Tracing the molecular and evolutionary determinants of novel functions in protein families

by

Andrew Charles Doxey

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis explores the limits of homology-based inference of protein function and evolution, where overall similarity between sequences can be a poor indicator of functional similarity or evolutionary relationships. Each case presented has undergone different patterns of evolutionary change due to differing selective pressures. Surface adaptations and regulatory (e.g., gene expression) divergence are examined as molecular determinants of novel functions whose patterns are easily missed by assessments of overall sequence similarity. Following this, internal repeats and mosaic sequences are investigated as cases in which key evolutionary events involving fragments of protein sequences are masked by overall comparison. Lastly, virulence factors, which cannot be unified based on sequence, are predicted by analysis of elevated host-mimicry patterns in pathogenic versus non-pathogenic bacterial genomes. These patterns have resulted from unique co-evolutionary pressures that apply to bacterial pathogens, but may be lacking in their close relatives. A recurring theme in the proteins/genes/genomes analyzed is an involvement in microbial pathogenesis or pathogen-defense. Due to the ongoing "evolutionary arms race" between hosts and pathogens, virulence and defense proteins have undergone—and will likely continue to generate—evolutionary novelties. Thus, they demonstrate the necessity to look beyond overall sequence comparison, and assess multiple dimensions of functional innovation in proteins.

Acknowledgements

Numerous individuals have contributed directly and indirectly to the work presented in this thesis. First, I would like to thank my supervisor, Brendan McConkey, who has provided the academic freedom and expertise that has allowed me to explore numerous disciplines. I would also like to express a great deal of thanks to Barb Moffatt, Elizabeth Meiering, Mike Lynch, Kirsten Müller, Marilyn Griffith, Bernard Duncker, Trevor Charles, Gabriel Renaud, Gerald Quon, Mahmoud Yaish, Owen Woody, Aron Broom, Tony Facciuolo, Zhenyu Cheng, Brian Ingalls, Josh Neufeld, Gabriel Moreno-Hagelsieb, and Nicholas Provart, for collaborations, scientific discussions, support, and/or involvement in my thesis committee.

I am also grateful to my funding sources, NSERC and OGS, for supporting my work throughout all of graduate school.

I cannot express enough thanks to my wife, Juliana, for her patience, understanding, and support. Lastly, I thank my family for their support and guidance, and my father for providing the intellectual spark that began my interest in science and many other things.

Contents

List of Tables

List of Figures

List of Abbreviations

β-1.3-G - β-1,3-glucanase

AFP - Antifreeze protein

BoNT - Botulinum neurotoxin

CBM - Carbohydrate-binding module

CNT - Clostridial neurotoxin

ECM - Extracellular matrix

GBP - Glycan-binding protein

GO - Gene ontology

GPI - Glycosylphosphatidylinositol

HA - Hemagglutinin

HCRc - BoNT C-terminal receptor binding domain

HCRn - BoNT N-terminal receptor binding domain

HMM - Hidden Markov Model

HVR - Hypervariable region

LDA - Linear discriminant analysis

LTP - Lipid transfer protein

ML - Maximum-likelihood

NJ - Neighbor-joining

NCBI - National Center for Biotechnology Information

NMR - Nuclear Magnetic Resonance

NTNH - Non-toxin non-hemagglutinin

MS - Mass spectrometry

OSC - Ordered surface carbon

PCA - Principal components analysis

PDB - Protein Data Bank

PPOD - Putative peroxidase

PR - Pathogenesis-related

SAS - Solvent accessible surface area

VFDB - Virulence factors database

Chapter 1

Introduction

In the post-genomic era of biology[1], we are faced with an astounding and continually increasing amount of sequence data. It is the goal of bioinformatics and computational biology to make sense of this data, and place it in a meaningful biological context. This ultimate goal, which falls under the label of *functional genomics*, is to ascribe function to genomes, genes, proteins, promoters, regulatory motifs, epigenetic traits, and ultimately, every functional base pair of DNA.

It has been suggested that modelers can be divided into two types of thinking: there are *lumpers* who tend to recognize similarities between things and thus prefer grouping them together; and *splitters* who tend to recognize differences, and thus prefer to split things into more narrowly defined subclasses to avoid model overgeneralization[2]. The lumper versus splitter dichotomy provides a useful perspective for assessing the major developments in bioinformatics and computational biology over the last decade.

It is not surprising that the lumper type of thinking is and has been predominant in bioinformatics. Given the vast apparent complexity of sequence data and the enormous challenge of functional annotation, it is natural to want to organize this complexity by grouping together related sequences, and categorize genes into a neatly defined set of functional groups. Indeed, the most profound developments in bioinformatics thus far have been from this perspective.

[1] The post-genomic era of biology concerns the tasks of analysis and interpretation of large-scale genomic sequence data.

[2] See Endersby et al. (2009) for a description of these terms as they apply to taxonomy.

Arguably, the most important and widely used tool in bioinformatics is the BLAST search (Altschul et al., 1990). BLAST and other sequence similarity search algorithms provide a means to detect homologous[3] relationships between sequences. Grouping together evolutionarily related sequences is an extremely powerful approach for reducing the complexity of functional genomics because "related sequences often have related functions" (Thomas et al., 2003). Altschul et al.'s (1990) paper on BLAST has since been cited over 30,000 times[4], and has become the standard approach among bioinformaticians and biologists for answering the frequently recurring questions, "what is my sequence related to?" and, "what is the function of my sequence?" (Stevens et al., 2001).

Since the original BLAST algorithm, increasingly sensitive methods have been developed to detect remote homologies between sequences and sequence families [e.g., PSI-BLAST (Altschul et al., 1997), HMMER (Eddy, 1998)]. These tools, paired with large online databases [e.g., PFAM (Finn et al., 2006), Genbank], have resulted in large-scale classifications of proteins into a surprisingly small number of groups. Proteins can be grouped into protein families or larger superfamilies which share a common three-dimensional structure (fold). Using available completed genome sequence data, Wolf et al. (2000) have estimated that all proteins can be classified into as few as 1,000 folds and 5,000 families. As protein families are derived from a common ancestor, the functions of their derived members are often thought to be modifications of an ancestral function. At its extreme, this argument leads to the presupposition of a small set of ancient folds and functions from which the modern protein universe evolved.

There is, however, a potential danger of oversimplification in homology-based annotation and classification of proteins. As stated by Sjölander (2004), "evolution not only conserves function, it also generates new functions". Annotating a common function for a family of proteins may work well when dealing with orthologs[5] that have not diverged appreciably in function. However, functional divergence can occur despite there being detectable, significant sequence similarity. The long term survival of paralogs[6], for instance, is thought to depend largely on whether new beneficial functions can develop before they are silenced by degenerative mutations (Ohno, 1970). In this model of functional divergence,

[3]Homologous sequences share similarities due to common ancestry.

[4]Citation number retrieved using Google Scholar on 03/01/2010.

[5]Orthologs are homologs separated by speciation.

[6]Paralogs are homologs separated by within-genome duplication.

called *neofunctionalization*, one copy retains its ancestral function where the other duplicate adopts a novel function[7]. More recent models have proposed that *subfunctionalization* occurs, whereby the ancestral gene function is split among the daughter copies (Hughes, 1994; Lynch and Force, 2000). Regardless of the underlying evolutionary forces, *functional divergence between duplicate genes* is a major driving force of evolutionary innovation, and it is neglected by high-throughput homology-based inference of function (Sjölander, 2004).

Ultimately, while grouping of related sequences and examining patterns of homology and conservation is useful, it does not provide a full picture of how novel functions originate and diversify. This thesis explores the limits of homology-based functional annotation of proteins, and examines different ways in which overall sequence similarity does not fully capture function and, in some cases, evolutionary relationship. Each case (i.e., proteins, genes, families, functional groups, genomes) examined presents unique problems to homology-based annotation because various evolutionary and functional innovations/changes have occurred. These evolutionary innovations (surface adaptations, expression divergence, internal repeats, recombination, etc.) complicate the basic assumption that similar sequences are directly related and thus have related functions. Thus, bioinformatic characterization of function or evolutionary history for these cases requires a shift from analysis of conservation patterns to the analysis of the underlying determinants of functional novelty and change. This involves analysis of various features (e.g., surface patterns, motifs, gene expression traits) that provide deeper clues into the origins of novel protein functions and proteins themselves. Pinpointing the key (though sometimes subtle) functional/evolutionary changes between homologous sequences and similarities between unrelated sequences will be critical for proper annotation of gene/protein function in the post-genomic era.

A recurring theme among the cases explored in this thesis is an involvement in microbial pathogenicity and pathogen defense. Virulence factors and pathogen-defense proteins are ideal cases for exploring the limits of homology-based annotation because their existence is largely due to the ongoing battle between pathogens and hosts, which requires the continual development of functional novelties. This idea is related to van Valen's original "Red Queen Hypothesis", which states "for an evolutionary system, continuing development is needed

[7]Ohno (1970) suggested that one duplicate partner retains the ancestral function, where the other duplicate gene loses selective constraint and becomes a pseudogene. Further mutations can cause the pseudogene to adopt a novel function, which may be subsequently maintained by purifying selection.

just in order to maintain its fitness relative to the systems it is co-evolving with" (van Valen, 1973). A dominant theme that arises from analysis of these cases is that virtually all aspects of protein structure are evolutionarily plastic, and are used by evolution to generate new proteins and new functions.

1.1 Homology-based inference of function

Here, I will provide a brief introduction to various methods and major concepts used in the area of protein classification and functional annotation. Most of these have an influence or are used directly in later chapters. The fundamental concepts of homology-based annotation using sequence and structural similarity are presented, as well as alternative approaches that assess specific patterns of conservation (e.g., evolutionary trace) or change, or molecular determinants of function (e.g., linear motifs, 3D surface patterns, and gene expression data). It is largely these alternative approaches that are explored and built upon throughout this thesis as they provide deeper clues into protein function and evolution.

1.1.1 Sequence, homology, and function

Two commonly recurring questions concerning a sequence of interest are: "what is the function of the sequence?" and, "what is the sequence most closely related to?" The second question is often motivated by the first because it is generally assumed that closely related sequences have similar functions. The most widely used approaches for attempting to answer these questions are based on the principles of *sequence alignment*.

Sequence alignment, the foundation of sequence-based bioinformatics, involves computing an optimal linear arrangement of two or more sequences based on a scoring function[8]. In most cases, the desired output of a sequence alignment is the inferred evolutionary relationship between the two sequences, where each amino acid (or nucleotide in DNA) is aligned with the related residue in the related sequence. Gaps represent insertions or deletions that, assuming the alignment represents the true evolutionary relationship, have occurred in either sequence since their most recent common ancestor (MRCA).

As any two sequences can be aligned, statistical methods are required to determine if an alignment reflects homology. Homology can be detected by comparing the alignment score to that expected from a distribution of scores obtained by aligning random sequences, which is known to approximate an extreme value distribution.

In a BLAST database search of a query sequence, the expected distribution is used to compute an *E-value* for each hit, which is the number of hits expected by chance in a

[8] A scoring function includes penalties for gaps and scores for each amino acid aligning with each other amino acid.

database of the given size. Identification of related sequences through BLAST E-values is fundamental to many modern bioinformatics approaches and tools.

If an alignment score is not significantly greater than that expected by chance, however, this does not necessarily mean that the two sequences are unrelated. There is always a possibility that the two sequences have a common ancestor, but have diverged to a point where they no longer possess significant sequence similarity. On the other hand, because sequence space is so vast, when two sequences are determined to be significantly similar according to an alignment, this is generally considered sufficient evidence of homology.

An example pairwise alignment of a segment of the flagellin protein from *Bacillus subtilis* (query) and a homologous segment from *Clostridium botulinum* (subject), produced via a BLAST search, is shown below:

```
Score = 36.2 bits (82), Expect = 1.1, Method:  Compositional matrix adjust.
Identities = 20/43 (46%), Positives = 27/43 (62%), Gaps = 3/43 (6%)

Query  1    TEFNTKKLLDGTAQN---LTFQIGANEGQTMSLSINKMDSESL  40
            TEFNT KLL+  A +   +  QIGANEGQ   + +  M+S +L
Sbjct  128  TEFNTIKLLNANAVDAGEVKLQIGANEGQFFGIKLQNMNSAAL  170
```

NOTE: The E-value of 1.1 indicates that 1.1 hits with the given alignment score (36.2 bits) are expected by chance in a database of the given size. While lower E-values (e.g., 0.01 and below) are typically used as cutoffs for determining homology, these two protein segments are nonetheless homologous, and their homology becomes more evident when larger segments of the protein sequence are aligned.

Pairwise alignments provide starting points for grouping together related sequences. A group of related proteins that are all derived from a common ancestor is a called a protein *family*, and members of the family can be aligned together in a multiple sequence alignment (MSA). Distantly related members of a protein family can be identified using profile-based methods such as PSI-BLAST, which uses a position-specific scoring matrix (PSSM), or HMMER, which forms a profile Hidden Markov Model (HMM) representation of a sequence family as a whole.

Profile representations of sequence families, unlike individual sequences, can model the

sequence variation within a family for each alignment position. Thus, they are extremely useful in bioinformatics and are the basis of major sequence databases. PFAM, for instance, is a database of protein families in which each protein family is represented by an HMM (an example MSA from PFAM is shown in Figure 1.1). Even more remote homologies between families can be inferred through profile-profile alignments [e.g., COMPASS (Sadreyev et al., 2007)] or through comparing HMMs [as in HHsearch (Söding, 2005)].

Figure 1.1: **Portion of a multiple sequence alignment from a protein family obtained from the PFAM database.**

1.1.2 Databases and homology-based annotation transfer

> "If there is a standard method for predicting protein function, it is the detection of similarity of amino-acid sequence by database searching, and assuming that the molecules identified are homologs with similar functions" (Whisstock and Lesk, 2003).

