Genome Biology and Evolution Advance Access published February 16, 2012 doi:10.1093/gbe/evs015

Genome Biology and Evolution

SMBE

The Plant Proteome Folding Project: Structure and Positive Selection in Plant Protein Families

M.M. Pentony[1], P. Winters[1], D. Penfold-Brown[1], K. Drew[1], A. Narechania[2], R. DeSalle[2], R. Bonneau[1*], M.D. Purugganan[1*]

[1]Center for Genomics and Systems Biology, Department of Biology, 12 Waverly Place, New York University, New York, NY 10003

[2]American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024

*co-corresponding authors: mp132@nyu.edu and bonneau@nyu.edu

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

© The Author(s) 2012. Published by Oxford University Press on behalf of the *Society for Molecular Biology and Evolution*.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Despite its importance, relatively little is known about the relationship between the structure, function and evolution of proteins, particularly in land plant species. We have developed a database with predicted protein domains for five plant proteomes (http://pfp.bio.nyu.edu), and used both protein structural fold-recognition and *de novo* Rosetta-based protein structure prediction to predict protein structure for Arabidopsis and rice proteins. Based on sequence similarity, we have identified ~15,000 orthologous/paralogous protein family clusters among these species, and used codon-based models to predict positive selection in protein evolution within 175 of these sequence clusters. Our results show that codons that display positive selection appears to be less frequent in helical and strand regions, and are over-represented in amino acid residues that are associated with a change in protein secondary structure. Like in other organisms, disordered protein regions also appear to have more selected sites. Structural information provides new functional insights into specific plant proteins and allows us to map positively selected amino acid sites onto protein structures and view these sites in a structural and functional context.

Keywords: Adaptation, Protein Structure, Plant Evolution, Fold Prediction

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

INTRODUCTION

While genomes remain complex structures that encode a large number of genetic entities, protein-coding genes arguably represent the largest and most important component of eukaryotic genomes. A large fraction of eukaryotic proteins are encoded by gene families which evolve by gene duplication and diversification . In the model plant species *Arabidopsis thaliana*, for example, nearly 1,000 gene families have been identified, which together account for >8,000 (~33%) of protein-coding loci, and the numbers for rice are comparable .

Adaptive evolution in organisms can proceed through the diversification of these protein-coding genes and gene families, and understanding the nature of evolutionary change requires us to understand how proteins evolve, both in structure and function. Methods of phylogenetic analysis that can reconstruct protein domain families are well-described, including maximum parsimony and Bayesian maximum likelihood methods, and several methods have been developed that can identify positive selection in key amino acid sites in evolving proteins .

Despite the intense interest in protein-family diversification however, detailed evolutionary analyses have only been undertaken for a few plant gene families, including those that encode the myb-like , homeodomain , MADS-box and proteasomal proteins. The potential exists for studying how selection of amino acids can occur in a structural context, and a few studies have started to incorporate structural information in the evolutionary analyses of gene families .

A major obstacle to studying the structural evolution of proteins is the lack of well-defined structures for the vast majority of eukaryotic proteins. While genomics projects have become adept at obtaining the primary sequence of the entire complement of protein-coding genes in genomes, annotations that depict secondary or tertiary structures remain sparse. It is

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

thus imperative that we develop methods to extend the annotations of genomes by incorporating protein structural information .

We have developed methods that take whole-genome protein sequences and provide structural annotation of these proteins. These methods include matching regions of protein sequence to known structures in the Protein Data Bank (PDB) [http://www.rcsb.org/pdb] using PSI-BLAST and fold recognition methods [FFAS] in order to predict domain boundaries and annotate predicted domains with structure information. Select domains that elude identification by BLAST and fold recognition are then predicted using Pfam (http://pfam.sanger.ac.uk) and heuristic approaches, and are considered contenders for Rosetta *de novo* structure prediction . Using these approaches, we have previously completed structural/functional annotation of 94 proteomes and Rosetta-generated structure predictions have specifically produced functional insights even when function is not evident from sequence-based analyses alone (Bonneau, Tsai et al. 2001; Bonneau, Strauss et al. 2002).

To date, several plant proteomes have been fully sequenced, with several more currently underway. Like other organisms, however, this increase in availability of sequence information has not been matched with an increase in known plant protein structures or known protein functions. In this work we have applied our structure-based annotation method to plant genome data to examine the evolution of selected amino acids on plant protein families in the context of their structure. We focused on the angiosperms, which are arguably the most diverse major plant group on the planet, with over 260,000 living species, in more than 450 families .

We have predicted structural domains for all known proteins in five flowering plant proteomes - *Arabidopsis thaliana, Oryza sativa, Populus trichocarpa, Sorghum bicolour* and *Vitis vinifera*. To our knowledge, this is the largest database of inferred protein structures

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

currently available for plants. Using OrthologID , we have also identified ~15,000 gene families (categorised as alignments with at least 2 sequences) within these five proteomes. Using codon-based models, we have done selection analysis on amino acid sites for 2,120 gene families and examined the structural context of these positively selected sites. Finally, we placed these positively-selected sites in a structural context by highlighting and visualizing these sites on corresponding three-dimensional predicted protein structures.

MATERIALS AND METHODS

Genomic and proteomic data

Phylogenies and alignments were generated from analysis of five plant species which have complete genome sequences available: *Arabidopsis thaliana* (http://www.arabidopsis.org/ Version: 9), *Oryza sativa (cv. nipponbare)* (http://rice.plantbiology.msu.edu/, Version: MSU6), *Vitis vinifera* (www.Gramene.org, Version: 2007-12-IGGP), *Populus trichocarpa* (http://genome.jgi-psf.org/ Version: 2004-12-JGI) and *Sorghum bicolour* (http://genome.jgi-psf.org/, Version: Sbi1). The complete proteomic and genomic sequences were downloaded from the Gramene website (http://www.gramene.org/) using BioMart (www.biomart.org). Both *A. thaliana* and *O. sativa* annotations were listed as fully complete, with the remaining three taxa in draft assembly.

Protein/gene family choice

To identify gene families, we modified OrthologID [OID] , which is a semi-automated homology search and phylogeny reconstruction pipeline. OID was modified to remove the MAFFT alignment refinement step (Kato et al. 2002), which removes ambiguously aligned

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

regions. This allowed for the OID protein alignments to be mapped correctly to their respective nucleotide alignments. Using OID, gene family alignments and phylogenies were generated from the annotated protein-coding genes in our study species using BLAST , with an expectation cut-off of e^{-20}. OID produced 14,822 putative gene families, with family sizes ranging from n = 2 to 1,600 gene sequences (see Figure 1). We define gene family based on sequence information with a BLAST analysis cut-off e^{-20}, a stringent cut-off has been used in previous studies (for example, Kinsella et al. 2003; Xiao et al. 2007) to delineate gene/protein families. Not included in this count are the 66,478 orphan sequences (n = 1) produced by OID. Due to the high number of OID results that contained less than 10 gene sequences, we limited our final analysis to only those gene families with at least 10 sequences, regardless of whether each alignment included a representative from each species. This reduced the dataset to 2,230 gene families, each corresponding to a plant protein domain family.

