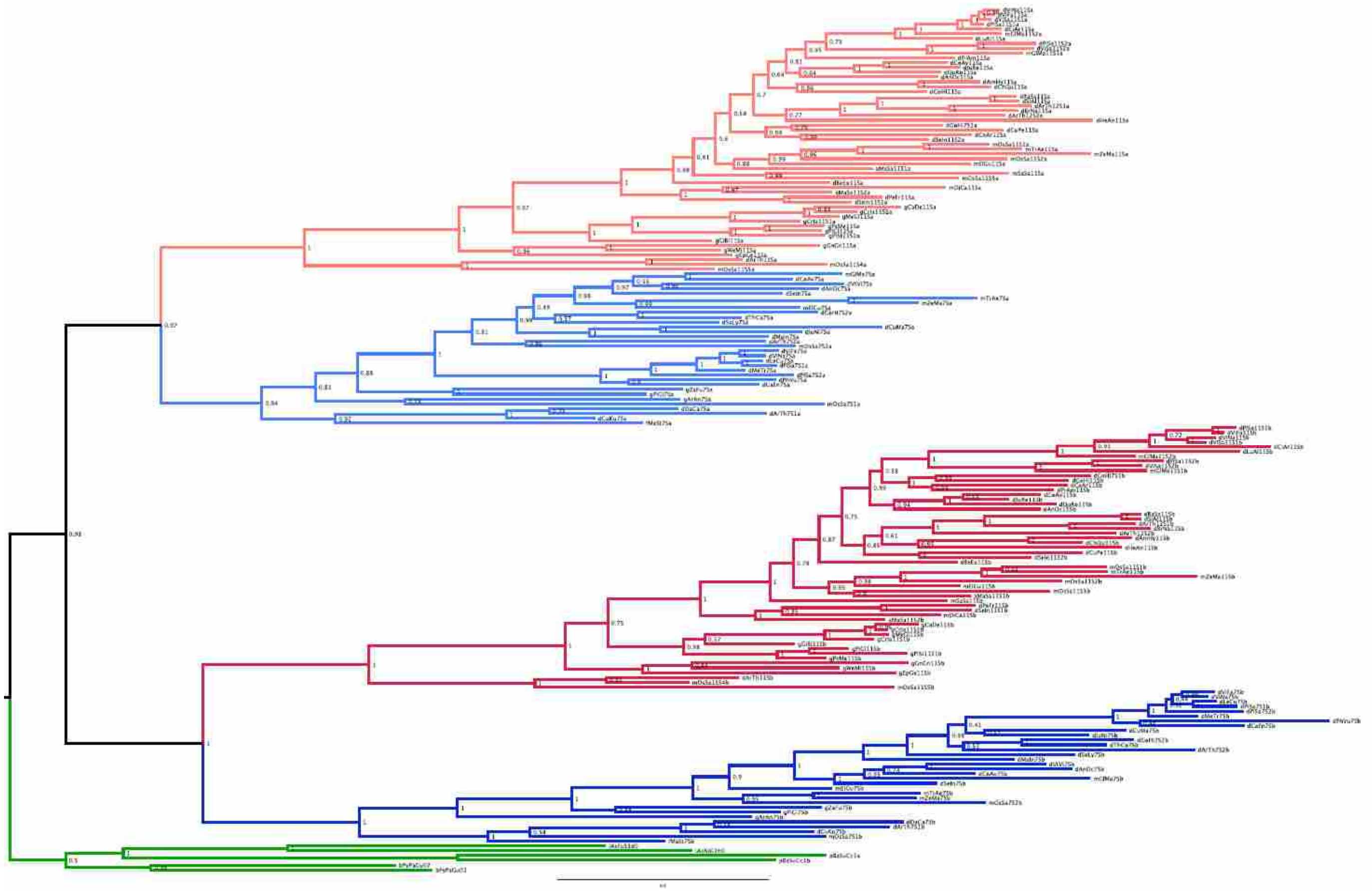


Figure 11. Phylogenetic tree reconstruction generated using MUSCLE alignment of cupin domains from globulin storage proteins. Tree is rooted using a non-globulin out-group (green) including two GLP sequences from *Physarum polycephalum*, a spherulin-like and a GLP sequence from the fungal genus *Aspergillus*, and the two cupin domains from an oxalate decarboxylase gene of *Bacillus subtilus*. N-terminus ("a" group) domains are light colored, and C-terminus ("b" group) domains are dark colored. Domain sequences pertaining to the 11S-type storage proteins are red, and domain sequences pertaining to the 7S-type storage proteins are blue. Nodes are labeled with pP values from Bayesian estimation of posterior distribution.

