

A bioinformatics-based approach to the identification, classification, and analysis of genes encoding plant cell wall hydroxyproline-rich glycoproteins (HRGPs)

Allan M. Showalter¹, Jason Yerardi², Tom Conley², Brian Keppler¹, and Lonnie R. Welch²

¹Department of Environmental and Plant Biology, Molecular and Cellular Biology Program, Ohio University, Athens, OH 45701-2979

²School of Electrical Engineering & Computer Science, Center for Intelligent, Distributed and Dependable Systems, Ohio University, Athens, OH 45701-2979

Hydroxyproline-rich glycoproteins (**HRGPs**) are a superfamily of plant cell wall proteins involved with diverse aspects of plant growth and development. The HRGP superfamily consists of three subfamilies: the hyperglycosylated arabinogalactan-proteins (**AGPs**), the moderately glycosylated extensins (**EXTs**), and the lightly glycosylated proline-rich proteins (**PRPs**). In order to “mine” genomic databases for these HRGP family members and guide future research directions, a “BioOhio” software program was developed that identifies and sub-classifies AGPs, EXTs, and PRPs from proteins predicted from DNA sequence data. This bioinformatics program is based in part on searching for biased amino acid compositions (e.g., 50% of an AGP’s amino acid composition consists of proline (P), alanine (A), serine (S), and threonine (T)) and in part on searching for particular protein motifs associated with known HRGPs (e.g., the pentapeptide sequence S-P-P-P-P is characteristic of EXTs). In addition, potential signal peptide sequences (which are generally associated with HRGPs and the secretory pathway) and glycosylphosphatidylinositol (GPI) lipid anchors (which are frequently associated with AGPs and responsible for their transient plasma membrane localization) are identified by the program. HRGPs identified by the program are subsequently analyzed: 1) to reveal any novel repeating amino acid sequences, 2) to create protein phylogenies, 3) to elucidate expression patterns of their genes using public databases, and 4) to determine whether genetic mutants are available for their corresponding genes. To date, the BioOhio software was used to identify and classify various AGPs (e.g., classical AGPs, lysine-rich AGPs, AG peptides, and fasciclin-like AGPs), EXTs (e.g., EXTs characterized by various repeating amino acid sequences, leucine-rich repeat-EXTs, EXT kinases, and EXT peptides), and PRPs (sub-classification work in progress) from protein databases derived from the genomic databases of two completely sequenced model plant species, *Arabidopsis thaliana* (a member of the mustard family) and rice.