Due to the large evolutionary distances between the organisms used in this study (~120-200 myr between monocots and dicots), the alignments were modified to remove excessive gapped regions. CodeML estimates ancestral sequences for alignments when estimating whether positive selection has occurred within the protein family. Excessive gapped regions make this task computationally difficult (i.e., computational run times >600 hours), and increase the possibilities of misalignments between homologous sequences; thus, large gapped alignments are not suitable for use in CodeML. Using multiple cut-off options, for both columns and rows, we determined the threshold percentage of gaps that allowed for feasible use with CodeML. Based on this, we culled each alignment in two ways. If the length of an aligned sequence contained at least 70% gaps, the sequence was removed. Further, if more than 30% of a column in the resulting alignment was comprised of gaps, that column was removed from the alignment.

SMBE

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Inspection of the data at this point showed alignments with greater than 100 sequences were still excessively gapped and would not be suitable for CodeML analysis, so we removed these alignments from our analysis. This further reduced our dataset by 110 gene families, to 2,120 families. This left a count of 46,667 sequences in our final analysis, with alignment sizes ranging from 10 – 100 sequences.

Gene family groups with known function were initially identified using the known families listed on The Arabidopsis Information Resource (TAIR) website (http://www.arabidopsis.org). In addition, we annotated the gene families in our analysis with known functional groups gleaned from the literature, as well as homologous relationships found in the Gramene website. This led to the identification of 192 families with previously annotated/ known functions and containing at least 2 members of each of the 5 species (approximately 30,000 sequences).

Positive selection analysis

Positively selected sites were predicted using CodeML from the PAML package . CodeML uses different evolutionary models to account for differences in transition/transversion rates in DNA, and also codon usage biases found within degenerate codons, and uses maximum likelihood to estimate the fit of levels of sequence divergence in homologous sequences given these models . Five models were used in this analysis: M0-M3 and M7-M8 with ambiguous residue positions included and using the F3X4 codon frequency model.

CodeML calculates the ratio of synonymous (dS) and non-synonymous (dN) changes that have occurred at each codon with a dN/dS > 1 suggestive of positive selection. Each CodeML model builds on the preceding one, and adds additional dN/dS (ω) classes. The most basic model

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

M0, assuming all sites are undergoing negative / deleterious selective pressure (dN/dS < 0; 1 class of sites). M1 allows for some sites to be under neutral selection (dN/dS = 1), while M2 allows for some sites to be under positive selection (dN/dS > 1). As each model is applied to the data more complex parameters are applied, allowing for multiple dN/dS classes in the dataset. The most complex model we used was M8, which allows for 13 classes of ω sites. If no positive selection was found using the basic M2 and M3 selection models, we did not proceed to the detailed codon selection analysis using models M7 and M8. For a more detailed description of CodeML classes, models and the statistical analyses involved, see Yang et al. .

Protein domain prediction

Ginzu , was used to predict domains for all proteins in our five proteomes (213,587 proteins). Ginzu first searches for sequence matches to known three-dimensional protein structures using PSI-BLAST , providing structural information for predicted domains. Ginzu searches then for matches to experimental structures using fold recognition . Domains in regions of protein not matched by PSI-BLAST or fold-recognition are predicted using matches to Pfam and a heuristic that predicts domain boundaries based on patterns within multiple sequence alignments. These latter domains, which lack structural information, are then exported (if < 165 amino acids) for external structure prediction via Rosetta, resulting in domain and structural predictions of varying confidence for all proteins considered in the Plant PFP. Using these methods resulted in 409,017 predicted domains. Although we predicted domains for all five proteomes, protein-folding structure prediction was performed only on *A. thaliana* and *O. sativa* due to the large computational overhead required for folding proteins.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Prediction of protein structural elements and positive site mapping

PSIPRED and DISOPRED2 was used to predict protein folds and disordered regions respectively. PSIPRED uses neural-networks to analyse the position-specific scoring matrices produced from PSI-BLAST to infer secondary structure and is one of the top secondary structure prediction methods available. Disordered regions are defined as those regions that do not fold into a three-dimensional structure in their native state. Disordered regions are flexible, dynamic and can be partially or completely unfolded in solution. DISOPRED2 uses known structural information, coupled with sequence records, to infer disordered regions.

For each residue position in an alignment, we use PSIPRED and DISOPRED2 to categorise specific amino acid sites into what secondary structure element they were found. If at least 80% of residues at a particular alignment position were predicted to be of the same fold / residue type, we classed the amino acid site as belonging to that type. For those sites that did not show outright support for a particular protein fold / class, we classified it as a mixed site.

DSSP was also used for additional secondary structure annotation information. Using PDB atomic co-ordinates, DSSP defines secondary structure, geometrical features and solvent exposure of proteins. We also obtained SCOP structural information. The Structural Classification of Proteins (SCOP) database uses manual inspection, with the help of automated methods, to predicted structural and evolutionary relatedness.

RESULTS

Database overview

The Plant Proteome Folding Pipeline (Plant PFP) website is available at http://pfp.bio.nyu.edu and currently represents 213,587 proteins from our five chosen organisms

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

along with their respective protein structure predictions. The Ginzu pipeline as described by Drew et al. was used to analyse 211,140 of these proteins, skipping 2,447 proteins due to their excessive length or high percentage of residues predicted to be in disordered regions. From these 211,140 proteins, Ginzu produced 409,017 domains (listed as 'Domains' in Table 1). The 173,820 domains predicted by PSI-BLAST and fold recognition methods were automatically associated with their top matching PDB structure. The remaining 235,197 domains were considered for Rosetta *de novo* structure prediction. 29,202 domains were returned from Rosetta with predicted structures, 4,769 of which are considered to be high-confidence, where high confidence is determined by an MCM score of 0.8 or greater, which correlates to the high atomic accuracy of the predicted structure in relation to the native structure [see Table 2] . To evaluate the accuracy of high-confidence structure predictions, a double-blind benchmarking of the structural analyses methods were used, and these correctly predicted 47% of structures using SCOP (v1.67) superfamily classifications , which is high for computational structure prediction. Comparison of the predicted and experimentally determined structures showed a strong correlation in structure prediction .

Data visualisation

To facilitate the exploration of predicted sites of positive selection mapped onto structures and the exploration of our predicted domain families, we have constructed a web interface to our resource (http://pfp.bio.nyu.edu). This web interface allows searching via accession or ontology term, by sequence using BLAST, or by searching the list of predetermined functional families by name (i.e. "bHLH Transcription Factor"). User selection of a family group

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

produces an initial page showing the phylogeny and sequence identifiers, with each identifier listing the mapped protein domains.