Transfer of known protein functions between identified homologous sequences has become standard practice in bioinformatics. A considerable number of online bioinformatics databases have automated this process, including PFAM, CDD (Marchler-Bauer et al., 2009), SMART (Schultz et al., 1998), InterPro (Apweiler et al., 2001), and many others. As proteins are composed of one or more *domains*[9], most databases annotate proteins according to domain composition, where individual domains belong to domain families (commonly represented by an HMM).

[9]Domains are modular, stable, folded structures within proteins.

Since phylogenetic trees increase the accuracy and detail regarding evolutionary relationships between sequences, mapping annotations onto subtrees, an approach known as *phylogenomics*[10] (Eisen, 1998), is also widely used. Combined with accurate experimental data as well as methods to distinguish orthologous from paralogous relationships, phylogenomic inference of protein function can be highly accurate despite being technically more demanding (Brown and Sjölander, 2006). While homology-based annotation is the simplest and most widely used approach for functional annotation, it has several major pitfalls (listed below). These also largely apply to phylogenomic methods of functional inference.

- Inaccuracy of annotations: The first problem concerns the inherent error in the experimental data itself (Valencia, 2005). Inaccuracies regarding protein functions described in the literature are frequently transferred to databases. Incorrect annotations can further propagate themselves from one database to another.
- Lack of experimental annotations: All computational annotations may be derived from an extremely small set of known protein sequences (perhaps less than 5%) (Valencia, 2005).
- It does not work for proteins without homologs of known function.
- It does not adequately account for the possibility of functional divergence between family members.
- Incorrect annotations based on protein modularity (Karp, 1998): Homology may be transferred between proteins with one or more—but not all—protein domains in common. This can be addressed by some phylogenomic methods.

[10]Phylogenomics involves selecting homologs, building a phylogenetic tree, mapping annotations onto the tree, distinguishing orthologs and paralogs, and inferring functions based on this information (Eisen, 1998).

1.1.3 Structural homology

The 3D structure of a protein may provide clues of homology in cases where the researcher is unable to detect homology at the sequence level (Watson et al., 2005). This works well because 3D structure is more conserved in the evolution of a protein family than is the sequence (Todd et al., 1999). This is due to the fact that sequence can change extensively while retaining key structural requirements of the protein fold[11]. Thus, organizing protein space by structure, as an alternative to sequence, can be useful for classifying distant evolutionary (or functional) relationships between proteins. The well known databases, SCOP and CATH, provide a hierarchical classification system for protein structural domains. In CATH, four levels (Class, Architecture, Topology, and Homologous superfamily) are used to classify protein domains. Structural alignment methods such as DALI (Holm and Sander, 1995) and VAST (Gibrat et al., 1996), can be used to recognize common folds for proteins that have diverged considerably in sequence.

Many homologous families, in which structure is conserved, possess a common and often exclusive function (Todd et al., 1999). In these cases, the function of a newly identified homologous family member is likely the same or very similar. Studies of sequence-to-function relationships in enzymes suggest that a percentage identity cutoff of ∼40% can be used to reliably predict the first three digits of an Enzyme Classification (E.C.) number, while the fourth digit varies within homologous families (Lee et al., 2007).

However, structure-based inference of function is complicated by cases in which protein function can converge and diverge in evolution. The TIM barrel fold, a so-called protein *superfold*[12], is found in proteins of diverse function [over 60 different E.C. numbers (Watson et al., 2005)] and thus highlights the limits of structure-based functional annotation. Convergence of function in unrelated folds is also well documented. For instance, the serine protease catalytic triad (Asp-His-Ser) has evolved independently in numerous protein

[11]Structures are very likely to be similar if their sequence identity is greater than 30-40% but structures can even retain similarity below 10% in the so-called "midnight zone" of homology (Rost, 1999). The midnight zone may also include structural "analogs" which have evolved similarities by convergent evolution.

[12]Superfolds are extremely abundant, recurring folds found in proteins that do not necessarily share sequence or functional similarity and may arise by convergent or parallel evolution (Orengo et al., 1994). Convergent evolution is the acquisition of similar traits in unrelated lineages. Parallel evolution is the independent acquisition of similar traits in related lineages.

lineages (Doolittle, 1994). Furthermore, as it is estimated that there are approximately 1,000 unique folds and 5000 protein families (Chothia, 1992; Wolf et al., 2000), a single fold-function or family-function model cannot possibly account for the functional diversity of proteins.

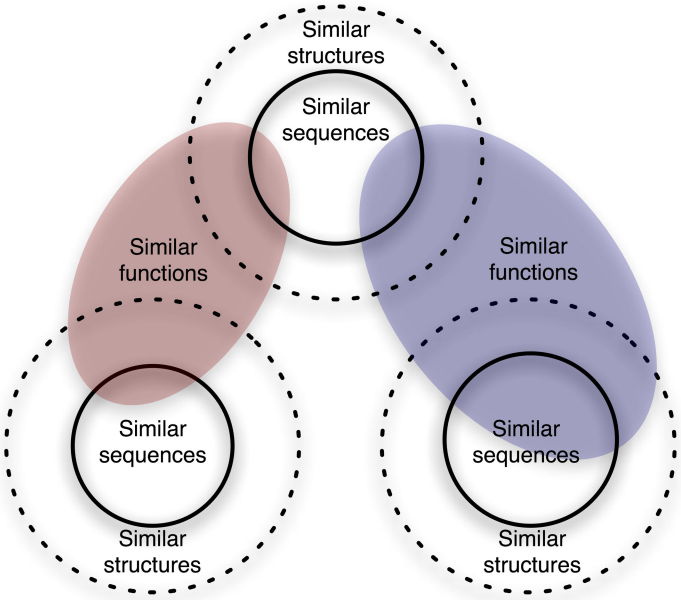


Figure 1.2: **An illustration of protein sequence, structural and functional space.** Unrelated structures and sequences can have similar functions. Related structures and sequences can have different functions. Adapted from Whisstock and Lesk (2003).

Ultimately, the relationship between sequence, structure, and function is highly complex and likely unique for each protein family. A general outline of the type of sequence/structure/function relationships that can occur in proteins is shown in Figure 1.2. As is depicted, similarity of sequence and structure does not correlate directly with similarity of function because, in addition to divergent evolution of function within a homologous protein family, similar functions can arise independently in different structures through convergent evolution. It has been suggested that overall structural similarity between proteins is ultimately no more reliable a predictor of functional similarity than sequence similarity (Gutteridge et al., 2003).

1.2 Modularity of protein function and evolution

The complex patterns by which function relates to the underlying amino acid acid sequence of a protein, and the complexity of function itself, explain why *overall sequence similarity* is by itself insufficient to sort out gene and protein functions in the post-genomic era. A step up in the sophistication of function-prediction and annotation tools is therefore achieved when the idea of sequence or structural similarity/homology is broken down and analyzed in more detail.

Function itself is a highly complex, multidimensional entity, and is difficult to define. A large-scale initiative to create a common vocabulary of function is the Gene Ontology (GO) consortium, which classifies function into three main hierarchical levels: cellular components, molecular function, and biological process (Ashburner et al., 2000). In reality, even GO terms do not capture all aspects of protein or gene function.

Furthermore, overall sequence similarity is not an adequate predictor of functional similarity because different aspects of function evolve at different rates in protein structures (Valencia, 2005). The catalytic site and structural core of a protein evolve very slowly due to enzymatic functional and structural constraints, respectively. Substrate pockets which affect ligand-specificity and regulatory regions on the protein surface likely evolve at faster rates. As function is tied to each of these properties, attempting to correlate a single entity called *function* to overall sequence or structural similarity is overly simplistic. That is, overall similarities may be maintained while individual aspects of function change, or overall similarities may be lost while individual aspects of function are maintained (but are no longer detectable through alignments).

Functionally constrained regions of proteins such as catalytic sites evolve slowly, and thus can often be identified as highly conserved positions within a multiple sequence alignment. This approach is commonly used to define *signatures* of key functional regions (e.g., a consensus sequence for an active site). The power of evolutionary conservation becomes even greater when structural information is incorporated, as demonstrated by the highly effective *Evolutionary Trace* method (Lichtarge et al., 1996). The Evolutionary Trace Method and related methods like ConSurf (Glaser et al., 2003) map the evolutionary conservation of each residue (as determined by the MSA) onto the structure or surface of a protein. Conserved residues can be clustered, or discriminated from those whose conservation is due to structural as opposed to functional constraints. As illustrated in Figure

1.3 for a β-1,3-glucanase family protein (examined in Chapter 3), conservation mapping is able to reveal the catalytic site and substrate pocket as the most evolutionarily constrained region of the protein.

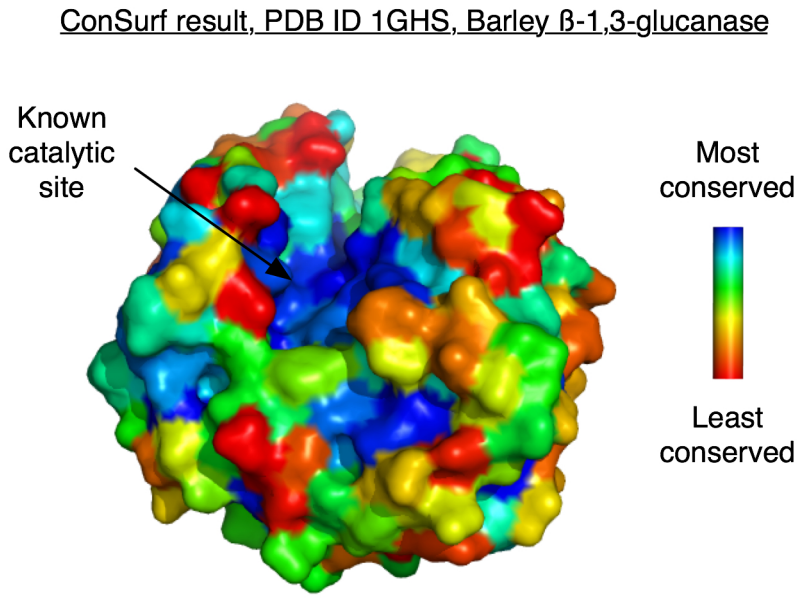


Figure 1.3: **Results of the ConSurf method applied to a barley β-1,3-glucanase (PDB ID 1GHS:A).** Mapping of evolutionary conservation onto protein structures reveals the known catalytic site.

Since both sequence and function generally diverge over time in the evolution of a protein family, conservation mapping can be expanded on by identifying different patterns of conservation in different subfamilies. This involves the identification of *tree determinant residues*, which are conserved within a subfamily but are different between subfamilies, and thus may play a role in the diversification of protein function. This general approach has been applied on a large scale as in the FunShift database (Abhiman and Sonnhammer, 2005). Statistical analyses of rate shifts within the framework of a phylogenetic tree can also be used to infer regions of functional divergence (Gu, 1999; Knudsen and Miyamoto, 2001), and mapped onto a protein structure (Gu and Vander Velden, 2002).

In addition to mapping amino acid conservation patterns onto a protein surface, related methods analyze the underlying selection pressures operating at the nucleotide level. Selection analysis, which forms the foundation of many studies of molecular evolution, typically involves analysis of non-synonymous/synonymous substitution rate ratios (K_a/K_s)

in protein-coding DNA sequences (Hurst, 2002). Two sequences that have diverged primarily due to selectively neutral changes will exhibit K_a/K_s ratios close to 1. K_a/K_s ratios significantly less than 1 indicate *purifying (negative) selection* which are under functional constraint, and K_a/K_s values significantly greater than 1 indicate *positive selection* and thus adaptive, function-altering changes. Functionally important residues can therefore be identified by site-specific selection analysis (Stern et al., 2007). The Selecton method (Stern et al., 2007) not only estimates site-specific selection but can take into account physicochemical properties of amino acids.

While sequence analyses of conservation and functional divergence are effective at inferring functional regions without any prior knowledge of the structure-function relationship, there are drawbacks. Primarily, while they indicate conservation or shifts in function, they do not indicate function itself. Moreover, they do not work well if a significant number of mutations have occurred (e.g., for a typical nuclear encoded gene K_a/K_s is useful as far back as 150 million years, Gaucher et al., 2002), if there has been recombination (Yang and Bielawski, 2000), or large-scale structural rearrangements have occurred.

1.3 Examining molecular determinants of function

> "...it is clear that functional annotation from sequence, even at the biochemical level, will require much more than just identification of homology" (Thornton et al., 1999).

The ideas surrounding annotation and analysis of protein function described thus far largely rely on evolutionary conservation of sequence and structure. Even analysis of functional divergence as inferred using a multiple sequence alignment requires that the positions of aligned residues be homologous. The next wave of tools move beyond the notion of homology, and instead identify *patterns* that can arise independently in unrelated proteins. These patterns, including *linear motifs*, *3D or structural motifs* and *gene expression* patterns, are often direct determinants of function.

It is primarily these types of patterns that are studied throughout the remainder of this thesis because, unlike conserved patterns due to homologous relationships, these types of patterns can represent functional developments in protein families and do not require homology between sequences.