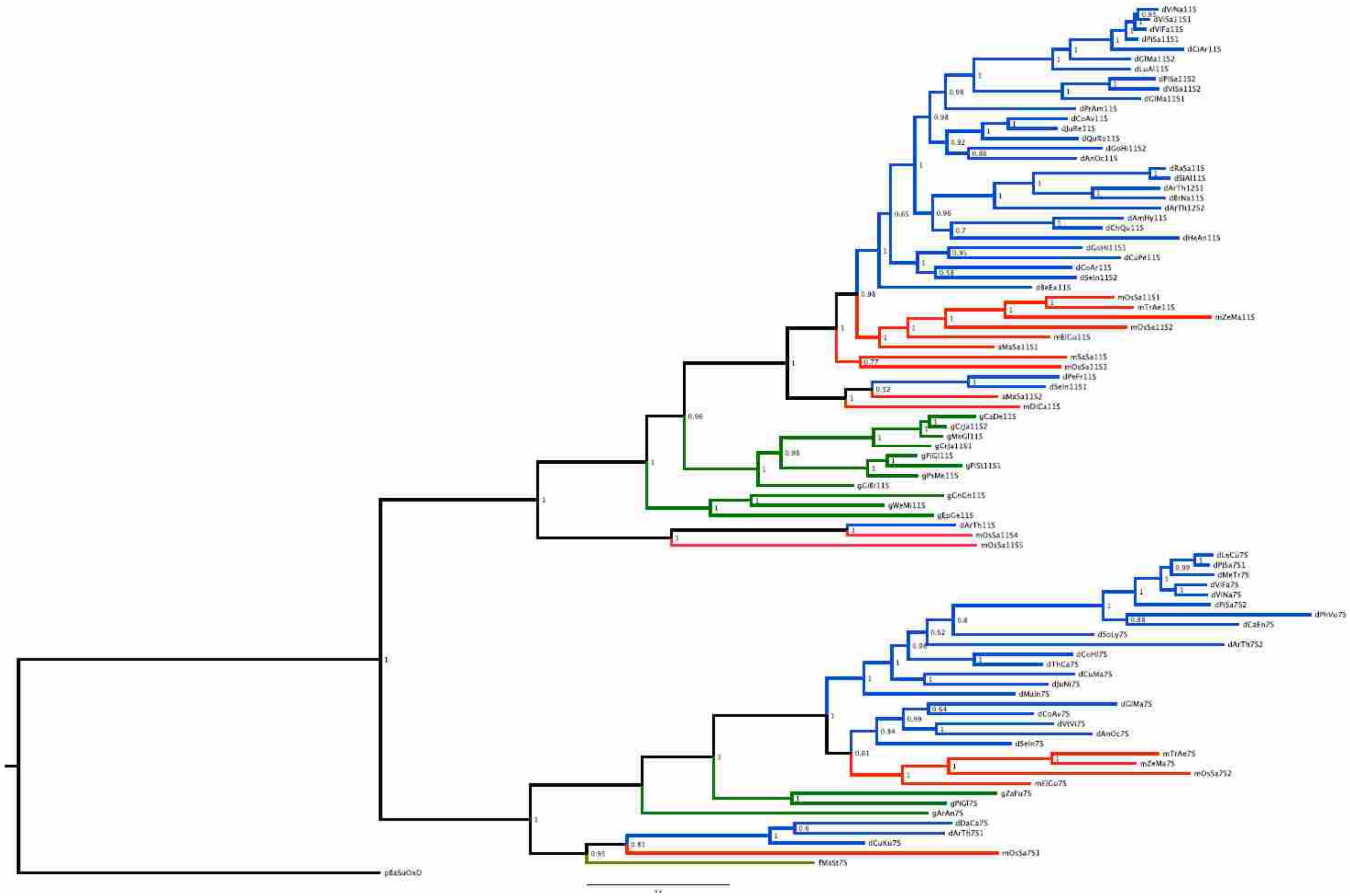


Figure 12. Phylogenetic tree reconstruction based on the MUSCLE MSA of globulin storage protein CDS's. Tree is rooted using the oxidase decarboxylase gene sequence from *Bacillus subtilis*. Branches are colored according to taxonomic groups (brown – filicophyta, green – gymnospermata, red – monocot, blue – dicot). Nodes are labeled with pP support values from Bayesian estimation of posterior. With the exception of the sequences from the bacterial out-group and the fern *Matteucia struthiopteris*, all other sequences are from the spermatophyte group.