This initial window contains three tabs - JALview, JMol and Phylowidget. JALView is used to display the multiple sequence alignment (MSA) of the proteins in the family, Phylowidget to display the family's phylogenetic tree and JMol (http://www.jmol.org) to view the three-dimensional structure (either a PDB structure or a Rosetta *de novo* predicted structure) of the individual proteins of the family. The website also provides a display that shows the division of family proteins into domains via the Ginzu program, and displays the methods by which Ginzu predicted each domain.

Selecting a mapped protein domain within a sequence that has a structure annotation and then selecting the 'Load in JMol' link loads the domain into the JMol tab, with the sites of positive selection highlighted (see Figure 2 for a JMol view). If a domain maps to a known protein structure in PDB, a link to this protein is included. Positively selected sites are highlighted along the structure in blue, while the atoms of selected residues are circled in yellow, which enables the choosing of specific positively selected residues within the alignment to be viewed easily on the structure.

Positive selection prediction for protein families

Using Likelihood Ratio Tests (LRT) and dN/dS values from model M8 of CodeML, we determined selection had occurred in 175 gene families, which indicated that 8.2% of families show selection. In total, there were over 4,000 sites that showed positive selection. Ignoring families that did not contain any selected sites with >95% confidence in prediction ($p > 0.95$ for $dN/dS > 1$), resulted in 938 selected sites in 122 gene families. The majority of families (97) had

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

10 or less sites of positive selection. Although this created a very conservative dataset, it increased confidence in the resulting analyses. There were 10-95 sequences per alignment, with a mean size of 20. Alignment length ranged from 51 to 1,372 residues, with a mean sequence length of 321 residues. Mapping these results back to the families where a gene function has been associated, we found 43 gene families mapped to 19 known functions (Supplementary Table 1).

Initial codeML undertook analyses using sequence alignments for gene families that contained both paralogues as well as orthologues. With the high level of gene duplication and polyploidization that has occurred within plant genomes, identification of orthologous genes can prove difficult. Previous work on a comparison of paralogous and orthologous evolutionary rates by Conant et al. (2007) found that the evolutionary rates do not differ significantly between orthologs and paralogs. Nevertheless, we also expanded on the current analysis, by separating gene families with positively selected sites into separate paralogous sequence alignments from each individual species. We used only alignments that (i) originally contained more than 1 species and (ii) contained at least 7 sequences from a species, and undertook codeML analyses using the same parameters as previously described. From the 122 alignments, we found that 101 met our criteria. We then separated out 142 separate paralogous gene families of individual species from these 101 alignments, and re-analyzed the individual sequence alignments for positive selection. A re-analysis for sites of positive selection that included alignments from multiple species (i.e., potential orthologues) had 923 sites of positive selection ($p > 0.95$). We find that 722 identified selected sites overlapped in the two analyses, which suggests that our original analyses is able to identify ~80% of the positively-selected sites found in paralogous gene families (while also identifying sites that were selected between putative orthologues). It

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

must be noted that analysing only the paralogues also identified sites that were not observed in the larger datasets; however, we feel that the results from the larger datasets that also include putative orthologues provide greater power and confidence in the selection analyses.

One other issue is whether the sequences we were analyzing were too divergent and the synonymous sites were saturated. Previous work suggests that codon-based models estimating selection on protein sequences are valid only when synonymous site divergence of $dS < 5$. We calculated the mean pairwise dS for each of our gene families and only 12.4% having $dS < 5$. However, sequence alignments for which we found positive selection had 45.8% with a mean pairwise $dS < 5$, and the majority of these had dS values between 0-1. If we only consider sequence alignments at the range of $dS < 5$, we find that ~21% of these gene families show evidence for positive selection.

Amino acid properties of positively selected sites

Using amino acid properties from Lise and Jones , we categorised selected and non-selected residues using eight properties (hydrophobic, polar, small, aliphatic, aromatic, positive, negative and unique). We changed the property 'proline' from Lise and Jones, to 'unique', which included glycine and proline, due to their distinctive properties.

Comparing selected versus non-selected residues for each property showed statistically significant differences in 5 properties ($p < 0.0001$): hydrophobic, polar, small, aliphatic and positive. There was a significant increase in selected residues with polar, small and positive categories and a less than expected number of residues within aliphatic and hydrophobic categories. There was a slight increase in negatively charged residues within selected sites, but the results were not much different from expected (see Table 3). These results indicate that

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

hydrophilic, small and positively charged residues are possibly more prone to evolve with positive selection than other residue types.

Clustering of selected sites

To examine the possibility that the length of the protein sequence impacts the relative number of selected sites, we examined the distribution of selected sites in relation to the length of the sequence alignments, both the original alignments and the alignments that were culled of excessively gapped regions. We used a Spearman's rank correlation test to examine the relationship between sequence length and number of selected sites (i.e., as protein length increases, does also the number of positively selected sites). For proteins with the same length, the mean number of selected sites was used. The closer the score is to ±1, the more correlated the two variables. Spearman's rank correlation scores of 0.16 ($p = 0.07$) and 0.33 ($p = 0.0002$) were found for the culled alignment lengths and the original alignment lengths respectively. These results indicate that, at least when considering the unculled alignment lengths, there does appear to be a moderately low but significant correlation between increased length and number of positively selected sites.

We looked to see whether the sites under positive selection showed evidence for clustering along the sequence. Using a pairwise distance measure, for each alignment we found the average distance between sites of selection along the primary sequence. We chose an equal number of sites at random and found the average pairwise distance in the dataset, and we repeated this random selection of sites 1,000 times. Using a 95% cut-off, we found 40% of alignments had a smaller pairwise distance between selected sites compared to 95% of the random distances. This increased to 58% if we use a 90% cut-off. This suggests that there is

SMBE

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

significant sequence-space clustering of positively selected sites in a substantial number of proteins.

To corroborate this result, we did a sliding window analysis using a window size of 5 residues moving one amino acid at a time. Comparing the location of selected sites, with 10,000 permutations based on random site selection, we found that in only one case did the number of windows containing random amino acid sites equal the number of windows containing actual sites of selection. Together, these results indicate that the majority of selected sites are clustered within plant protein gene families.

Secondary structure prediction of positively selected sites

We predicted the secondary structure of all sequences using PSIPRED and created secondary structure alignments. When comparing the secondary structure distribution of positively selected versus non-positively selected residues, we found a significant reduction in predicted helical ($p < 0.0001$) and strand residues ($p < 0.0001$) that contain selected sites (see Figure 3). There appears to be slight reduction of predicted coiled residues with selected sites, although the results were not significantly different than expected. Interestingly, 66% of sites could not be classed within any definitive secondary structure type, and were classed as of mixed structure.

Looking more closely at this 'mixed' group, we divided the data into amino acid positions that showed only combinations of coils and sheets (CS), coils and helices (CH), sheets and helices (SH) and positions that still contained all types of secondary structure elements (ALL). In each of these mixed group categories, the positively selected sites were significantly over-represented (CH, $p < 0.006$; CS and ALL, $p < 0.0001$; SH, $p < 0.03$).