Linear motifs and 3D surface motifs

Small components of proteins that are localized in 1D (linear motifs) or 3D (structural motifs) often act as determinants of protein function. Programs such as PSORT and TargetP, for instance, predict protein subcellular localization through detection of signal peptides. Short linear motifs (SLiMs) have been defined as short (3-8 residue) amino acid functional motifs in proteins that can act as phosphorylation sites, targeting signals, protein-protein interaction motifs or other determinants of function (Figure 1.4). Given their short length and the diversity of folds in which they are found, SLiMs appear to have a propensity to arise through convergent evolution. Thus, they can be identified by correcting for common ancestry and identifying overrepresented motifs that have convergently arisen in unrelated structures, which is the basis of the approach used by programs such as DILIMOT (Neduva and Russell, 2006), and SLiMDisc (Davey et al., 2006). Because of the convergent nature in which these motifs arise in proteins, they represent "evolutionary interaction switches", and demonstrate how sudden changes in protein function may occur through subtle point mutations (Neduva and Russell, 2005). Moreover, they outline how overall sequence similarity between proteins may be a poor indicator of protein function, especially if the primary protein function involves, for instance, a key protein-protein interaction. Examples of well-known linear motifs include the SH3-binding motif (PxxP) and the phosphorylation motif (KDEL). Known linear motifs can be found in resources such as the Eukaryotic Linear Motif (ELM) resource (Puntervoll et al., 2003) and within PROSITE (Hulo et al., 2006).

One issue with linear motifs is that functional sites in proteins are not necessarily contiguous in the amino acid sequence. Thus, structure-based methods that aim to identify 3D motifs and surface patches within protein structures have been developed. Several of these operate from a similar evolutionary perspective as the previously mentioned approaches for identifying linear sequence motifs, as they intend to correct for homology and find 3D motifs that have developed independently in unrelated structures (convergence). The program FEATURE generates a statistical model of a binding site using physicochemical, spatial and residue information obtained from known binding sites (Bagley and Altman, 1995; Liang et al., 2003). Any structure can then be searched for sites that are significantly similar to those in the model library, where each site is evaluated by a naive Bayes scoring function. PINTS (Patterns In Non-homologous Tertiary Structures) (Stark and Russell,

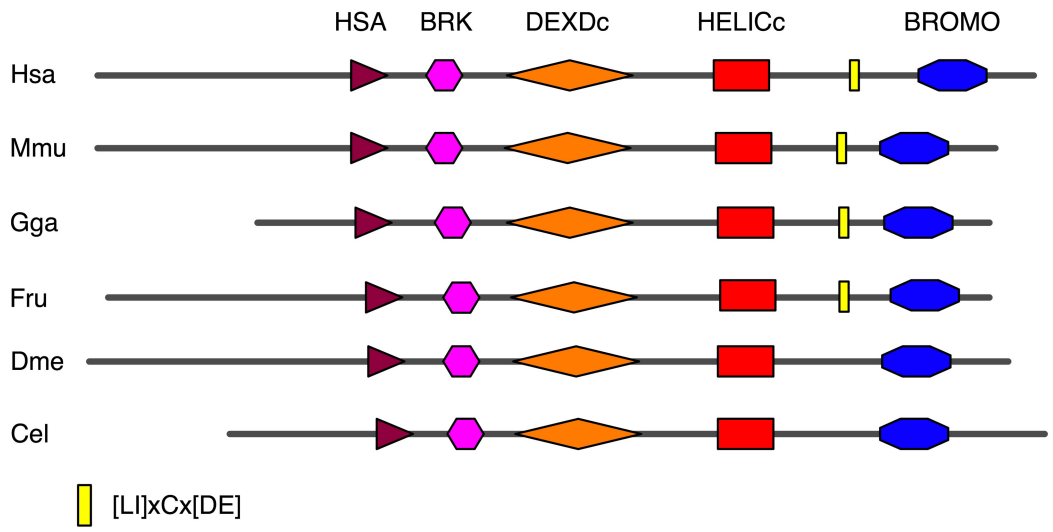


Figure 1.4: **An example short linear motif (SLiM) found in a subset of SNF2-alpha like proteins.** While similar protein domains are found in each member of the protein family, the SLiM (a determinant of protein function) is not found in all sequences. Species abbreviations are: Hsa, *Homo sapiens*; Mmu, *Mus musculis*; Gga, *Gallus gallus*; Fru, *Fugu rubripes*; Dme, *Drosophila melanogaster*; Cel, *Caenorhabditis elegans*. SMART domain annotations are indicated above the alignment. Adapted from Neduva and Russell (2005).

2003) identifies overrepresented sidechain patterns found in a group of unrelated structures, which can identify functional sites. Other approaches such as GASPS (Polacco and Babbitt, 2006) and FUNCLUST (Ausiello et al., 2008) also perform 3D motif searching. An example structural alignment of a functional 3D motif identified using FUNCLUST is shown in Figure 1.5.

Methods based on the identification of more general features of protein binding sites, including detection of binding clefts (Laskowski et al., 1996), patch analysis (Jones and Thornton, 1997) and identification of binding "hot spots" (Burgoyne and Jackson, 2006) have been developed. While methods that recognize general features of binding sites recognize direct physicochemical determinants of protein function, the functions themselves (i.e., ligand specificity, protein interaction partner) cannot be accurately predicted. Even if a 3D binding site model, trained on known examples, is used to predict new binding sites, it will only predict binding sites that conform to the model. However, different proteins may have evolved different molecular mechanisms for binding the same target or performing a

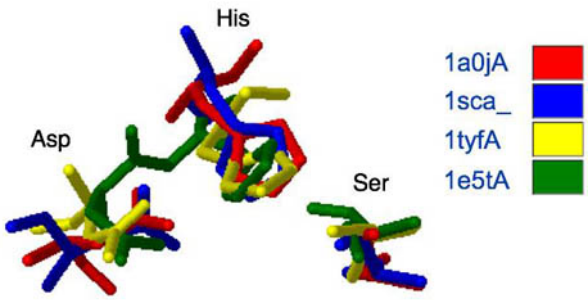


Figure 1.5: **An example of 3D motif matching performed by the FunClust server.** The example shown is the His/Asp/Ser catalytic triad from four non-homologous serine endopeptidases (PDB IDs 1A0Ja, 1SCA, 1TYFa, 1E5Ta). The example is one of three available preconfigured choices from the FunClust server (Ausiello et al., 2008), and the result was obtained using the default "active sites" parameter settings.

similar function.

Gene expression

An entirely different perspective for studying protein and gene function is the analysis of gene/protein *expression* as measured through microarrays or proteomics systems. *Gene expression analysis*, pioneered by Eisen et al. (1998), organizes genes not by evolutionary relatedness, sequence, or structural properties, but by their expression patterns which result from their underlying gene regulation.

Just as sequences can be classified and grouped by similarity, *gene expression profiles*, patterns of measured gene abundance from a variety of microarray experiments, can be clustered. For example, the result of a two-way (gene-wise and experiment-wise) hierarchical clustering experiment is shown in Figure 1.6, which reveals similar gene expression patterns common to different genes. Clustering of expression data is a useful approach for grouping together sequences with related function, as *co-expressed* genes often perform in the same pathway and play similar biological roles. Similarity of gene expression profiles[13] can be used to identify co-expressed gene clusters, and predict function by analyzing the known functions and annotations of cluster members (Hughes et al., 2000; Wu et al., 2002).

[13]Similarity of expression profiles can be measured simply through a Pearson correlation coefficient

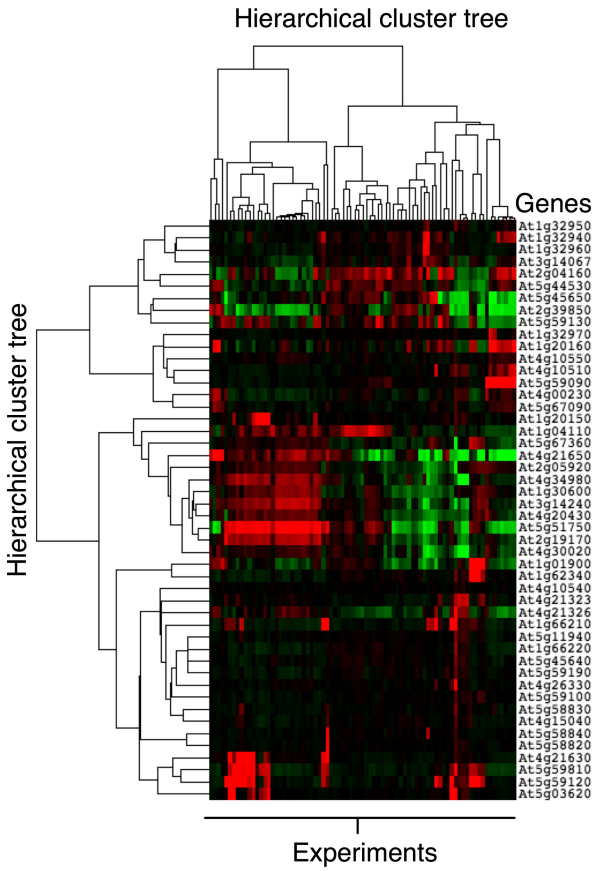


Figure 1.6: **Clustering of microarray data.** The figure illustrates an example of two-way hierarchical clustered microarray data. Listed genes are from the *Arabidopsis* subtilisin-like gene family. The microarray data is developmental expression data from Schmid et al. (2005). The red/green color scale indicates over-expression and under-expression, respectively, in various developmental stages (unlabeled).

Studies of how gene expression changes (*expression divergence*) or remains conserved between duplicate genes have therefore provided valuable insights into functional evolution (van Noort et al., 2003; Gu et al., 2005; Li et al., 2005; He and Zhang, 2005; Duarte et al., 2006). Expression divergence provides one avenue for understanding how proteins can stay highly similar in sequence and structure, with no apparent alterations of their molecular function, but change their biological function through changes in their temporal or spatial expression (Woody et al., 2008). Indeed, even proteins with 100% identity can have altered

(PCC). Cutoff values of PCC have been used to identify co-expressed gene clusters.

functions through diverged expression [e.g., eye lens proteins and lactate dehydrogenase, described in Whisstock and Lesk (2003)].

Studies of isoenzymes over thirty years ago first demonstrated that duplicate genes undergo specialization by diversifying expression in different tissues (Markert and Møller F, 1959). Such studies suggested that regulatory divergence is the first step in functional divergence of a newly duplicated gene (Ohno, 1970). Since this classical work, microarray data has allowed for large-scale studies of gene expression evolution in duplicate genes, gene families, and genomes. Large-scale studies of microarray data have shown that expression divergence generally increases over evolutionary time (measured by K_s). Expression diversity is also significantly greater among gene families than for single genes. However, the relative roles of subfunctionalization and neofunctionalization remain unclear (Li et al., 2005). He and Zhang's (2005) large-scale analysis of protein-protein interaction data in yeast suggests that there may be an initial period of subfunctionalization following gene duplication which is in turn followed by neofunctionalization.

Given that expression divergence contributes to functional divergence between gene duplicates, integrating gene expression analysis should improve sequence-based function prediction. Indeed, studies have shown that combining sequence conservation with expression conservation improves function prediction (van Noort et al., 2003).

1.4 Genomic context and interaction prediction

An overview of protein functional inference is incomplete without a discussion of methods for predicting protein-protein associations based on *genomic context* (Huynen et al., 2000). These methods largely assess various characteristics of homologous genes (in most cases, orthologs) across different genomes. *Gene fusion* can be used as a predictor of protein-protein associations because proteins that interact are often found as fusions (so-called "Rosetta Stone Sequences") in certain genomes (Marcotte et al., 1999). Assessment of *gene neighborhood*, which involves analysis of a gene's operon structure (in prokaryotes), is based on the idea that genes in close genetic proximity have similar functions or operate in similar pathways. *Phylogenetic profiling* is based on the idea that, if there is an interaction between two proteins they should display correlated occurrence in genomes (Pellegrini et al., 1999). The STRING database uses these features of genomic context, as well as

experimentally determined protein-protein interaction data (e.g., from large-scale yeast-two-hybrid studies), to predict functional associations between proteins (von Mering et al., 2003).

As many of these methods analyze genomic context across multiple genomes and transfer annotations/predictions between homologous genes, they are subject to many of the same limitations as homology-based methods. For instance, if phylogenetic profiling predicts a potential interaction between two proteins, that interaction does not necessarily exist in all cases where both proteins are detected in a genome.

1.5 Thesis outline: Tracing the origins of functional novelty

> "How does an entirely new function originate after gene duplication? More detailed molecular studies of model gene families are needed to look into the emergence of novel gene function" (Zhang, 2003).

The previous sections introduced the concepts of homology-based annotation, which relies on the transfer of functional annotations between homologous sequences. It was then described how modularity of protein function and evolution complicate the basic assumption that overall sequence similarity implies a related function or correlates with functional similarity. Approaches that break down sequences into evolutionary patterns of conservation or change, examine fundamental determinants of function, and analyze genomic context were then introduced.

Ultimately, none of the approaches described are necessarily good or bad, but rather depend on the sequences in question and their evolutionary history and function. In situations where there is detectable homology between sequences and conservation of all aspects of function, homology-based transfer is appropriate. However, if there is detectable homology between sequences that have undergone functional divergence, or if homologs of known function cannot be identified, homology-based annotation is inappropriate. In order to properly characterize these harder cases, which is the primary focus of this thesis, it becomes necessary to examine the molecular and evolutionary determinants of functional change or novelty.

Eight cases are examined, including proteins, protein families, gene clusters and entire genomes, whose function or evolutionary history defies complete explanation using approaches based on evaluation of overall sequence similarity. Each chapter deals with a different way in which evolutionary changes have occurred such that overall sequence similarity does not properly capture function or evolutionary history. Each example requires going beyond standard homology-based annotation to instead analyze additional molecular traits that provide deeper clues to existing or ancestral functions. This ultimately involves tracing the molecular evolutionary roots of new functions or new protein families.