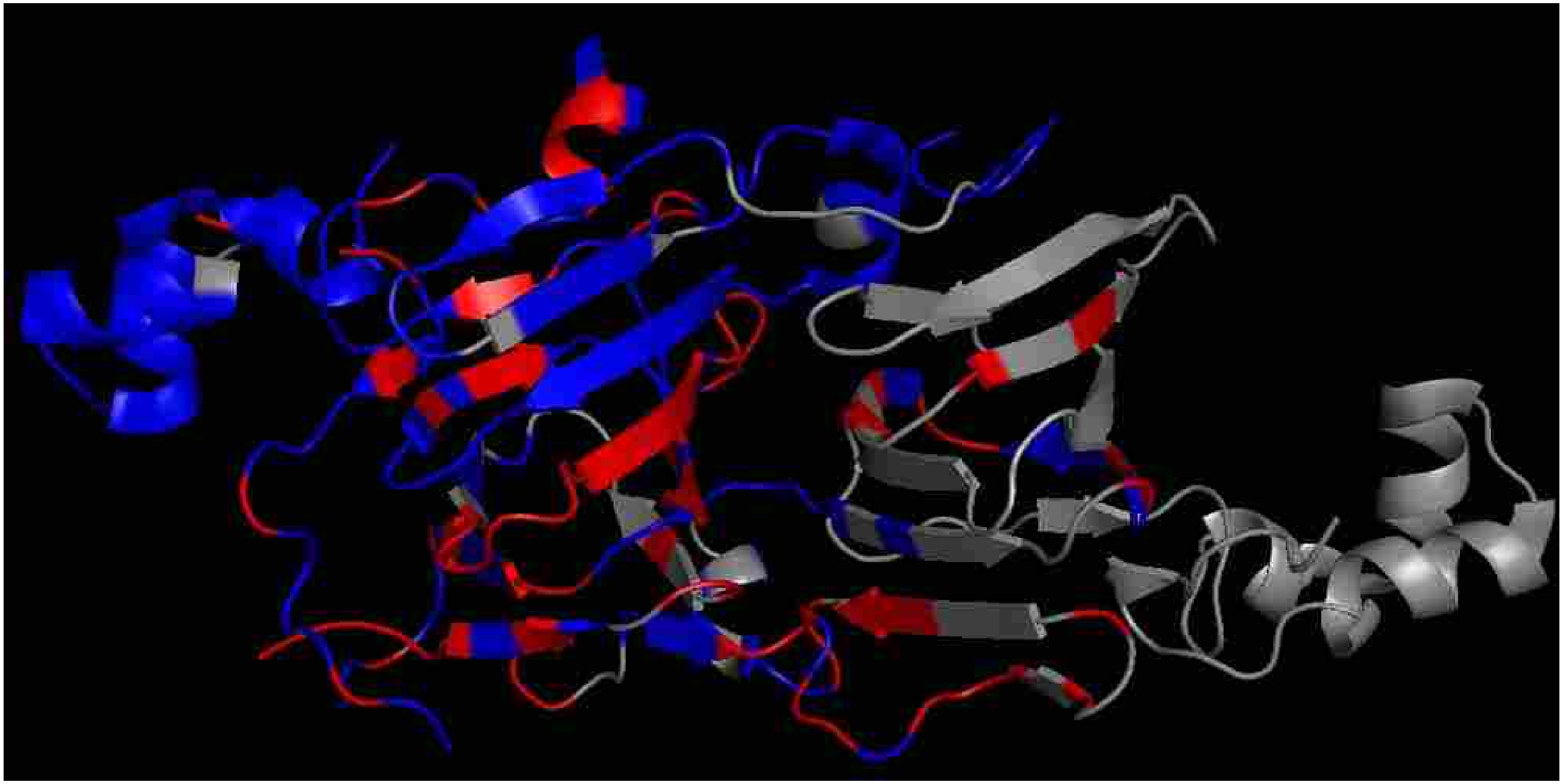


Figure 13. 11S (1fxz) shows differential in selection patterns between the two domains. Destabilizing (red) selection and stabilizing, or minimal, (blue) selection appears to have occurred with greater frequency along the N-terminus domain (left), while the C-terminus (right) domain shows much less significant change. This pattern is suggestive of functional and/or structural differentiation between the duplicated domains.