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Initially if an amino acid position was found to occur in particular secondary structure type in at least 80% of the sequences in an alignment, we classified that amino acid position as being of that type. Modifying this to include 60, 70, 90 and 100% cut-offs, we found all secondary structure element counts remained the same, except for counts for coils, and within the mixed group CS. As the stringency of the cut-off increased, the number of coiled-only regions decreased and the number of CS regions increased within both selected and non-selected positions, indicating that the change from coils to sheets or vice versa may be more flexible than other structural changes.

Using DISOPRED2 we found that positively selected amino acid sites (compared to non-selected sites) were found largely in disordered regions. We found an increased number of sites of positive selection predicted to be disordered than expected ($p < 0.0001$) [see Figure 3]. As disordered regions can undergo different folding conformations, our results could indicate that such flexible regions are under higher selective pressure. We found 65.7% of sites to be categorised as 'mixed'. Residues with a mixture of different secondary structure element types, and order/disordered etc. suggest that areas where structure changes from one fold type to another may be targets of adaptive evolution.

Solvent accessibility

Using DSSP, we obtained the Relative Accessible Surface Area (RASA) for the plant proteins in our analysis. RASA provides a measure of how exposed to a solvent an amino acid residue is within a protein structure; the lower the RASA score, the more buried the residue. The RASA score is between 0 – 1 with 0 being completely buried. We had RASA scores for 17,738 amino acids in our data. Of the 938 positively selected residues, we found RASA scores for 454.

SMBE

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

If a particular residue had more than one RASA score, we averaged the scores for this residue, unless the residue has a PDB score, in which case we used score that only. Multiple RASA scores were the result of the multiple methods used in the structure prediction (Rosetta, Ginzu). We divided the residues from all alignments into 3 groups, selected sites, conservative sites (those sites which showed no amino acid replacement) and all others.

A Wilcoxon test showed a significant difference ($p = 0.0001$) between the distribution of RASA scores within selected sites, compared to conservative sites, with conservative sites being more buried, suggesting more selective pressure occurs on more solvent-exposed residues. This is consistent with our analysis on amino acid properties, which showed selected sites to be more hydrophilic than no selected sites. Comparison of selected sites to all other sites did not show a significant difference ($p = 0.2$) (see Figure 4).

Case study 1: A mis-annotated C2H2 zinc-finger transcription factor family include TPR domain proteins

To demonstrate some of the capabilities of our structure database, we examine in greater detail two examples of gene families where we found evidence for positive selection. The first family is an OID group (http://pfp.bio.nyu.edu/family/18894) that is a family of 11 *Arabidopsis thaliana* proteins that are annotated on the TAIR family website as comprising C2H2 transcription factor proteins. C2H2 transcription factor proteins contain a zinc-finger domain and are involved in a wide range of functions, including DNA and RNA binding and protein-protein interactions . Of these 11 proteins, TAIR lists seven of them as containing a zinc-finger domain. Our analysis of the domain architecture of this family of 11 proteins suggests it has possibly been

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

mis-annotated. Searching each sequence manually against GenBank revealed the sequences did not return a hit to zinc-finger domain containing proteins.

Our Ginzu and Rosetta analyses predicted that these sequences had between 4-7 domains, with all but two N- and C- terminal regions mapping to known domains in the PDB. Non-terminal regions were predominantly unassigned, as domain boundaries were identified using less confident heuristic methods. Seven unique domains were mapped to the N-terminal region. Checking the PDB structures for these domains, all were from proteins characterised as TPR-domain containing proteins. TPR domains, first classified in yeast, are a repeating helix-turn-helix motif containing approximately 34 amino acids that are involved in multiple biological functions . Functionally, TPR domains have been linked to many roles including Hsp90 mediation, scaffolding proteins and transcription . Within this alignment, we found 54 sites of positive selection with greater than 95% confidence, of which 41 occurred within the predicted N-terminal domain (see Figure 5). Within the C-terminal, domains from 4 PDB structures were mapped, which were described as having ubiquitin-related functions.

Case study II: Selection in an F-box protein family is C-terminal to the F-box domain

Our second case example is a family (http://pfp.bio.nyu.edu/family/19187), which contains 23 sequences from *Sorghum bicolour* and *Oryza sativa* and are annotated as F-box proteins. F-box proteins, first described as part of the SCF (Skp, Cullin, F-box) complex, are involved in ubiquitination and are characterised by a structural motif of approximately 50 residues and constitute one of the largest plant multigene families . F-box proteins are comprised of an N-terminal F-box domain, which interact with Skp1, a linker domain and varying C-terminal domains, which are used to recruit substrates .

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Ginzu and Rosetta mapped domains from 14 proteins in this family to domains in three PDB structures. Twelve proteins mapped to the same PDB structure: 2ovr, along the entire length of the sequence, which includes a WD40 domain at the C-terminal end. Of the other two mapped structures, 1p22 and 2e32, 1p22 also contains a WD40 domain. The third structure, 2e32, differs by having a sugar binding domain (SBD) rather than a WD40 domain at the C-terminal end. Although these proteins differ in the C-terminal region, they were found to be in the same family due to the presence of a conserved F-box domain (see Figure 6).

The complete protein sequence alignment of this family is 408 residues in length, and we inferred 24 sites with positive selection at the 95% confidence level, all of which occurred within the latter half of this alignment. Adaptive evolution does not appear to target the N-terminal F-box domain, but appears to be concentrated on the C-terminal domain. Analysis of the positive selection found within the SBD of the PDB structure 2e32, showed selection occurring in proximity to the substrate binding region, although the residues involved in substrate binding where not themselves under selection .

The PDB structure 2ovr is the Fbw7-Skp1-CyclinE Complex . This protein complex is part of the CyclinE degradation pathway and is important in cell cycle regulation . The other PDB structure mapped, 1p22 is a β-TrCP1-Skp1-β-Catenin Complex, also important in cell cycle regulation . The WD40 domain contains sites of positive selection (see Figure 7). Although both 2ovr and 1p22 contain WD40 domains, they bind different substrates at this position, namely CyclinE and β-catenin respectively. Prediction of positive selection within this domain could indicate that proteins with multiple binding affinities may be under increased selective pressure. Mapping positive selection to the PDB structure 2ovr, we found three amino acid sites involved in protein binding that are predicted to be positively selected.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

DISCUSSION

Generating structure predictions for these proteomes allows us to delve further into the underlying occurrences of selection and selective pressure within plant proteins. Positive selection appears to be a non-random occurrence within proteins, occurring in cluster along the alignment length, which could be indicative of pressure on a particular fold or protein region to change. Our results demonstrate that positive selection also occurs more often within particular elements and areas of structural fold change within protein structures. In particular, the numbers of selected residues were less than expected in helical and strand elements. Interestingly, there were significantly more selected sites among residues that were associated with a change in protein structure (e.g. coil to strand or vice versa). Disordered residues also showed an increase in positively selected sites. It is known that many disordered regions become ordered upon binding , as well as having affinity to bind multiple proteins . Within our analysis of plant protein structures, we found, similar to work previously done on Drosophila , that certain secondary structure elements, plus disordered residues have more positive selection than others. Our results indicate that areas of possible evolutionary change, be it as a disordered region or a secondary structure region, may be under greater positive selective pressure than more structurally conserved, ordered regions of the protein.