Chapter 2 concerns functional changes that develop via mutations of the *protein surface*. In these cases, a protein's fold stays conserved, but subtle surface patterns (e.g., binding sites) can arise through a small number of point mutations leading to a change in function. These changes may cause related proteins to have different functions or unrelated proteins to have similar functions, and thus represent an excellent example of a scenario in which standard homology-based annotation does not work. Two examples are explored in this chapter: ice-binding (antifreeze) proteins and glycan-binding proteins. While proteins in each of these functional categories cannot be unified by sequence similarity, it is shown that a 3D, algorithmic representation of their binding site patterns can effectively discriminate these two classes of proteins from other types of proteins, and predict novel cases where sequence-based approaches cannot.

Chapter 3 deals with functional changes resulting from shifts in *gene expression*. The β-1,3-glucanase family from *Arabidopsis* is examined in order to study how changes in gene regulation can underlie functional changes *within* a homologous gene/protein family. In this case, the family members are similar in sequence and 3D structure, and share active sites and likely enzymatic activity, but have adopted new functions due to changes in their regulatory patterns. This example demonstrates yet another way in which the transfer of a functional annotation between related sequences can be overly simplistic. While the *molecular function* of these sequences may be similar or even identical, their roles in terms of biological process have changed. One example outlined in detail is a proposed series of events through which ancestral β-1,3-glucanases involved in normal plant development evolved into a subfamily of pathogen-defense proteins.

Chapter 4 examines a common protein fold (the β-trefoil) to study how *protein structures* themselves can originate and evolve through *internal repeats*. It is shown that even folds/domains, which are commonly thought to be basic (and relatively indivisible) units

of protein evolution, can originate and evolve by repetition of internal repeat elements (supersecondary structural elements). These repeat events are not accounted for by standard approaches that examine overall sequence or structural similarity at the protein domain level. Through a large-scale analysis of internal repeats in β-trefoil proteins, numerous cases are identified whereby evolution has independently assembled a β-trefoil domain or family from a distinct precursor repeat. These events have occurred most frequently in the ricin-type family of lectins, many of which have roles as carbohydrate-binding toxins. One family (the *Hydra* PPOD family) that has undergone unique internal repeats is then analyzed in detail, which reveals its evolutionary origin, a function in carbohydrate-binding, and putative mechanism. The results suggest that ongoing internal repeat events are important for maintaining symmetry-related functions (i.e., multivalent carbohydrate-binding), which are masked by bioinformatic approaches that operate at the level of protein domains.

Chapter 5 explores protein sequences that evolve extremely rapidly, have likely undergone *recombination* and thus exhibit *mosaicism*[14]. As a consequence, their evolutionary signal has become scrambled, sequence similarities to other proteins are too distant to recognize statistically, and so homology-based methods do not work. Two examples are assessed in this chapter [clostridial neurotoxins (CNTs) and bacterial flagellins], both of which play major roles in bacterial virulence. In order to uncover ancestral functions and relationships regarding CNTs, I examine not only the individual neurotoxin protein but also adjacent genes in the neurotoxin gene cluster, which reveal additional clues to a recurring ancestral function. By examining recurring traits within the neurotoxin gene cluster, including sequence, structural, and small motif similarities, an ancestral function in collagen adhesion and degradation is inferred, which links CNTs to a potentially ancestral family of toxins.

The middle, antigenic hypervariable region (HVR) of bacterial flagellin is then studied as an additional example of rapidly evolving mosaic sequence family. In this sequence family, the HVR can evolve independently of the rest of the sequence which complicates assessments of flagellin similarity. In addition, due to rapid evolution, HVRs are too divergent to assess by standard alignment methods. Thus, an approach is developed to cluster HVRs into types and map them onto a phylogenetic tree of a conserved flagellin

[14]Mosaic sequences have been assembled by recombination and thus contain fragments with dissimilar phylogenetic histories.

domain, which reveals their evolutionary patterns and recombination events. The results show that many distantly related flagellin sequences are likely closer in function than their phylogenetic neighbors due to HVR recombination, and ultimately that overall sequence similarity between flagellins may in many cases be a poor indicator of functional similarity.

Lastly, in Chapter 6, I look at *host-mimicry* relationships of bacterial pathogens and their virulence factors, where considerable evolutionary novelties have occurred recently and frequently. Thus, the analysis shifts from individual proteins, families and gene clusters, to entire genomes. Just as the preceding chapters demonstrate how overall sequence similarity does not adequately predict function, overall genome similarity (i.e., species relatedness) does not adequately define a role in pathogenicity/virulence of bacteria towards humans. Furthermore, virulence factors themselves cannot be unified by sequence as protein virulence mechanisms have evolved independently in many unrelated protein lineages. In order to uncover a more fundamental determinant of bacterial pathogenicity, a large-scale comparative analysis of *host-protein mimicry* is performed. The hypothesis that similarities to host (i.e., human) proteins may be overrepresented in virulence factors compared to random proteins is tested. This feature is shown to be a significant predictor of known virulence factors, and is used to infer numerous novel proteins involved in bacterial pathogenesis of humans as well as their potential mechanisms.

A recurring theme in each chapter is that overall sequence similarity can be a poor indicator of function or evolutionary history when functional innovations have occurred. This is shown to apply to individual gene/protein sequences, as well as entire genomes. The cases examined largely involve virulence factors and pathogen-defense proteins, which, because of the nature of host-pathogen coevolution, have undergone—and continue to generate—many functional changes. Each case presented has undergone different types of evolutionary change due to differing evolutionary pressures. Thus, while all cases require going beyond standard homology-based approaches, they also require analysis of different functional traits (e.g., surface patterns, gene expression characteristics, repeats, HVRs). As each protein/gene/genome is unique, no single approach to functional and evolutionary characterization is generalizable. However, most if not all of the traits examined have a convergent/recurring or evolutionarily plastic quality. A summary of the analyses is shown in Table 1.1, each of which can be described as: a problem that cannot be solved by examining overall sequence similarity, a proposed solution to that problem involving analysis of additional functional traits, and an evolutionary scenario through which the

group in question has changed in function.

Lastly, tailored approaches are used for analyzing a series of individual examples, as opposed to applying a single, general approach to each case. This was motivated by two ideas. First, it is unreasonable to expect that functional diversification always follows the same patterns in different cases (Hughes, 2005). Second, more studies of real protein families are needed to understand the emergence of novel gene function (Zhang, 2003). Very few examples exist outlining the molecular details underlying functional novelty in protein/gene families (Hughes, 2005). A common philosophy of each analysis is that the *specifics* concerning the unique structure-function relationships of the cases examined are extremely important. These specifics, which are different for each case examined, are taken into account computationally and used to investigate their unique patterns of evolution change, functional diversification and novelty.

Table 1.1: **Summary of analyses presented in this thesis.**

Group analyzed	Problem	Approach Used	Molecular Adaptation(s)
Chapter 2			
Antifreeze Proteins	Cannot be unified by seq. similarity	Identify repetitive **surfaces**	Convergent evolution of ordered surfaces in unrelated folds
Glycan-binding proteins	Cannot be unified by seq. similarity	Identify **surface** aromatic motifs	Convergent evolution of aromatic motifs in unrelated folds
Chapter 3			
β-1,3-glucanase family	Diversification of function unclear based on sequence	Trace divergence of gene **expression** traits	Functions have diversified through changes in expression
Chapter 4			
β-trefoil superfold	Origins unclear and symmetries diverse	Analyze individual **repeats**	Parallel/convergent evolution of fold/domain from structural repeats
"PPODs"	Origins and function unknown	Analyze **repeats** and recurring surface motifs	Internal repeats, triplication of aromatic motif, and lateral gene transfer
Chapter 5			
Flagellins	Too divergent to analyze; functions and history unclear	Trace HVR types on phylogenetic tree	Functions diversified through HVR **recombination**, loss, and duplication
Neurotoxin gene cluster	Origins of toxin and ancestral functions unknown	Analyze recurring structural, sequence, and motif similarities throughout gene cluster	Extreme divergence and **recombination** from an ancestral toxin gene cluster
Chapter 6			
Bacterial pathogens and their virulence factors	Poor classification based on genome/sequence similarity	Analyze patterns of host-protein similarity/**mimicry**	Coevolution with host proteins, repeats, lateral gene transfer, convergent or parallel evolution

Chapter 2

Protein surface adaptations

The following chapter includes published material from Doxey et al. (2006)[1].

Protein surfaces are key to understanding the structure-function relationships of proteins. Residues exposed on the surface of proteins are capable of forming interactions with other molecules, while residues in the interior of the protein are restricted to internal interactions. Thus, analyzing the geometry and residue characteristics of a protein surface is critical for understanding how function is specified in the underlying protein sequence, and how function can change through mutation. Function-altering surface mutations may be extremely subtle in nature, and involve a small number of single amino acid substitutions.

Important surface determinants of function may be missed by homology-based approaches to functional annotation, which focus on overall similarity as opposed to subtle differences on protein surfaces. Moreover, a homology-based approach is incapable of unifying a functional class that includes proteins of different evolutionary histories. Antifreeze proteins (AFPs) and other ice-binding proteins are an excellent example of a protein functional class that cannot be unified by existing bioinformatic methods, and a "textbook example" of convergent evolution (Logsdon and Doolittle, 1997). In section 2.1, an algorithm is developed to detect a more fundamental molecular determinant of AFPs (*ordered surface carbons*). This feature is able to recognize and unify existing AFPs based on structure alone, and predict novel AFPs where sequence-based methods cannot.

[1]Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology (Doxey et al., "Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions", 24, 7, 852-5), copyright (2006).

Similar to ice-binding proteins, glycan-binding proteins and other carbohydrate-binding proteins are extremely diverse in terms of structure, with no single fold, domain family or sequence motif that can unify them and recognize novel cases. By detecting a common molecular feature of glycan-binding sites (*coplanar surface aromatics*), it is shown in section 2.2 that a wide array of existing glycan-binding sites as well as novel binding sites can be recognized based on structure. These two examples demonstrate the power of using recurring surface patterns that have arisen by convergent evolution to classify and predict protein function.

2.1 Antifreeze Proteins

2.1.1 Introduction

AFPs are found in cold adapted organisms and have the unusual ability to bind to and inhibit the growth of ice crystals. Understanding how AFPs bind to ice has been a fundamental question in AFP research since their discovery in fish and insects over 30 years ago (Duman and DeVries, 1974) and is crucial in the development of potential biotechnological applications (cryogenic storage, cryosurgery, food preservation, freeze-resistant crops, Knight, 2000) and rational AFP design.

AFPs present several experimental and computational challenges. Since AFPs cannot be crystallized with ice (Jia and Davies, 2002), there is no direct approach for identifying AFP-ice interfaces. Thus, structure-function studies rely mainly on site-directed mutagenesis and computational methods, which have led to several different hypotheses regarding AFP-ice interactions. Bioinformatic classification of AFPs is also challenging because they do not belong to a single protein family and thus cannot be unified based on sequence alone (i.e., using a profile or motif-based approach) or even fold. Thus, AFPs are an example of a protein functional class that has evolved independently in different protein lineages through convergent evolution.

In this work, the problem of computational prediction and analysis of AFP function is tackled by development of a structural pattern detection algorithm. The pattern detection algorithm is able to unify existing AFPs and predict a novel AFP from the cold-adapted plant, winter rye, as well as its putative ice-binding region.

2.1.2 Recognizing the physicochemical basis of AFP function

As asked previously (Jia and Davies, 2002), what distinguishes AFPs from ∼99.9% of proteins that have no observable affinity for ice? To answer this question, a generic model could be used to compare a full set of AFP structures with other proteins in the Protein Data Bank (PDB, Berman et al., 2000). One interaction model proposes that AFPs bind ice via relatively flat, hydrophobic surfaces that are complementary to surfaces of ice (Jia and Davies, 2002). Testing such a model, however, is not feasible using current algorithmic methods (e.g., molecular docking) as they are computationally intensive for just a single AFP-ice system, and are highly dependent on choice of energy function and definition of the ice surface (Jorov and Zhorov, 2004). To circumvent this issue, a rapid algorithm has been developed to score the protein directly based on a set of physicochemical surface features, allowing for comparative evaluation of multiple AFPs and thousands of PDB structures.

The basic tenet of the algorithm is that spatially regular surface atoms should have an increased probability of docking to an ordered substrate such as ice and, therefore, should be most concentrated on the ice-binding surface. Spatial regularity is scored explicitly by computing vectors between solvent accessible atoms and determining which atoms are associated with a local, repetitive vector pattern (Figure 2.1). This approach will recognize highly ordered, relatively planar surfaces stemming from a repetitive geometric pattern, thus allowing for detection of ice-binding surfaces regardless of their ice-plane specificity and binding orientation. Extending the "hydrophobic surface" model (Harding et al., 1999; Jia and Davies, 2002), the algorithm assesses the spatial regularity of solvent accessible, non-polar carbons instead of polar atoms. This is consistent with an ice-binding mechanism in which carbons (e.g., methyl groups) penetrate regular spacings (e.g., ice cages, grooves, hexagonal rings) on the ice-surface, forming favorable van der Waals contacts (Yang et al., 1998; Jorov and Zhorov, 2004). Carbons from aromatic and charged residues have been omitted, as aromatic groups are too bulky to be accommodated within interstitial spaces of ice and have been shown to reduce antifreeze activity by mutational studies (DeLuca et al., 1998), and interactions of charged residues energetically favor liquid water over ice (Gallagher and Sharp, 2003). An additional desirable feature of an ice-binding surface is that it is sufficiently hydrophilic to be solvent-exposed in its native state. Therefore, only surface carbons in close proximity to polar atoms are selected. According to this model, polar atoms are required primarily for solubility, but a supplementary role in hydrogen-

bonding interactions with ice cannot be ruled out. The final output of the algorithm for a putative AFP is an identified set of spatially regular, hydrophobic carbons, termed "ordered surface carbons" (OSCs).