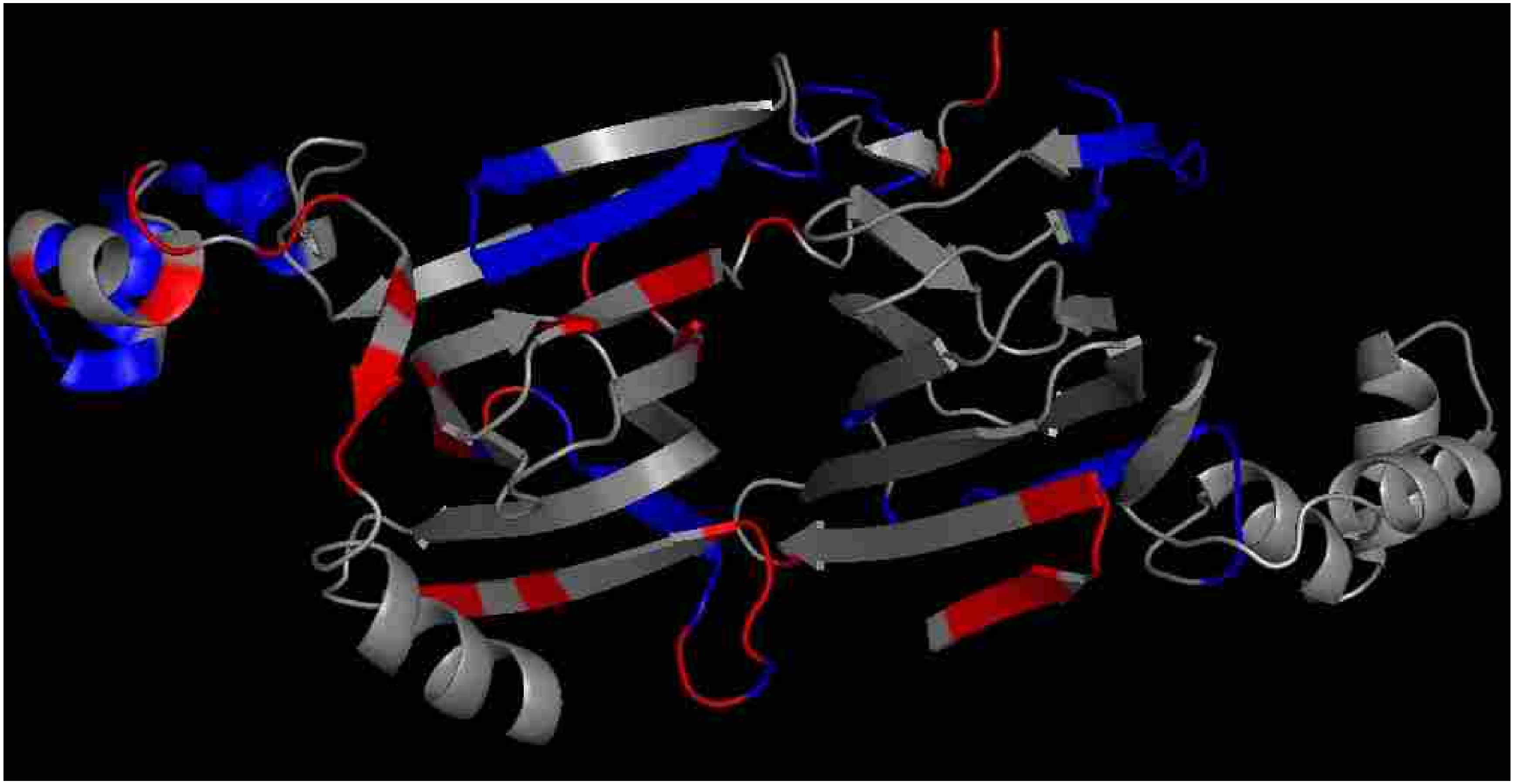


Figure 14. Patterns of differential selection in the two cupin domains of the vicilin (2phl) protein are shown, with destabilizing (red) and stabilizing, or minimizing (blue). Note the N-terminal helix region is under stabilizing selection, while the C-terminal region shows no pattern of non-synonymous substitution that tends towards stabilizing or destabilizing selection.