Previous studies on positive selection within plant genomes have usually focused on single, or small number, of plant families . To our knowledge, the largest current study published on plant adaptive evolution is by Roth and Liberles who used The Adaptive Evolution Database (TAED) to predict positive selection in 4,216 seed plant gene families. They found 87 families

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

showed positive selection; however most of these families contained only 2 proteins. We have analysed over 2,000 gene families with at least 10 sequences per family for positive selection, one of the largest plant protein family analyses to date, and found selection in 175 families.

In agreement with previous work , the amount of positive selection we found within plant species appears to be low. Eight percent of our plant protein families appear to be undergoing selection, which is much smaller than estimates within non-plant species . This could be indicative of the effective population size of plants being smaller, relative to other species .

Mapping of positively selected sites to our known structures suggest that amino acid sites which show evidence for positive selection cluster within a protein sequence, a result also found in a previous analysis on Drosophila (Ridout *et al.* 2010). Ridout *et al.* (2010) however, found that N-, C- terminal regions appear to contain more positively selected sites, something we do not find in our plant analysis, where selection appears to be spread through out the protein. Due to the high amount of divergence between our original sequences, we culled our sequences by removing extraneous gapped regions. Most of these gapped deletions where in N-, C- terminals, giving us a possible sampling bias, which may have removed sites that could be undergoing selection and therefore missing additional selected regions, which could have mirrored the results found in Drosophila.

We also find that sites undergoing positive selection appear to be less buried than wholly conserved sites within protein structures. Previous work by Liu et al. , Roth and Liberles and Petersen et al. , working with human SNP data, seed plants and *E. coli* respectively, all found similar results in their data, suggesting that less buried residues are under less selective pressure in multiple species.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Linking protein structure, function and evolution has been one of the key goals of molecular evolution. The availability of whole genome sequences has allowed investigators access to an inventory of all proteins in an organism, and the availability of data across multiple species allows for a comparative analysis in a phylogenetically informed approach. We have used several tools for protein structure domain identification and prediction that we had previously applied to multiple proteomes , and have now developed a similar database for plant proteins. As we describe in this study, this provides us with both general trends in the evolution of plant protein families, as well as allow us to highlight evolution of specific examples – in this case a C2H2 and an F-box family. As investigators exploit this database, we may be able to identify even more compelling trends in the structural evolution of plant proteins.

ACKNOWLEDGEMENTS

We would like to thank the volunteers participating in the Human Proteome Folding Project on IBM's World Community Grid and the FFAS03 development team, and the Rosetta Commons and the Yeast Resource Center. This work was supported in part by grant NSF DBI-0820757.

REFERENCES

Altschul, SF, W Gish, W Miller, EW Myers, DJ Lipman.1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Altschul, SF and Koonin, EV 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem. Sci. 23: 444-447.

SMBE

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Alvarez-Ponce, D, Aguade, M and Rozas, J 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. Genome Biol. Evol. 3: 87-101.

Angiosperm Phylogeny Group 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical J. Linnean Soc. 141: 399-436.

Baliga, NS, SJ Bjork, R Bonneau, M Pan, C Iloanusi, MC Kottemann, L Hood, J DiRuggiero 2004. Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1. Genome Res. 14: 1025-1035.

Bharathan, G, Janssen, BJ, Kellogg, EA and Sinha, N 1999. Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. Mol. Biol. Evol. 16: 553-563.

Bishop, JG, Dean, AM and Mitchell-Olds, T 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc. Natl. Acad. Sci. USA 97: 5322-5327.

Bustamante, CD, R Nielsen, SA Sawyer, KM Olsen, MD Purugganan, DL Hartl. 2002. The cost of inbreeding in Arabidopsis. Nature 416: 531-534.

Chevalier, BS, T Kortemme, MS Chadsey, D Baker, RJ Monnat, BL Stoddard. 2002. Design, activity, and structure of a highly specific artificial endonuclease. Mol. Cell 10: 895-905.

Chiu, JC, EK Lee, MG Egan, IN Sarkar, GM Coruzzi, R DeSalle. 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics 22: 699-707.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Chivian, D, Robertson, T, Bonneau, R and Baker, D 2003. Ab initio methods. Methods Biochem. Anal. 44: 547-557.

Conery, JS and Lynch, M 2001. Nucleotide substitutions and the evolution of duplicate genes. Pacific Symposium on Biocomputing: 167-178.

Conrad, B and Antonarakis, SE 2007. Gene duplication: a drive for phenotypic diversity and cause of human disease. Ann. Rev. Genomics Hum. Genet. 8: 17-35.

Das, AK, Cohen, PW and Barford, D 1998. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. EMBO J. 17: 1192-1199.

Doolittle, RF 1995. The origins and evolution of eukaryotic proteins. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 349: 235-240.

Drew K, P Winters, GL Butterfoss, V Berstis, K Uplinger, J Armstrong, M Riffle, E Schweighofer, B Bovermann, DR Goodlett, TN Davis, D Shasha, L Malmström, R Bonneau 2011. The Proteome Folding Project: proteome-scale prediction of structure and function. Genome Res. 21: 1981-1994.

Dunker, AK, MS Cortese, P Romero, LM Iakoucheva, VN Uversky. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS j. 272: 5129-5148.

Eisenberg, D, Marcotte, EM, Xenarios, I and Yeates, TO 2000. Protein function in the post-genomic era. Nature 405: 823-826.

Englbrecht, CC, Schoof, H and Bohm, S 2004. Conservation, diversification and expansion of C2H2 zinc finger proteins in the Arabidopsis thaliana genome. BMC Genomics 5: 39.

Gossmann, TI, BH Song, AJ Windsor, T Mitchell-Olds, CJ Dixon, MV Kapralov, DA Filatov, A Eyre-Walker. 2010. Genome wide analyses reveal little evidence for adaptive

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

evolution in many plant species. Mol. Biol. Evol. 27: 1822-1832.

Gray, JJ, S Moughon, C Wang, O Schueler-Furman, B Kuhlman, CA Rohl, D Baker. 2003a. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. 331: 281-299.

Gray, JJ, SE Moughon, T Kortemme, O Schueler-Furman, KM Misura, AV Morozov, D Baker. 2003b. Protein-protein docking predictions for the CAPRI experiment. Proteins 52: 118-122.

Hahn, MW, Han, MV and Han, SG 2007. Gene family evolution across 12 Drosophila genomes. PLoS Genet. 3: e197.

Halligan, DL, F Oliver, A Eyre-Walker, B Harr, PD Keightley. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6: e1000825.

Hao, B, S Oehlmann, ME Sowa, JW Harper, NP Pavletich. 2007. Structure of a Fbw7-Skp1-cyclin E complex: multisite-phosphorylated substrate recognition by SCF ubiquitin ligases. Mol. Cell 26: 131-143.