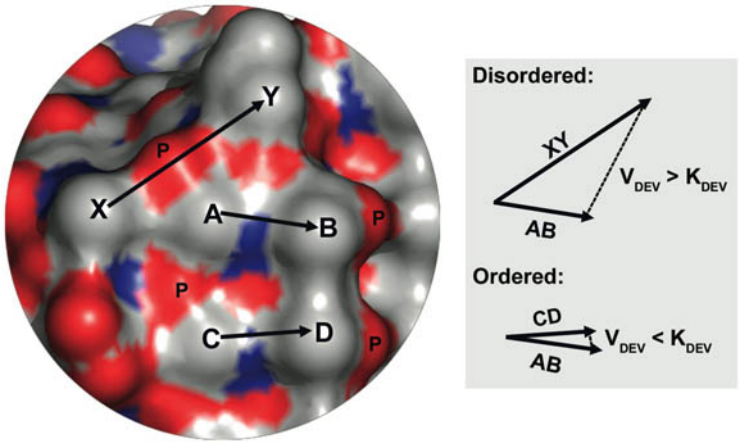


Figure 2.1: **Identification of ordered surface carbons using vector comparison algorithm.** Identification of ordered surface carbons using vector comparison algorithm. Atoms {A,B,C,D,X,Y} are sufficiently solvent exposed carbons (SAS > K_{SAS}). Atoms {A,B,C,D} are spatially regular since |AB − CD| < K_{DEV} and A,B is near C,D (within distance K_{HIGH}). These atoms are in also close proximity (within distance Kpolar) to polar atoms (some of these have been labeled P). A,B,C,D are ordered surface carbons (OSCs).

2.1.3 Methods

Algorithm

The algorithm (see Figure 2.2) was developed in the Perl programming language, and uses a standard PDB file as input. Per-atom solvent accessible surface area (SAS) is calculated using a rapid Voronoi procedure (McConkey et al., 2002). Sufficiently solvent accessible carbons are isolated using a cutoff value (K_{SAS} = 15.1 Å^2). Carbons from uncharged, non-aromatic amino acids within distance K_{POLAR} (4 Å) of a solvent accessible polar atom (oxygen or nitrogen) are selected. Two carbon pairs {A,B} and {C,D} meeting the above criteria are scored for regularity if the minimum distance between atoms in separate pairs is within K_{LOW} to K_{HIGH}. These atoms are deemed ordered if they produce equivalent

Input: Standard PDB file (hydrogen atoms excluded)
Output: Largest cluster of ordered surface carbons
Parameters: K_{SAS} = 15.1 Å^2, K_{POLAR} = 4 Å, K_{LOW} = 4.5 Å, K_{HIGH} = 7.66 Å, K_{DEV} = 1.09 Å

```
1   for all carbon atoms from uncharged, non-aromatic residues do
2       compute SAS(1)
3       if (SAS > K_SAS) and (carbon is within K_POLAR of a surface polar atom)
4           add to list L_Solvent Accessible Carbons
5       endif
6   endfor
7   for all carbon pairs (a,b) in L_Solvent Accessible Carbons
8       compute interatomic distance d_ab
9       if (K_LOW < d_ab < K_HIGH)
10          add (a,b) to list L_Surface Carbon Pairs
11      endif
12  endfor
13  for all carbon pairs (a,b) and (c,d) in L_Surface Carbon Pairs
14      compute min distance q between (a,b) and (c,d)(2)
15      compute deviation V_DEV between vectors ab and cd(3)
16      if (K_LOW < q < K_HIGH) and (V_DEV < K_DEV)
17          add a, b, c and d to list L_Ordered Surface Carbons
18      endif
19  endfor
20  connect any two carbons (a,b) in L_Ordered Surface Carbons if d_ab < K_HIGH
21  return largest connected set of ordered surface carbons
```

(1) Solvent-accessible surface area is calculated using a rapid Voronoi procedure[23].
(2) Min{d_{ac},d_{ad},d_{bc},d_{bd}}
(3) V_{DEV} = |**ab** - **cd**|

Figure 2.2: **Algorithm for prediction of OSCs forming putative ice-binding surfaces.**

vectors ($|\mathrm{AB} - \mathrm{CD}| < \mathrm{K_{DEV}}$). OSCs from equivalent vectors are clustered together if their minimum inter-atomic distance is within distance $\mathrm{K_{HIGH}}$. The largest cluster of OSCs is defined as the putative ice-binding surface. A quantitative measure of ice-binding potential is estimated as the surface area occupied by these carbons. Both fractional and total surface area is calculated. As carbons should also be adequately separated to prevent steric interference with ice-binding, an upper bound ($\mathrm{K_{HIGH}}$ = 7.66 Å) and lower bound ($\mathrm{K_{LOW}}$ = 4.50 Å) limits the length of carbon-to-carbon vectors. $\mathrm{K_{DEV}}$ (1.09 Å), the maximum deviation allowed for two vectors to be considered equivalent, controls the stringency of the algorithm, and is optimized based on the resolution of target structures. Parameter values were initially set based on theoretical considerations (i.e., $\mathrm{K_{HIGH}}$ and $\mathrm{K_{LOW}}$ correspond closely to the unit cell dimensions of hexagonal ice [7.345 Å and 4.516 Å, Madura et al. (2000)], and were locally adjusted to increase stringency while maintaining

ice-binding surface prediction accuracy. Three previously characterized AFP structures (PDB IDs 1WFA, 1EZG, and 1MSI) were used to validate parameter values.

3D structure dataset

A set of 3,196 non-redundant structures was generated using the PISCES server (Wang and Dunbrack, 2003). Protein chains between 40 and 10,000 residues with less than 25% pairwise sequence identity, R-factors below 0.3 and with resolution below 3Å were selected. NMR entries, C-α only structures, and error-containing PDB files were excluded to obtain a set of high quality X-ray structures. A representative set of AFP structures from the PDB was obtained by selecting the highest resolution structures (or most representative NMR model) from separate crystallization or NMR studies, and then selecting the top three (if available) structures for separate AFP classes (type I, type II, type III, and β-helical insect AFPs). The most representative model for NMR studies was determined using the NMRCLUST procedure from the OLDERADO database (On-Line Database of Ensemble Representatives and DOmains (Kelley and Sutcliffe, 1997). AFP structures not already in the non-redundant set were added manually.

AFP prediction and homology modeling

A set of protein sequences was obtained in previous studies of the freezing-tolerant plants winter rye (*Secale cereale* L. cv Musketeer) and Saltwater cress (*Thellungiella salsuginea*). All sequences were identified from cold-acclimated plant tissue but the corresponding proteins had not been purified or tested for ice-binding activity. Of these, twenty novel sequences were homologous to proteins of known structure and permitted comparative modeling. 3D models were constructed using a fully automated procedure available through the SWISS-MODEL server (Schwede et al., 2003). No further manipulations were made to the models. Entries with the lowest E-value in a PDB-BLAST search were selected as structural templates for homology modeling. Non-specific lipid-transfer protein from maize (PDB ID 1FK2) was used as a template for both LTP1 and LTP2 models. Swiss PDB Viewer version 3.7 was used for structure visualization (Guex and Peitsch, 1997), and WHATCHECK (Hooft et al., 1996) was used for structure verification and revealed no major errors.

Cloning and antifreeze assays

NOTE: Cloning and antifreeze assays were performed by Dr. Mahmoud Yaish.

LTP1 and LTP2 clones were inserted in the pGEX-KG expression vector and fused with GST tags. Positive colonies were isolated, and the recombinant fused protein in the lysate was purified using 1 mL GSTrap^{TM}FF and HiTrap Benzamidine FF columns (Amersham Biosciences) according to established protocols.

Antifreeze activity was assayed qualitatively by observing ice-crystal growth morphology in solution using a Clifton nanoliter osmometer (Hartford, NY) mounted on a phase-contrast photomicroscope (Olympus BHT, Tokyo). Solutions were flash frozen and melted until a single ice-crystal remained, and cooled to observe changes in morphology as the ice crystal grew. Protein concentrations were 33 μg/mL. Multifaceted (e.g., hexagonally shaped) ice-crystals are produced in presence of AFPs, while round, flat ice-crystals indicate lack of ice-binding activity (DeVries, 1986; Hon et al., 1994; Griffith and Yaish, 2004). Protease treatment was used to verify crystal morphology was due to presence of protein, as described previously (Bravo and Griffith, 2005).

2.1.4 Results

Discrimination of AFPs and binding site prediction

The algorithm was applied to a representative set of AFP structures from the PDB including three each of type I AFPs, type III AFPs, and β-helical insect AFPs, as well as the one available NMR model of type II AFP. Predicted OSCs (Table 2.1 and Figure 2.3) for nearly all AFPs were in excellent agreement with ice-binding residues identified by previous studies (Harding et al., 1999; Antson et al., 2001; Graether and Sykes, 2004).

Due to the structural regularity and amphipathic nature of α-helical type I AFPs, OSCs were found predominantly on their conserved, Ala-rich, "hydrophobic faces" (Harding et al., 1999). Moreover, OSCs detected on insect AFPs corresponded to their highly ordered, Thr faces, and OSCs recognized on type III AFP were in almost perfect correspondence with established ice-binding residues (Antson et al., 2001). Previous type III AFP ice-binding models involve many of these residues in H-bonding interactions with ice (Chao et al., 1994; Jia et al., 1996; Graether et al., 1999). It is shown here, however, that these

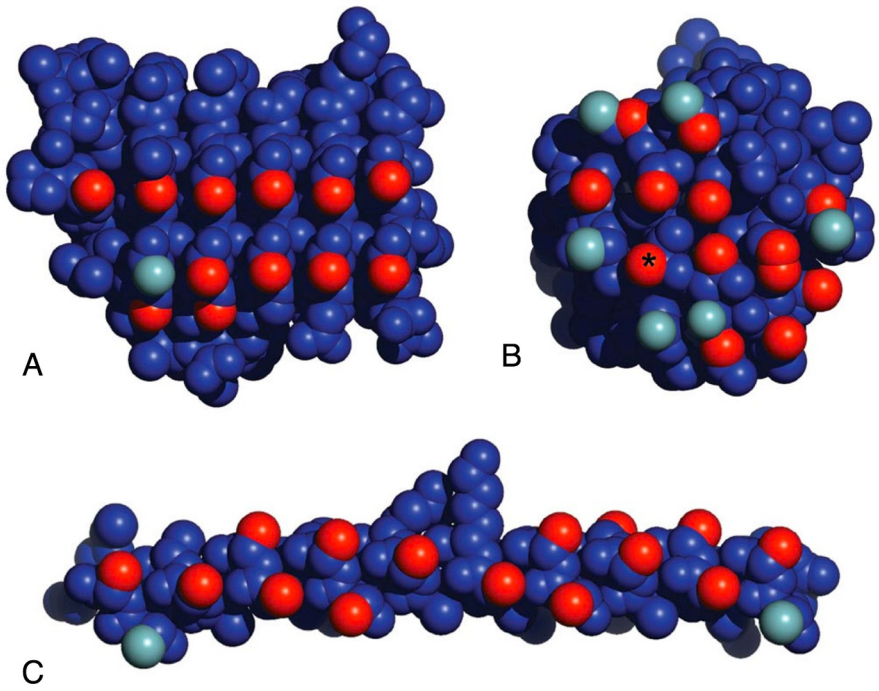


Figure 2.3: **Ordered surface carbons detected in known AFPs.** (A) Spruce budworm AFP (PDB ID 1L0S:A). (B) Type III AFP (PDB ID 1MSI). (C) Winter flounder AFP (PDB ID 1WFB:A). Atoms highlighted in red are OSCs detected by the algorithm. Atoms (of any type) highlighted in light blue share the grid-like arrangement formed by neighboring OSCs but were not explicitly recognized by the algorithm. The OSC from A16 (labeled with an asterisk) was not identified in 1MSI but was identified in 1UCS:A and 1HG7, type III AFP homologs of 1MSI.

sites can be detected by focusing on OSCs and that consideration of the orientation of polar atoms is not required. This is further reinforced by a relatively small reduction in performance when polar atom information was omitted.

In many cases, the identified ice-binding surfaces represent relatively flat regions on the protein surface, an important feature of ice-binding that was utilized by the "flatness function" algorithm (Yang et al., 1998). In addition to planar residues, the vector-based algorithm presented here also identified out-of-plane residues flanking the planar region of the ice-binding surface, which have been implicated in type III AFP's mechanism of action (Graether et al., 1999; Antson et al., 2001). Such peripheral residues may allow type III AFP to simultaneously bind multiple ice-surfaces (Jia and Davies, 2002) or may

Table 2.1: **Residues contributing predicted ordered surface carbons.**

AFP	PDB ID	Exp. Method	Residues contributing predicted OSCs and correspondence with previous models
Type I Winter Flounder AFP (wfAFP)	1WFB:A	X-ray	A3, S4, **A6**, A7, A8, **A10**, A11, L12, **T13, A14**, A15, **A17, A21**, **T24, A25, A28, A29, A32, A36**
Hydrophobic analogue of wfAFP	1J5B:2	NMR	A3, **A6, A9, A10, V13,** N16, A20, **V24**, N27, A30, A31, **A32**, A34, V35
Type I Shorthorn Sculpin AFP	1Y04:18	NMR	**P6, A9, A10, A11, T13, A14**, A15 **A17, L18, A21, A25, A29, A32, A34**
Type III Fish AFP	1MSI	X-ray	**L10, P12, I13, N14, T15, T18, L19 V20, M21, V41**, A48
Type III Fish AFP	1HG7	X-ray	**L10, P12, I13, N14, T15, A16, T18 L19, V20, V41**, A48, P50
Type III Fish AFP	1UCS:A	X-ray	**L10, P12, N14, A16, T18, L19, V20**, **M21, V41**, A48, P50
Yellow Mealworm Beetle AFP	1EZG:A	X-ray	**T4**, G5, **A15, T17**, G18, V26, **T27**, **T29**, N38, **T39, T41**, Q50, **T51, T53**, N62, **T63, T65, T77**
Spruce Budworm AFP	1L0S:A	X-ray	**T5, T7, T21, T23, T36, T38, T51**, **T53**, G54, **I68**, G71, **A75**
Spruce Budworm AFP isoform 501	1M8N:A	X-ray	**T21, T23, I37, T39, V52, T54, T67**, **T84, T99, T101**
Type II Sea Raven AFP	2AFP:3	NMR	None
Winter Rye LTP1	NA	Hom. model	**A1, T3, C4, A44, T48, N51**

Bolded residues correspond to: hydrophobic faces for type I AFPs [see Figure 8 in Harding et al. (1999)], Thr faces (Graether and Sykes, 2004) for insect AFPs, and ice-binding residues as described by Antson et al. (2001) for type III AFPs. The hydrophobic face of wfAFP (hydrophobic analogue) is assumed to be the same as type I wfAFP.

alter hydration structure to further impede crystal growth (Graether et al., 1999).