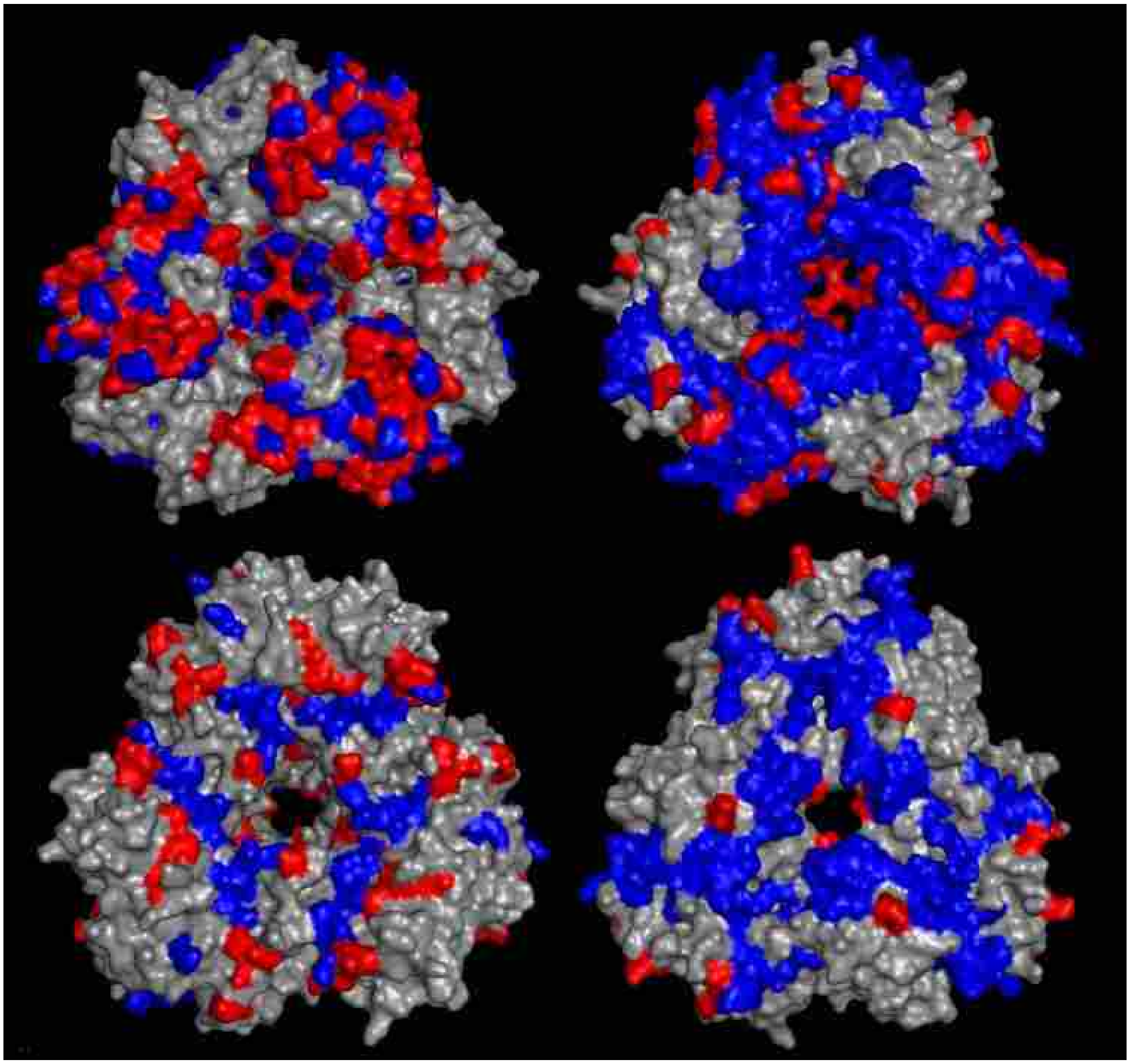


Figure 15. Legumin (1fxz, top) with regions of radical (red) and minimal (blue) selection highlighted. Vicilin (2phl, bottom) also shown with sites under destabilizing and stabilizing selection. Plane of legumin trimer involved in hexamer formation (top, left) and the corresponding face of the vicilin trimer (bottom, left) show considerable differences in pattern of selection. The opposite (outer) plane of the legumin trimer and corresponding face of vicilin show remarkably similar patterns of selection, being under predominantly stabilizing selection across similar regions of the protein. In both legumin and vicilin examples amino acid residues which protrude into or form part of central pore are under radical selection.