Hernandez-Hernandez, T, Martinez-Castilla, LP and Alvarez-Buylla, ER 2007. Functional diversification of B MADS-box homeotic regulators of flower development: Adaptive evolution in protein-protein interaction domains after major gene duplication events. Mol. Biol. Evol. 24: 465-481.

Jaroszewski, L, Rychlewski, L and Godzik, A 2000. Improving the quality of twilight-zone alignments. Protein Sci. 9: 1487-1496.

Jaroszewski, L, L Rychlewski, Z Li, W Li, A Godzik. 2005. FFAS03: a server for profile--profile sequence alignments. Nuc. Acids Res. 33: W284-288.

Jiao Y, NJ Wickett, S Ayyampalayam et al. 2011. Ancestral polyploidy in seed plants and

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

angiosperms. Nature 473: 97-100.

Jin, J, T Cardozo, RC Lovering, SJ Elledge, M Pagano, JW Harper. 2004. Systematic analysis and nomenclature of mammalian F-box proteins. Genes Dev. 18: 2573-2580.

Jordan, GE and Piel, WH 2008. PhyloWidget: web-based visualizations for the tree of life. Bioinformatics 24: 1641-1642.

Kabsch, W and Sander, C 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.

Kajander, T, Cortajarena, AL, Mochrie, S and Regan, L 2007. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. Acta Crystallogr. D Biol. Crystallogr. 63: 800-811.

Kapralov, MV, Kubien, DS, Andersson, I and Filatov, DA 2011. Changes in Rubisco kinetics during the evolution of C4 photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme. Mol. Biol. Evol. 28: 1491-1503.

Kelleher, ES, Swanson, WJ and Markow, TA 2007. Gene duplication and adaptive evolution of digestive proteases in Drosophila arizonae female reproductive tracts. PLoS Genet. 3: e148.

Kim, DE, Chivian, D, Malmstrom, L and Baker, D 2005. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61 Suppl 7: 193-200.

Kinsella, RJ, DA Fitzpatrick, CJ Creevey, JO McInerney. 2003. Fatty acid biosynthesis in Mycobacterium tuberculosis: lateral gene transfer, adaptive evolution, and gene

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

duplication. Proc. Natl. Acad. Sci. USA. 100:10320-10325.

Kipreos, ET and Pagano, M 2000. The F-box protein family. Genome Biol. 1: reviews3002.1-3002.7.

Kramer, EM, Jaramillo, MA and Di Stilio, VS 2004. Patterns of gene duplication and functional evolution during the diversification of the AGAMOUS subfamily of MADS box genes in angiosperms. Genetics 166: 1011-1023.

Kuhlman, B and Baker, D 2000. Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA. 97: 10383-10388.

Kuhlman, B, G Dantas, GC Ireton, G Varani, BL Stoddard, D Baker. 2003. Design of a novel globular protein fold with atomic-level accuracy. Science 302: 1364-1368.

Li, W, B Liu, L Yu, D Feng, H Wang, J Wang. 2009. Phylogenetic analysis, structural evolution and functional divergence of the 12-oxo-phytodienoate acid reductase gene family in plants. BMC Evol. Biol. 9: 90.

Lise, S and Jones, DT 2005. Sequence patterns associated with disordered regions in proteins. Proteins 58: 144-150.

Liu, J, Zhang, Y, Lei, X and Zhang, Z 2008. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol. 9: R69.

Lynch, M and Conery, JS 2003. The evolutionary demography of duplicate genes. J. Struc. Func. Genomics 3: 35-44.

Main, ER, Y Xiong, MJ Cocco, L D'Andrea, L Regan. 2003. Design of stable alpha-helical arrays from an idealized TPR motif. Structure 11: 497-508.

Martinez-Castilla, LP and Alvarez-Buylla, ER 2003. Adaptive evolution in the Arabidopsis

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

MADS-box gene family inferred from its complete resolved phylogeny. Proc. Natl. Acad. Sci. USA. 100: 13407-13412.

McGuffin, LJ, Bryson, K and Jones, DT 2000. The PSIPRED protein structure prediction server. Bioinformatics 16: 404-405.

Mizushima, T, T Hirao, Y Yoshida, SJ Lee, T Chiba, K Iwai, Y Yamaguchi, K Kato, T Tsukihara, K Tanaka. 2004. Structural basis of sugar-recognizing ubiquitin ligase. Nat. Struct. Mol. Biol. 11: 365-370.

Mizushima, T, Y Yoshida, T Kumanomidou, Y Hasegawa, A Suzuki, T Yamane, K Tanaka. 2007. Structural basis for the selection of glycosylated substrates by SCF(Fbs1) ubiquitin ligase. Proc. Natl. Acad. Sci. USA. 104: 5777-5781.

Mondragon-Palomino, M and Gaut, BS 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in Arabidopsis thaliana. Mol. Biol. Evol. 22: 2444-2456.

Mondragon-Palomino, M, Meyers, BC, Michelmore, RW and Gaut, BS 2002. Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res. 12: 1305-1315.

Moury, B and Simon, V 2011. dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of Potato virus Y. Mol. Biol. Evol. 28: 2707 - 2717

Murzin, AG, Brenner, SE, Hubbard, T and Chothia, C 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247: 536-540.

Nam, J, J Kim, S Lee, G An, H Ma, M Nei. 2004. Type I MADS-box genes have experienced

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc. Natl. Acad. Sci. USA. 101: 1910-1915.

Palme, AE, Pyhajarvi, T, Wachowiak, W and Savolainen, O 2009. Selection on nuclear genes in a Pinus phylogeny. Mol. Biol. Evol. 26: 893-905.

Petersen, L, JP Bollback, M Dimmic, M Hubisz, R Nielsen. 2007. Genes under positive selection in Escherichia coli. Genome Res. 17: 1336-1343.

Petroski, MD and Deshaies, RJ 2005. Function and regulation of cullin-RING ubiquitin ligases. Nat. Rev. Mol. Cell Biol. 6: 9-20.

Purugganan, MD 1997. The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. J. Mol. Evol. 45: 392-396.

Purugganan, MD, Rounsley, SD, Schmidt, RJ and Yanofsky, MF 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. Genetics 140: 345-356.

Ridout, KE, Dixon, CJ and Filatov, DA 2010. Positive selection differs between protein secondary structure elements in Drosophila. Genome Biol. Evol. 2: 166-179.

Rohl, CA 2005. Protein structure estimation from minimal restraints using Rosetta. Methods Enzymol. 394: 244-260.

Rohl, CA, Strauss, CE, Chivian, D and Baker, D 2004. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 55: 656-677.

Rosinski, JA and Atchley, WR 1999. Molecular evolution of helix-turn-helix proteins. J. Mol. Evol. 49: 301-309.

Roth, C and Liberles, DA 2006. A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol. 6: 12.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Rychlewski, L, Jaroszewski, L, Li, W and Godzik, A 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci. 9: 232-241.