To determine whether AFPs could be distinguished from other structures in the PDB based on algorithmic prediction of OSCs, 3,206 non-redundant structures including ten wild-type AFPs were scored (Figure 2.4). Scored by fraction of surface area occupied by predicted OSCs (FSA), 9/10 AFPs ranked above 99.3% of other PDB structures. Scored by total surface area of predicted OSCs (TSA) alone, these AFPs ranked above the 98.6th percentile.

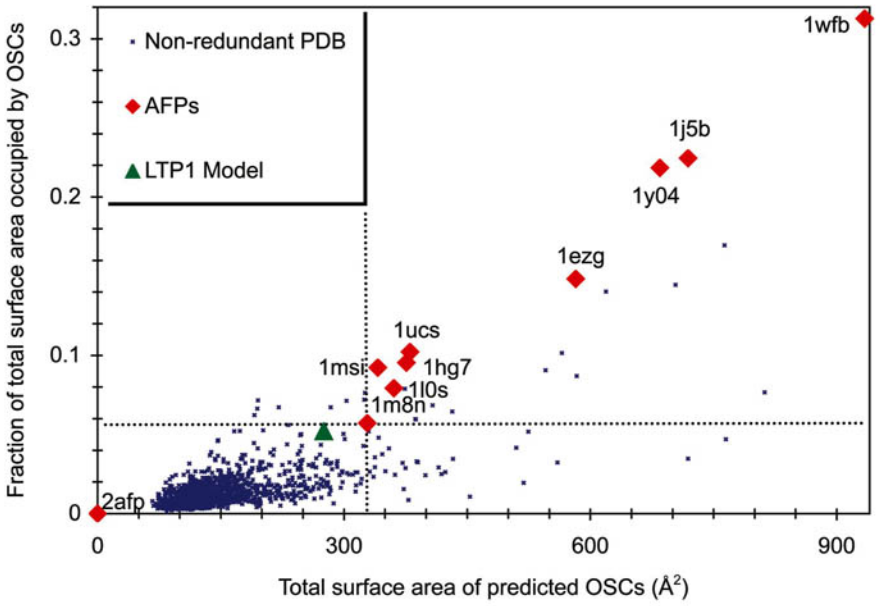


Figure 2.4: **Scoring of AFPs versus non-redundant PDB structures.** Each point represents a single structure. The structures are positioned along the y-axis by one scoring scheme (size of predicted IBS as fraction of surface area, FSA) and along the x-axis by another scoring scheme (size of predicted IBS, independent of total surface area, TSA). The plot has been divided in four quadrants at the data point representing structure 1m8n. The two left quadrants contain one AFP structure (type II NMR model) and 3165 others. The bottom-right quadrant contains no AFPs and 29 others, and the top right quadrant contains nine AFPs (seven X-ray structures and two NMR models) and 12 others.

Type III mutant structures also ranked significantly high relative to other structures in the PDB, with an average rank of 99.3% by FSA and 93.8% by TSA. In addition, when the algorithm stringency was relaxed to account for lower resolution models in the type III mutant data set, a correlation ($r = 0.66$) emerged between algorithm score and thermal hysteresis[2] (Figure 2.5). This is consistent with a previous neural-network study, which found non-polar solvent accessible surface area to be the best predictor of thermal hysteresis (Graether et al., 1999).

Discriminative ability and ice-binding surface prediction accuracy were generally lower

[2]Thermal hysteresis refers to a measured difference between the melting point and freezing point of a solution. Addition of certain AFPs to a solution reduces the freezing point and thus increases thermal hysteresis in a non-colligative fashion (Duman and Serianni, 2002).

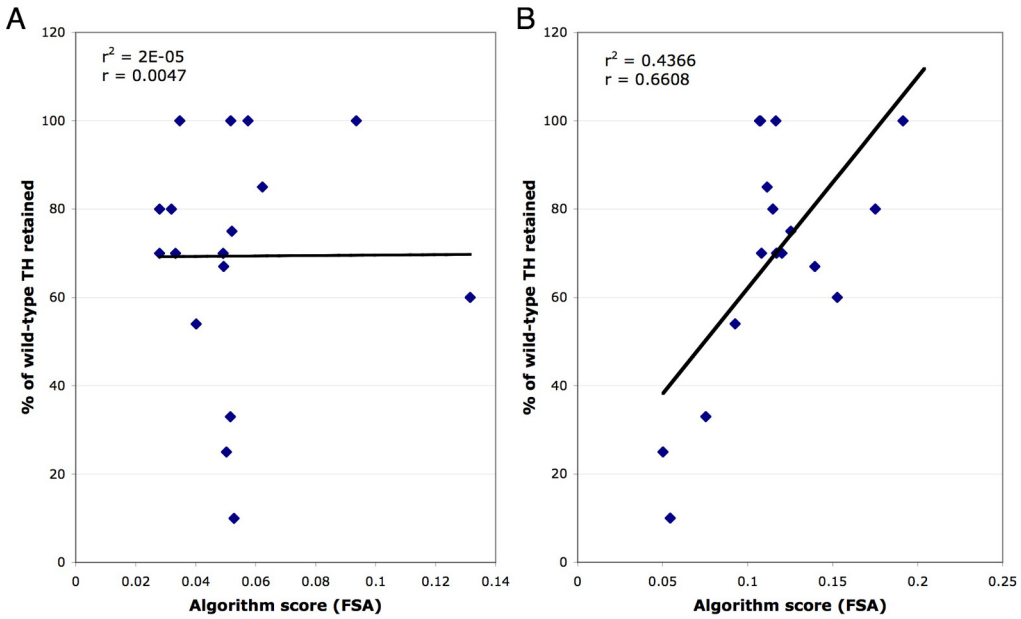


Figure 2.5: **Correlations between type III mutant thermal hysteresis values and algorithm score (fractional surface area occupied by OSCs).** Two plots were made for comparison using K_DEV parameter values of (A) 1.09 Å (B) 1.5 Å. See Table 3 in Graether et al. (1999) for the dataset (list of PDB entries and corresponding thermal hysteresis values.

for NMR models compared to X-ray structures. However, while the representative type II NMR model received a score of zero, one of the five models within PDB ID 2AFP received high scores of 96.2% (FSA) and 94.5% (TSA). Variable results for NMR models are not surprising as the algorithm relies on inter-atomic distances and parameters on the order of one to two Å, an atomic resolution not typically achieved using NMR. An inability to predict protein interaction sites using NMR data has been reported previously (Fernandez-Recio et al., 2005), and was also attributed to insufficient resolution of surface side-chains. In these cases, relaxing the parameters (i.e., lowering the stringency of vector-matching) can improve identification of true OSCs but also increases the rate of false positives. Other highly scored structures in addition to AFPs included viral coat proteins, membrane proteins, and subunits within protein complexes. In these cases, many high-scoring surfaces detected by the algorithm were protein-protein contact surfaces within the full biological unit and would therefore be unable to interact with ice.

Prediction of a novel ice-binding protein

As there were no obvious novel candidate AFPs within the top ranked structures, the algorithm was applied to homology models based on a set of sequences previously isolated from the freezing-tolerant plants winter rye (*Secale cereale* L. cv Musketeer) and Saltwater cress *Thellungiella salsuginea*. Winter rye is known to encode AFPs that are homologous to pathogenesis-related proteins, but their structure-to-function relationships are unclear (Griffith and Yaish, 2004). The isolated sequences included lipid transfer protein (LTP) homologs, some of which are known to be expressed in response to cold and have cryoprotective functions (Hincha et al., 2001) but have not been shown to have antifreeze activity. Where possible, structural models were generated using X-ray structure templates. From twenty constructed models, a lipid transfer protein homolog (LTP1) yielded a significant ice-binding surface score and was selected as a positive test. The algorithm recognized a set of six highly planar OSCs (four methyl groups from Ala and Thr residues and two β-carbons from Cys and Asn residues) forming the putative ice-binding surface. Scored among structures in the non-redundant set, LTP1 ranked higher than 99.2% of other PDB structures using FSA and 97.5% using TSA. LTP1's score was considered to indicate a high probability of ice-binding activity, particularly given that a homology model was used. Another modeled lipid transfer protein homolog (LTP2) received a score of zero despite having 70% sequence identity with LTP1 (Figure 2.6A). The lower score relative to LTP1 resulted from a small number of amino acid substitutions on the analogous surface (Figure 2.6B), which presented LTP2 as an ideal negative test. In place of several predicted ice-binding residues in LTP1 (Thr 3, Ala 44, and Thr 48) are Ser 3, Asp 44, and Ala 48, which disrupt regularity and introduce a charged side-chain to the putative ice-binding surface.

The predicted ice-binding activity was tested experimentally using the ice-crystal growth morphology assay (Figure 2.6C). LTP1 bound to ice, producing hexagonal shapes (characteristic of AFP-ice interaction) that grew further to produce unusual six-pointed star shaped crystals as the temperature was decreased. Similar morphologies have been observed in isolates from antifreeze-active fungi (Snider et al., 2000). In contrast, LTP2 did not exhibit ice-binding activity as evidenced by a round, flat ice-crystal morphology, which was indistinguishable from control samples. The ability of the algorithm to predict ice-binding activity for a new AFP fold provides additional experimental validation for the model.

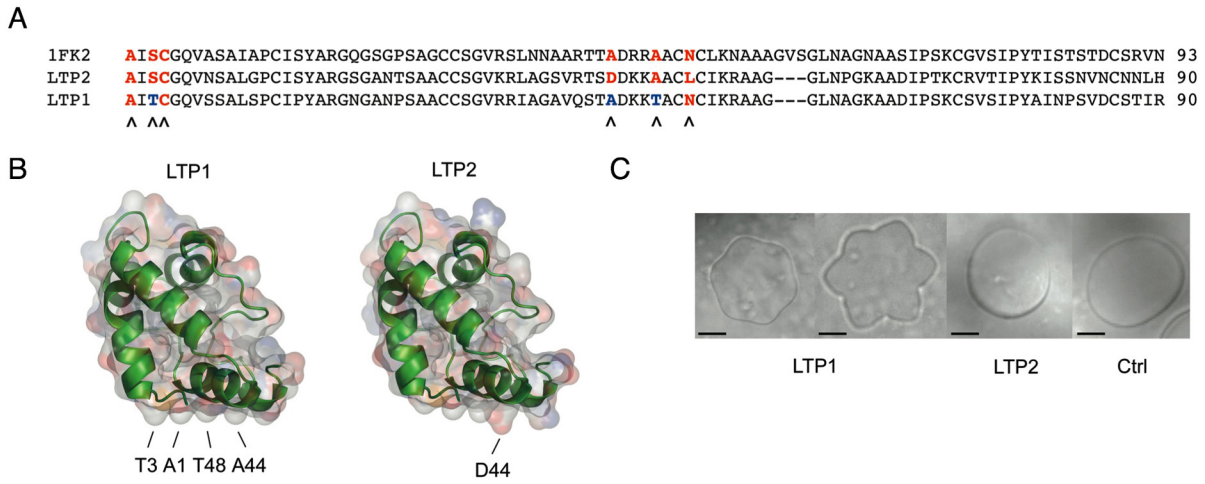


Figure 2.6: **Prediction of ice-binding activity from homology models.** Prediction of ice-binding activity from homology models. (A) Sequence alignment of LTP1, LTP2, and template 1FK2. Marked residues correspond to predicted ice-binding sites in LTP1. Key substitutions are colored in blue. (B) Structural models of LTP1 and LTP2, visualized using PyMOL (http://www.pymol.org). Both structures have identical folds but differ in the degree of spatial regularity of the predicted IBS region. Several OSCs in LTP1 are labeled as the corresponding residue and the key substitution A44D is labeled on LTP2. The putative IBS of LTP1 is perpendicular to the plane of this figure. (C) Antifreeze activity of recombinant LTPs by ice-crystal morphology assay. Ice-crystals in LTP1 solution (labeled LTP1) initially grew as hexagons (left), and developed into six-pointed stars (right) as the temperature was cooled and warmed in very slow cycles. No antifreeze activity was observed for the control sample or LTP2 as evidenced by round, flat ice-crystals (labeled LTP2 and Ctrl). Incubation with proteinase K eliminated LTP1 antifreeze activity. Scale bars represent 10μm.