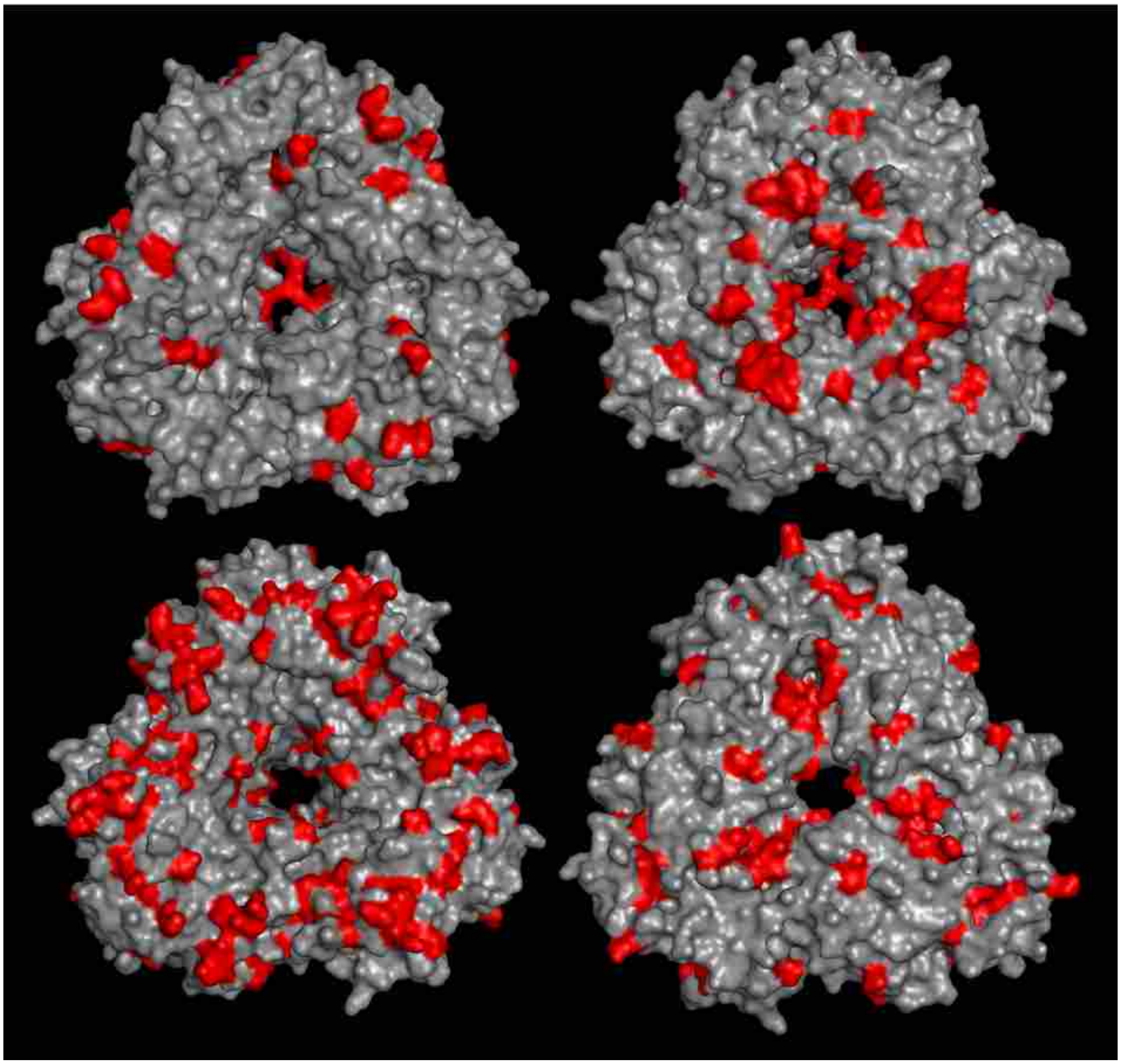


Figure 16. Views of patterns of sites under selection for 11S-like non-storage protein (superimposed on 1fxz) and 7S-like non-storage protein (superimposed on 2phl). Images are arranged as in Figure 15. Although patterns differ significantly between 11S-like and 7S-like groups, the radical selection occurring surrounding the central pore in both groups is notable.