Sikorski, RS, Boguski, MS, Goebl, M and Hieter, P 1990. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. Cell 60: 307-317.

Smith, NG and Eyre-Walker, A 2002. Adaptive protein evolution in Drosophila. Nature 415: 1022-1024.

Sterck, L, S Rombauts, K Vandepoele, P Rouze, Y Van de Peer. 2007. How many genes are there in plants (... and why are they there)? Curr. Opin. Plant Biol. 10: 199-203.

Strasburg, JL, NC Kane, AR Raduski, A Bonin, R Michelmore, LH Rieseberg. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol. Biol. Evol. 28: 1569-1580.

Swaffield, JC and Purugganan, MD 1997. The evolution of the conserved ATPase domain (CAD): reconstructing the history of an ancient protein module. J. Mol. Evol. 45: 549-563.

Wagstaff, BJ and Begun, DJ 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert Drosophila. Genetics 177: 1023-1030.

Ward, JJ, LJ McGuffin, K Bryson, BF Buxton, DT Jones. 2004. The DISOPRED server for the prediction of protein disorder. Bioinformatics 20: 2138-2139.

Waterhouse, AM, JB Procter, DM Martin, M Clamp, GJ Barton. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Weston, AD, Baliga, NS, Bonneau, R and Hood, L 2004. Systems approaches applied to the sudy of Saccharomyces cerevisiae and Halobacterium sp. In: *Cold Spring Harb. Symp. Quant. Biol.*, ed. S.D. Stillman, Cold Spring Harb. Symp. Quant. Biol.: Cold Spring Harbor Laboratory Press.

Wilson, CG, Kajander, T and Regan, L 2005. The crystal structure of NlpI. A prokaryotic tetratricopeptide repeat protein with a globular fold. FEBS J. 272: 166-179.

Wright, PE and Dyson, HJ 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293: 321-331.

Wu, G, G Xu, BA Schulman, PD Jeffrey, JW Harper, NP Pavletich. 2003. Structure of a beta-TrCP1-Skp1-beta-catenin complex: destruction motif binding and lysine specificity of the SCF(beta-TrCP1) ubiquitin ligase. Mol. Cell 11: 1445-1456.

Xiao, W, J Zhao, S Fan, L Li, J Dai, M Xu. 2007. Mapping of genome-wide resistance gene analogs (RGAs) in maize (Zea mays L.). Theor. Appl. Genet. 115:501-508.

Xu, G, Ma, H, Nei, M and Kong, H 2009. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc. Natl. Acad. Sci. USA. 106: 835-840.

Yang, J, SM Roe, MJ Cliff, MA Williams, JE Ladbury, PT Cohen, D Barford. 2005. Molecular basis for TPR domain-mediated regulation of protein phosphatase 5. EMBO J. 24: 1-10.

Yang, Z 2006. *Computational Molecular Evolution*. Oxford: Oxford University Press.

Yang, Z 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586-1591.

Yang, Z, Nielsen, R, Goldman, N and Pedersen, AM 2000. Codon-substitution models for

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

Yu, J, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296: 79-92.

Yu, J, et al. 2005. The Genomes of Oryza sativa: a history of duplications. PLoS Biol. 3: e38.

Zhang, L, Pond, SK and Gaut, BS 2001. A survey of the molecular evolutionary dynamics of twenty-five multigene families from four grass taxa. J. Mol. Evol. 52: 144-156.

Zhang, W and Koepp, DM 2006. Fbw7 isoform interaction contributes to cyclin E proteolysis. Mol. Cancer Res. 4: 935-943.

Zhao, JP and Su, XH 2010. Patterns of molecular evolution and predicted function in thaumatin-like proteins of Populus trichocarpa. Planta 232: 949-962.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

FIGURE LEGENDS

Figure 1. Size distribution of plant protein families based on initial OrthologID analyses.

Figure 2. Sample structure output in the Plant PFP. This example is of a MADS-box domain interacting with DNA, as depicted in PDB. The highlighted regions in blue spheres show regions of predicted positive selection. The gray indicates regions in the PDB structure that were not predicted to structurally map to the plant MADS-box domain or an alternate PDB chain.

Figure 3. Distribution of positively selected residue positions in protein structure elements. (A) Distribution according to secondary structural elements. (B) Distribution of positively selected residues in ordered and disordered protein regions. The gray bars are the observed and the black bars are expected numbers.

Figure 4. Distribution of RASA values across amino acid sites in protein structures. The distributions are shown for all sites, conserved and positively selected amino acid sites. Although the distributions are similar, a t-test showed significant differences in RASA values between conserved and positively selected amino acid sites

Figure 5. PDB structure 1na0:B, an idealized TRP domain, is one of the domains mapped to the N-terminal of this OID family. Highlighted regions in blue are sites of

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

positive selection with at least 95% confidence of prediction support. In this example, the gray regions were not predicted by our analysis to map structurally to our protein.

Figure 6. Sugar binding domain (SBD) of F-box protein 2e32. Areas highlighted in blue are positively selected sites with high confidence. In red, are those residues indicated to be involved in substrate binding . The gray regions were not predicted by our analysis to map structurally to our protein. Not included here is the F-box domain, which did not have sites inferred to have evolved with positive selection.

Figure 7. PDB structures of the C-terminal WD40 domain in the F-box protein. Residues highlighted in blue are positively selected residues. Residues highlighted in red are substrate-binding residues (Wao et al. 2003, Hao et al. 2007). The gray regions were not predicted by our analysis to map structurally to our protein. (A) The PDB structure of 1p22. (B) The WD40 C-terminal domain of the PDB structure 2ovr. Positively selected residues appear in close proximity to substrate binding residues. Highlighted in black are three residues that show evidence for positive selection and are also involved in substrate binding.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Genome Biology and Evolution

Table 1. The number of proteins and predicted protein domains for each organism within the Plant PFP database.

Organism	Number of Proteins	Number of Domains	Domains with known PDB structures[a]	Domains without known PDB structures[b]
A. thaliana	36,350	63,748	31,200	32,548
O. sativa	67,393	150,986	57,828	93,158
P. trichocarpa	43,000	71,171	32,443	38,728
S. bicolour	36,410	68,644	27,722	40,922
V. vinifera	30,434	54,468	24,627	29,841

[a]based on psi-blast, fold recognition methods; [b]domain structures based on pfam, msa, and heuristic methods.

SMBE

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Genome Biology and Evolution

Table 2. *De novo* structure predictions for domain that did not map to known PDB structures.

Organism	**All *de novo* domain predictions**	***de novo* domain predictions (>0.8 confidence)**
A. thaliana	9631	1618
O. sativa	19541	3146
P. trichocarpa	14	1
S. bicolour	16	4
V. vinifera	0	0

SMBE

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Table 3. Each amino acid in an alignment was categorised into 8 amino acid properties.