Interestingly, LTP1 is cysteine-rich and contains four disulfide bridges. High disulfide content is a characteristic of several AFP folds including β-helical insect and type II AFPs and has been proposed to provide a rigid conformation for enhanced stability at low temperatures (Gauthier et al., 1998). Since the LTP sequences were identified in cold-acclimated winter rye, ice-binding is quite likely related to the biological function of LTP1 and may be an important feature of some other LTPs. For example, the LTP homolog cryoprotectin is known to be cold-induced and is thought to prevent rupture of chloroplast thylakoids at freezing temperature (Sieg et al., 1996; Hincha et al., 2001). Therefore, the ability to bind ice and affect crystal growth may not be specific to LTP1, but may also offer insight into the function of additional members of the LTP family.

2.1.5 Conclusion and Discussion

As in the case of AFPs, some protein functional classes can be diverse in sequence, fold, and even mechanism. Thus, homology-based methods of functional annotation do not work in these cases, and a physicochemical surface pattern may better represent protein function. Using this approach, this work has computationally determined a set of features unique to AFPs from fish, insects, and plants, which may provide valuable insight into the nature of the AFP-ice interaction. Since the algorithm is not limited to particular vector orientations and surface geometries, it applies to a wide range of AFPs with different ice-binding surfaces, binding orientations, ice-plane specificities, and resulting ice-crystal growth morphologies. Future adaptations of the algorithm that include more detailed analysis of vector-geometry may prove useful in detecting additional details of ice-binding such as these. Outside the realm of AFP research, the general approach of vector-based surface analysis may be extended to develop and test hypotheses regarding structure-function relationships of other protein classes.

2.2 Glycan-binding proteins

2.2.1 Introduction

Structure-based algorithms are a promising approach for prediction and analysis of protein binding sites from first principles. Just as sequence motifs and domains may be used to infer function in uncharacterized sequences, structural patterns may be used to infer function in incompletely characterized structures (Jones and Thornton, 2004). For obvious reasons, 3D information can be more indicative than primary sequence alone of the underlying molecular mechanisms by which proteins function.

In the previous section, a surface pattern detection approach was developed to unify existing antifreeze or ice-binding proteins based on common physicochemical surface patterns despite their remarkably diverse sequences and folds. A structural pattern was used to identify a novel ice-binding protein that could not have been detected without analysis of 3D information. Glycan-binding proteins (GBPs) are another class of proteins that encompass a wide variety of structures with different sequences, structures, binding sites, and evolutionary histories (Boraston et al., 2004). Thus, while a sequence-based classification system such as the CAZY database (Cantarel et al., 2009) is useful for identifying additional homologs of known GBP families, sequence similarity alone cannot predict unrelated GBPs with novel binding sites and mechanisms, or unify GBPs from different evolutionary families into one common functional group.

A commonly studied class of sugar-binding proteins, of which GBPs are a subset, are carbohydrate-binding modules (CBMs). These are modular domains attached to a variety of carbohydrate-metabolizing enzymes that impart carbohydrate-binding activity. Glycans, which are oligosaccharides or polysaccharides, are substrates of type A (insoluble-polysaccharide binding) and B (glycan-chain binding) CBMs, respectively, while type C CBMs bind monosaccharides and small sugars. Several structure-based algorithms have been developed for predicting carbohydrate-binding sites in proteins (Taroni et al., 2000; Shionyu-Mitsuyama et al., 2003; Malik and Ahmad, 2007). These studies have predominantly focused on small sugar binding sites like those in type C CBMs, and have achieved reasonable prediction accuracy (up to 65%). However, there are few if any studies that focus on type A or B CBMs that bind crystalline polysaccharide surfaces and larger glycan chains. Because these types of CBMs bind to chains or large, crystalline carbohydrate surfaces,

their binding mechanisms are relatively unique compared to those of small-sugar binding CBMs, and so their binding sites are also thought to have unique properties (Boraston et al., 2004). Type A CBM binding sites often possess three surface aromatic residues whose side-chains are arranged in a coplanar orientation. Type B CBM binding sites typically contain two aromatic residues in either a coplanar or sandwich (clamp-like) orientation. The most common residues involved in carbohydrate-surface binding are tryptophan and tyrosine, with phenylalanine and histidine occurring to a lesser degree. Mutations of these residues have been shown to significantly decrease carbohydrate-binding activities (Bray et al., 1996; McLean et al., 2000; Uchiyama et al., 2001; Lehtio et al., 2003). This type of structural information could be used as search criteria in a structure-based method for predicting carbohydrate-binding sites, and predicting novel carbohydrate-binding proteins such as GBPs.

In this study, an algorithm has been developed that searches for coplanar aromatic motifs in 3D structures. It is shown that this simple 3D pattern can effectively distinguish a large number of different GBPs and their binding sites from random structures. The algorithm was then used to screen the PDB for novel binding sites outside of the training set, which identified a striking number of additional (and novel) binding sites. One of these predictions was tested experimentally through a cellulose-pulldown assay and mass spectrometry analysis, which verified the predicted binding activity. The method is fast, sensitive and, most importantly, widely applicable to proteins of any structure.

2.2.2 Materials and Methods

Datasets

Type A and B CBMs: Structures of six representative, non-homologous type A CBMs and 12 type B CBMs were retrieved from the Protein Data Bank (PDB), yielding a total of 17 unique protein folds. Type A CBMs included 1cbh (CBM1), 1exg (CBM2), 1nbc (CBM3), 1aiw (CBM5), 1e8r (CBM10), and 2bh0, a recently solved structure of a bacterial expansin possessing a binding site and binding activity analogous to type A CBMs (Kerff et al., 2008). Type B CBMs included 2xbd (CBM2), 1gui (CBM4), 1uxx (CBM6), 1gny (CBM15), 1j84 (CBM17), 1ac0 (CBM20), 1dyo (CBM22), 1of4 (CBM27), 1uww (CBM28), 1gwk (CBM29), 1uh2 (CBM34), and 1ux7 (CBM36). These structures were selected based

on the scheme presented in Boraston et al. (2004), a comprehensive review on CBMs.

Nh3d reference dataset: The Nh3d version 3 dataset (Thiruv et al., 2005) was retrieved for use as the "background" dataset. This dataset contains 806 structurally dissimilar protein chains from the PDB, and was built using the hierarchical CATH database classification. Nh3d was chosen because it was carefully constructed to contain structurally dissimilar protein chains without recognizable common ancestry, and so it lends itself to statistical, structural analysis. In addition, none of the proteins in this dataset are closely related (as determined through BLASTp searches) to the sequences in the type A and B CBM dataset.

nrPDB: For the purposes of screening, a large non-redundant database of 15,970 PDB structures was generated by retrieving a precomputed list of structures from the PISCES server (Wang and Dunbrack, 2003). The percentage identity cutoff was 90%, the resolution cutoff was 3.0 Å, and the R-factor cutoff was 1.0. Homologs of the type A and B CBM structures were not included in this dataset.

Algorithm and linear discriminant analysis

Based on previous literature, it was hypothesized that surface clusters of coplanar aromatic residues could be used as the basis of glycan-binding site identification. The analysis was performed using a combination of Perl and R programs, which are available upon request. The prediction method was developed in the following two stages.

1) Aromatic residue pairs were selected within known structures from the Type A and B CBM dataset based on previous literature and manual inspection. For type B CBMs, one pair of aromatic residues was used to describe a single ligand-binding site, and for type A CBMs composed of three aromatic residues, the two pairs with the shortest inter-residue distance were used. As a comparison "background" dataset containing random protein structures, the Nh3D dataset was used. All aromatic pairs passing initial parameter thresholds (identical to that later described in *Screening*) were selected, and assumed as non-binding-sites for algorithm training.

2) Key features (coplanarity, residue type, solvent-accessibility, and distance) were computed for all pairs of aromatic residues (Trp, Tyr, Phe, His) and used in a linear discriminant analysis (LDA) in an attempt to discriminate the known sites from the random sites

(background). LDA generates a scoring function based on a linear combination of the input variables that best separates the two classes of data. A linear discriminant function can be represented as shown below in Equation 2.1.

Equation 2.1:

$$D_i = \beta_1 Z_{i1} + \beta_2 Z_{i2} + ... + \beta_p Z_{ip}$$

Here, Z_{ip} are the values of the discriminating variables (standardized); B_p are the coefficients (standardized); and D_i is the score for the ith data point (in this case, an aromatic pair). The raw LDA score was used in subsequent searches for novel glycan-binding sites. The following features were used to train the parameters in LDA:

Coplanarity: measured as the angle (0 to 90 degrees) between the normal vectors of both aromatic rings.

Residue type: four parameters indicating the frequency of each aromatic residue type (Trp, Tyr, Phe, His). This value therefore ranges from 0 to 2 for each parameter.

Distance: the Euclidean distance, $\sqrt{(x2-x1)^2+(y2-y1)^2+(z2-z1)^2}$, between the centroids $[(x1, y1, z1)$ and $(x2, y2, z2)]$ of each aromatic ring.

Solvent-accessibility: the solvent accessible surface area (SAS) was calculated from a Voronoi tessellation (McConkey et al., 2002), and the average SAS of the two aromatic residues is used as a measure.

Screening

In the screening phase, a separate database of 15,970 non-redundant structures (nrPDB) was screened for potential binding-sites not included in the training set. This involved two steps:

1) Thresholding: aromatic pairs with feature values outside the allowed range were removed. The allowed range was simply defined based on the minimum and/or maximum values observed for known binding sites in the training set (Distance $\geq$ 6.16 Å, Distance $\leq$ 20.85 Å, Solvent-accessibility $\geq$ 33.34 Å^2, Coplanarity $\leq$ 83.55 degrees). The SAS cutoff, for instance, removed internal aromatic residues incapable of forming external interactions.

2) Scoring: The LDA scoring function was used to score all remaining candidate binding sites. The score of any aromatic pair can be compared to the "background" distribution of scores, which reflects its potential for being a binding site.

Gene Ontology (GO) analysis

PDB GO annotations were downloaded from the Gene Ontology Annotation database of the European Bioinformatics Institute (http://www.ebi.ac.uk/GOA). For structures identified by screening, GO term enrichment was tested for several GO categories related to carbohydrate-binding and metabolism. A binomial exact test was used to compute p-values.

Cellulose-binding assay

NOTE: The following experimental procedure was performed by Zhenyu Cheng. An interesting prediction (tobacco pathogenesis-related protein, PR-5d) was investigated experimentally through a cellulose-binding pulldown assay and mass spectrometry analysis. To purify tobacco proteins with insoluble-cellulose binding (type A CBM) activity, protein was isolated from tobacco roots, passed through a column containing microcrystalline cellulose (Sigma), run on a SDS polyacrylamide gel, and the strongest band was excised and identified by mass spectrometry. The experimental protocol is based on that in Doxey et al. (2008a), using cellulose instead of chitin.

Phylogenetic and sequence analysis of PR-5d

A BLASTp search of tobacco PR-5d (PDB ID 1aun) was used to identify related sequences. All sequences with $E < 0.001$ that aligned to the query with sequence coverage $> 90\%$ were used to build a second alignment using MUSCLE. Conserved regions of the alignment were used to generate a neighbor-joining tree using Seaview. A major clade of PR-5d proteins containing the PR-5d subclade was then selected for further analysis.

2.2.3 Results and Discussion

Discrimination of binding sites in type A and B CBMs based on features of their coplanar aromatic motifs

Twenty-six pairs of aromatic residues in the set of known type A and B CBM binding sites and 140,383 pairs found in 649/806 structures from the reference Nh3d dataset were analyzed. This set of 140,383 random pairs was further filtered to 6,254 pairs by selecting only those whose features were no worse than the worst-case values in the type A and B CBM training set (see thresholding section in Methods). This limited the set of cases from the Nh3d dataset to those with feature values within the same range as the known type A and B CBM cases, which improved later predictions.

The positive cases used to train algorithm parameters are shown in Figure 2.7A. These included 18 different type A and B CBM structures, each with distinct binding sites composed of two or more coplanar aromatic sidechains. As shown for the structures that contain bound ligands, coplanar aromatic motifs are oriented to dock on top of glycan-chains ("aromatic strip" motif) or around glycan-chains like a clamp ("aromatic sandwich" motif). As type A CBMs bind insoluble polysaccharides, the ligands are too large for co-crystallization and are thus not present in the stuctures, but the binding sites have been identified based on indirect methods such as mutagenesis experiments.