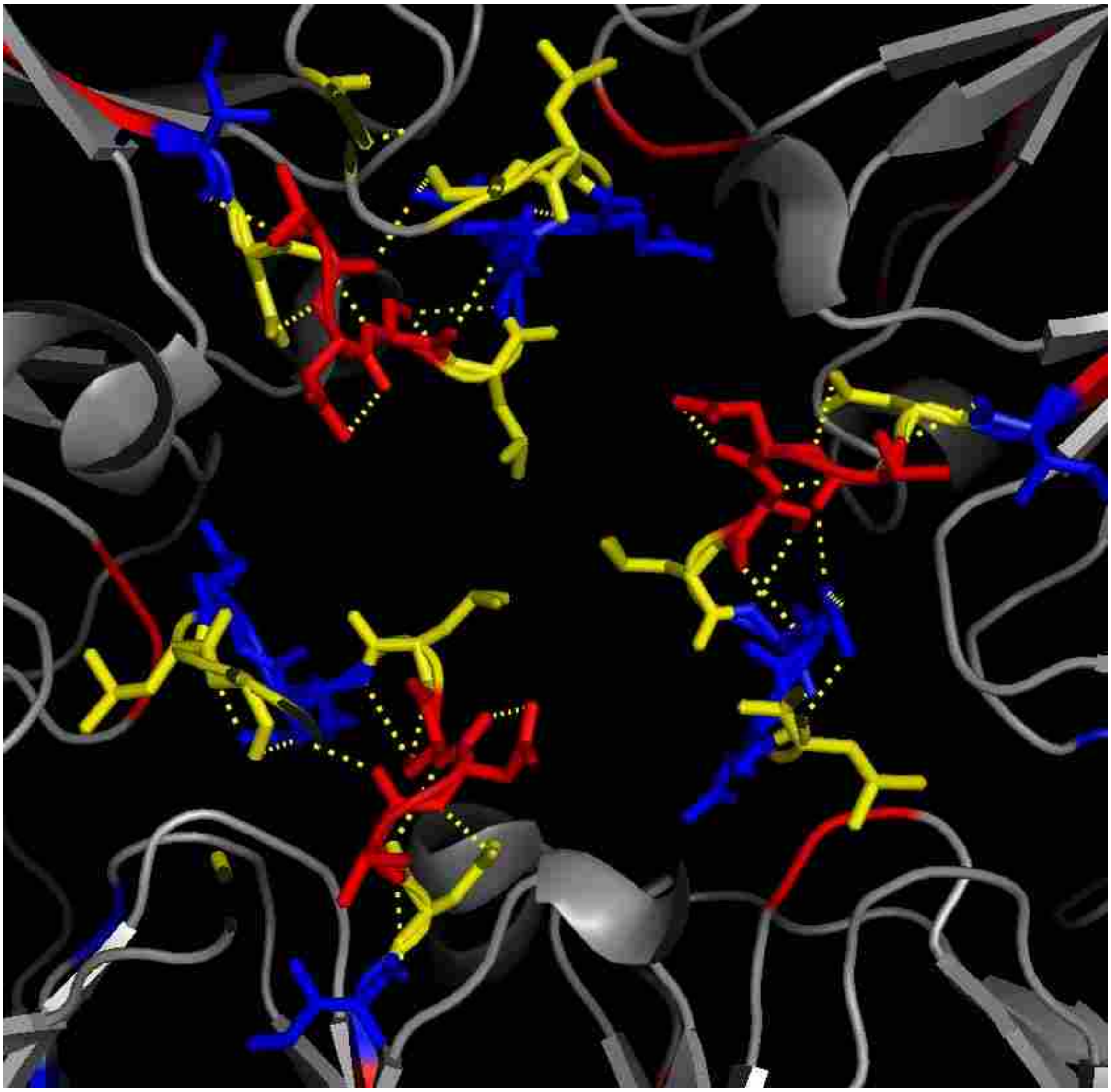


Figure 17. Inner core of legumin group (shown on 1fxz) with sites under selection colored for destabilizing selection in alpha-helical tendencies (yellow), and power to be at the C-terminal (red), and for stabilizing selection in average number of surrounding residues (blue). Predicted bonding between adjacent polar contacts shown by dashed lines. View is from internal side of legumin trimer.