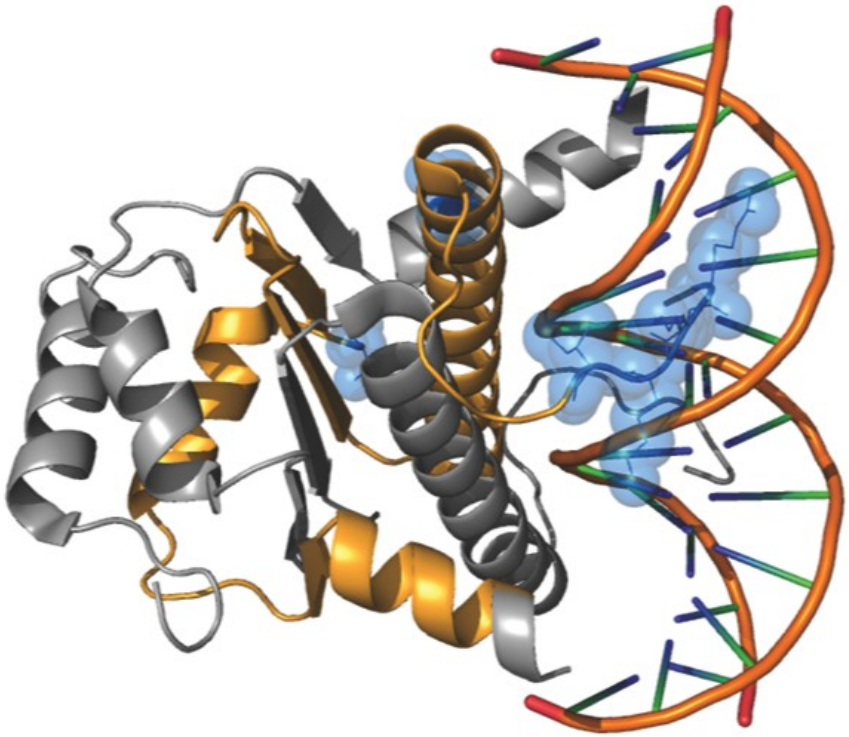
Properties	Observed	Expected	P-value[a]
Hydrophobic	12671	14589	< 0.00001
Polar	13021	10774	< 0.00001
Small	11231	10795	< 0.00001
Aliphatic	3175	4760	< 0.00001
Aromatic	2449	2459	< 0.8384
Positive	3695	3053	< 0.00001
Negative	2683	2511	< 0.0002237
Unique	2513	2489	< 0.5602

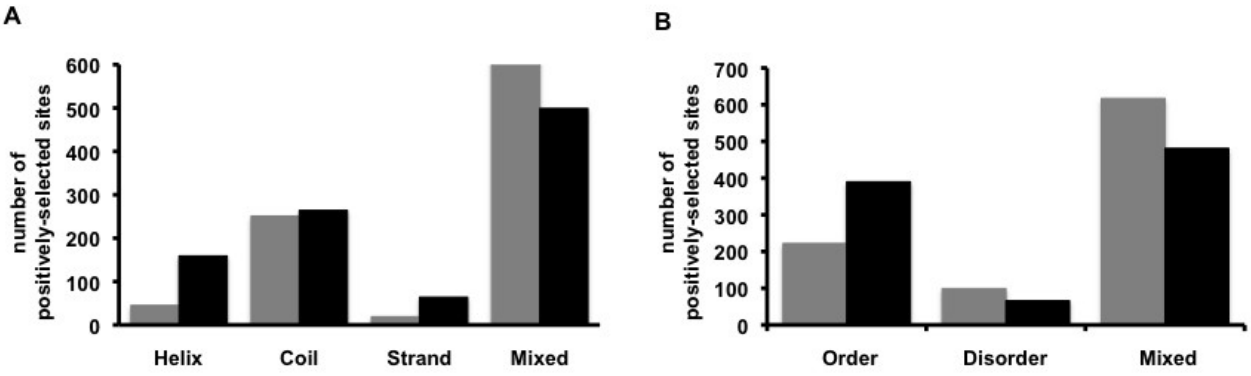
[a]Significance of positive correlation between number of positively selected sites and polarity, size and positive charge while hydrophobic and aliphatic residues showed a strong negative correlation with positive selection.

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

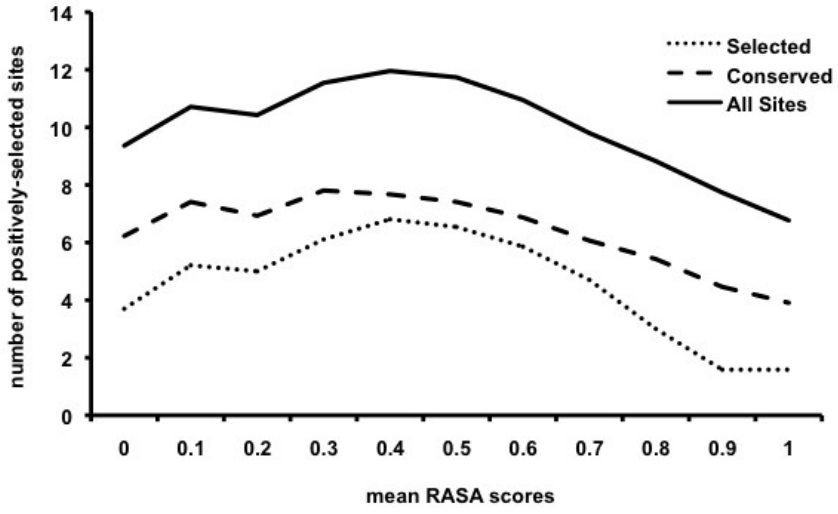
Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012



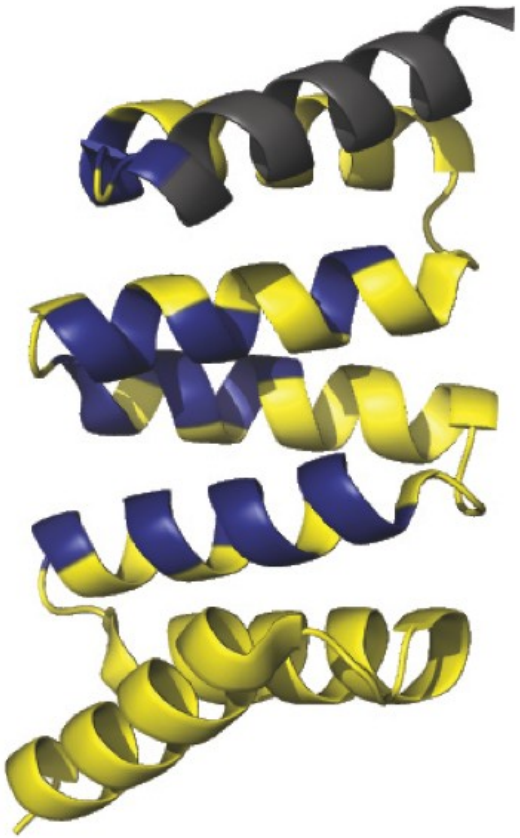


Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012



Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012



Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012

Downloaded from http://gbe.oxfordjournals.org/ at New York University on March 2, 2012