LDA was then used in an attempt to separate the positive cases (aromatic motifs from type A and B CBM binding sites) from background cases (aromatic pairs from the Nh3d dataset). As shown in Figure 2.7B, which displays the distributions of raw LDA scores, LDA was able to effectively separate the two classes based on the input features. 17/26 (65%) of the positive cases scored greater than 95% of background scores, and 24/26 (92%) scored greater than 90% of the background scores. This is likely an underestimate of the true degree of discrimination as the background dataset is expected to contain false negatives (i.e., coplanar aromatic motifs involved in glycan binding that have not been annotated). The positive case receiving the lowest score was aromatic pair (His 57, Trp 118) in structure 1NBC (middle residue, left residue, respectively), which makes sense as this is the most distant pair of the positive cases and contains a His residue, which has lower sugar-binding affinity than Trp and Tyr (Figure 2.7A). The positive case receiving the highest score was aromatic pair (Trp 61, Trp 102) in structure 1GUI, which forms

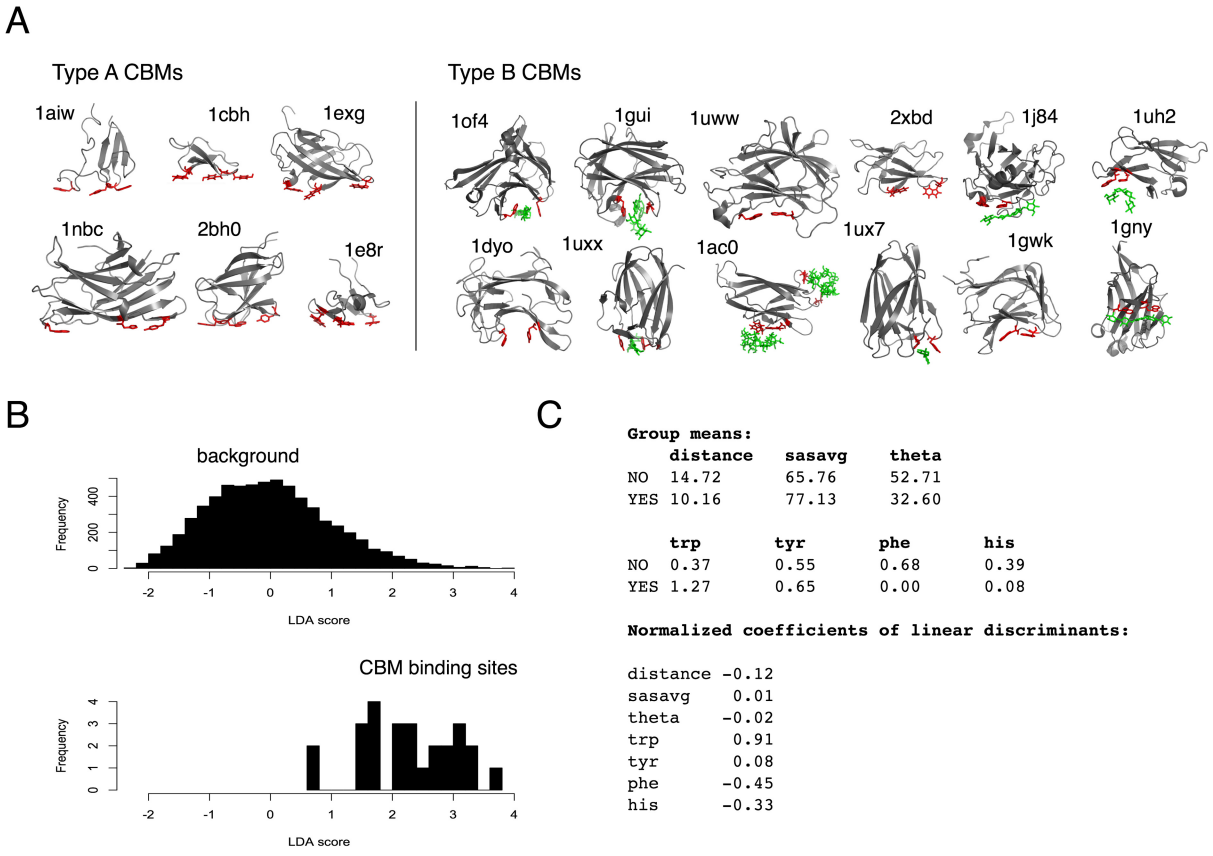


Figure 2.7: **Discrimination of glycan-binding aromatic motifs in type A and B CBMs from the reference Nh3d dataset.** (A) Structures of 6 type A and 12 type B CBMs and their glycan-binding aromatic motifs (highlighted in red) and ligands (highlighted in green if present in the PDB structure). (B) LDA-based discrimination of selected aromatic motifs in type A and B CBMs from aromatic pairs from the Nh3d dataset that passed the initial thresholds. Statistics regarding LDA are shown in (C).

a highly coplanar, accessible "aromatic sandwich" motif composed entirely of Trp, the optimal sugar-binding residue.

The normalized coefficients of linear discriminants, which provide an indication of the relative contribution and relationship of each variable to the discrimination, are shown in Figure 2.7C. The signs of each coefficient are as expected; the distance parameter is negative indicating a preference for aromatic pairs in closer proximity; solvent accessibility (sasavg) is positive indicating a preference for greater exposure to solvent; theta is negative indicating that lower angles (higher coplanarity) is preferred; and Trp and Tyr are positive

while Phe and His are negative indicating the preferences for Trp and Tyr residues. The residue-type parameter Trp had the largest coefficient magnitude (0.91). Thus, the type of aromatic residue appears to be the strongest discriminating variable between aromatic motifs found in glycan-binding sites and random, aromatic surface motifs. While the coefficients for coplanarity and solvent-accessibility are smaller, subtle differences in these parameters become increasingly important in the high end of the distribution. That is, Trp/Trp pairs (the best residue combination) with high coplanarity will be scored higher than those with low coplanarity, which may be important in screening for novel glycan-binding sites.

PDB screening

The algorithm was applied to 15,970 non-redundant structures from the nrPDB dataset, which resulted in 167,789 aromatic pairs from 14,723 different structures that passed the initial parameter thresholds. These were screened for pairs that received a raw LDA score greater than the 99th percentile score ($\sim$2.67) of all sites analyzed in the initial training set. This resulted in 1,054 high-scoring aromatic pairs from a total of 854 unique structures.

To determine whether the proteins identified by screening are enriched in carbohydrate-related functions, a statistical test was performed for enrichment of the GO terms for carbohydrate metabolism (GO:0005975) and carbohydrate-binding (GO:0030246). Of 380 structures in the nrPDB annotated with GO:0005975 (carbohydrate metabolism), 84 of these were identified in the screen. Of 48 structures in the nrPDB that contain the term GO:0030246 (carbohydrate-binding), 22 of these were identified. This is equivalent to an over four-fold enrichment ($p < 1e\text{-}15$), and over eight-fold enrichment ($p < 1e\text{-}13$) in carbohydrate-metabolism and binding functions, respectively.

Diversity of glycan-binding folds and binding sites identified in the PDB screen

Shown in Figure 2.8 are a structurally diverse set of proteins and their predicted binding sites identified in the PDB screen. An astounding structural diversity of binding sites and folds were detected, far beyond what can be detected by a sequence database search. While the training set contains predominantly beta-rich structures (Figure 2.7A), the structures

identified in the screen were of all types including all-α, all-β and mixed α-β structures, corresponding to a wide variety of folds, superfamilies, and families.

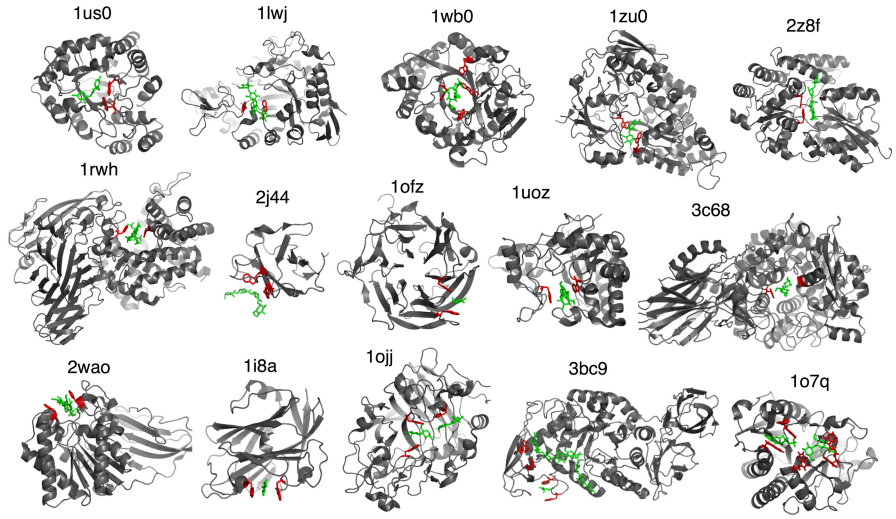
Fifteen examples of unique folds and binding sites for which the predicted coplanar aromatic residues are contacting a bound glycan, are shown in Figure 2.8 (upper panel). Some well defined fold classes of the proteins shown in Figure 2.8 include the 7-stranded beta/alpha barrel (1uoz), 6-bladed beta-propeller (1ofz), periplasmic binding protein-like II (1z8f, 1zu0), starch-binding domain-like (2j44), alpha/alpha toroid (3c68), and TIM barrel (e.g., 1us0, 1lwj, 1wb0). The TIM barrel, the second most common fold in proteins after immunoglobulins (Söding et al., 2006), is a recurring structure for many of the identified proteins. Binding sites were predicted in a wide variety of TIM barrel families, including many beta-glycanases (SCOP ID 51487), amylases (51466), type II chitinases (51534), beta-N-acetylhexosaminidases (51550), NagZ-like family (51553), Aldo-keto reductases (51431), and Xylose isomerases (51665). In TIM barrels and other fold families, the detected binding sites are diverse in terms of their spatial and sequence positions, but commonly clustered near or within active sites, or on adjacent binding domains.

Structural genomics predictions

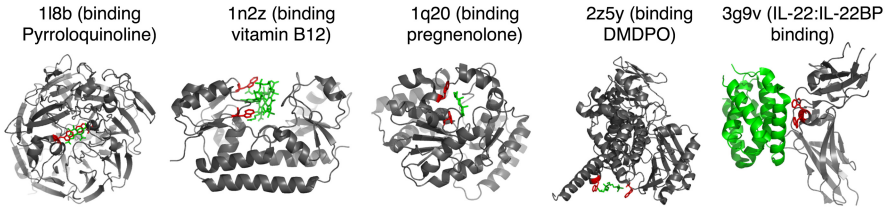
While the correctly detected binding sites (Figure 2.8, upper panel) provide support for the method, the most interesting cases detected in the PDB screen are binding sites that are previously uncharacterized and are thus potentially novel cases. The best examples of these are proteins whose structures have been determined by large-scale Structural Genomics consortiums, but whose functions are unclear or completely unknown. A striking number of binding sites in Structural Genomics proteins were identified, 20 of which are shown in Figure 2.8 (bottom panel). These binding-site predictions (see Table 2.2) provide starting points to guide future experimental work on these and any related proteins sharing the detected motif. While the detected binding sites suggest a role in carbohydrate-binding, as described in the next few sections, some of the detected sites likely bind additional types of ligands.

Figure 2.8 *(following page)*: **Selected binding-site predictions obtained by PDB-screening.** Previously characterized binding sites (top panel, middle panel) as well as novel predictions for structures of unknown function (bottom panel) are included.

Example binding site predictions in structures with bound carbohydrates



Other types of ligand-binding sites detected by the algorithm



Predicted binding sites in structural genomics proteins of unknown function

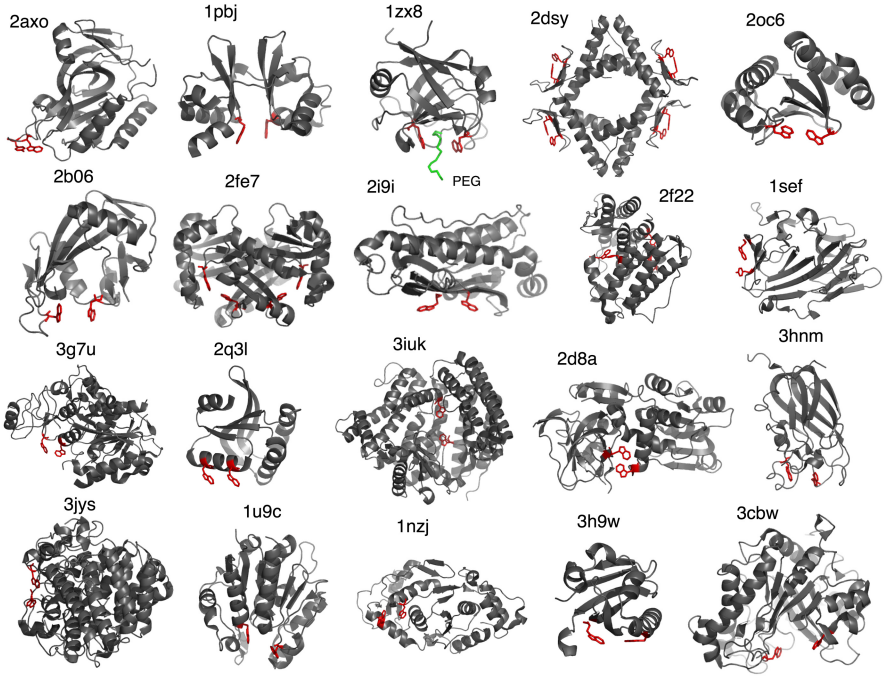


Table 2.2: **Predicted glycan-binding aromatic motifs in 20 structural genomics proteins.**

PDB ID	Aromatic Pair
2AXO	(Y85, W88)
1PBJ	(W50, W97)
1ZX8	(W39, W69)
2DSY	(Y30, W40)
2OC6	(W40, W121)
2B06	(W35, W150)
2FE7	(W75, W143)
2I9I	(W146, W177)
2F22	(Y8, W141)
1SEF	(W222, Y242)
3G7U	(W206, W339)
2Q3L	(W98, W105)
3IUK	(W414, W507)
2D8A	(W137, W297)
3HNM	(W39, W57)
3JYS	(W100, W327)
1U9C	(W24, W187)
1NZJ	(W91, W210)
3H9W	(W71, W101)
3CBW	(W198, W328)

Additional ligand specificities predicted by coplanar aromatics

One of the most surprising results obtained from the PDB screen is the degree to which the algorithm recognized binding sites for additional types of ligands, mediated by coplanar aromatic motifs [Figure 2.8 (middle panel) and Figure 2.9]. Five examples are shown in Figure 2.8 (middle panel), including proteins binding important ligands as vitamin B12, and pregnenolone. Protein-protein interaction sites were also detected by screening for coplanar aromatics [e.g., as shown for 3g9v, interaction between interleukin 22 (IL2) and IL2 binding protein)].

Another commonly identified interaction among the top-scoring cases was protein-nucleotide binding. As shown in Figure 2.9, numerous aromatic strip motifs identified by the algorithm are base-recognition motifs. For example, a coplanar aromatic motif detected in the prokaryotic Tn5 transposase (1mus, Figure 2.9) mediates an interaction with

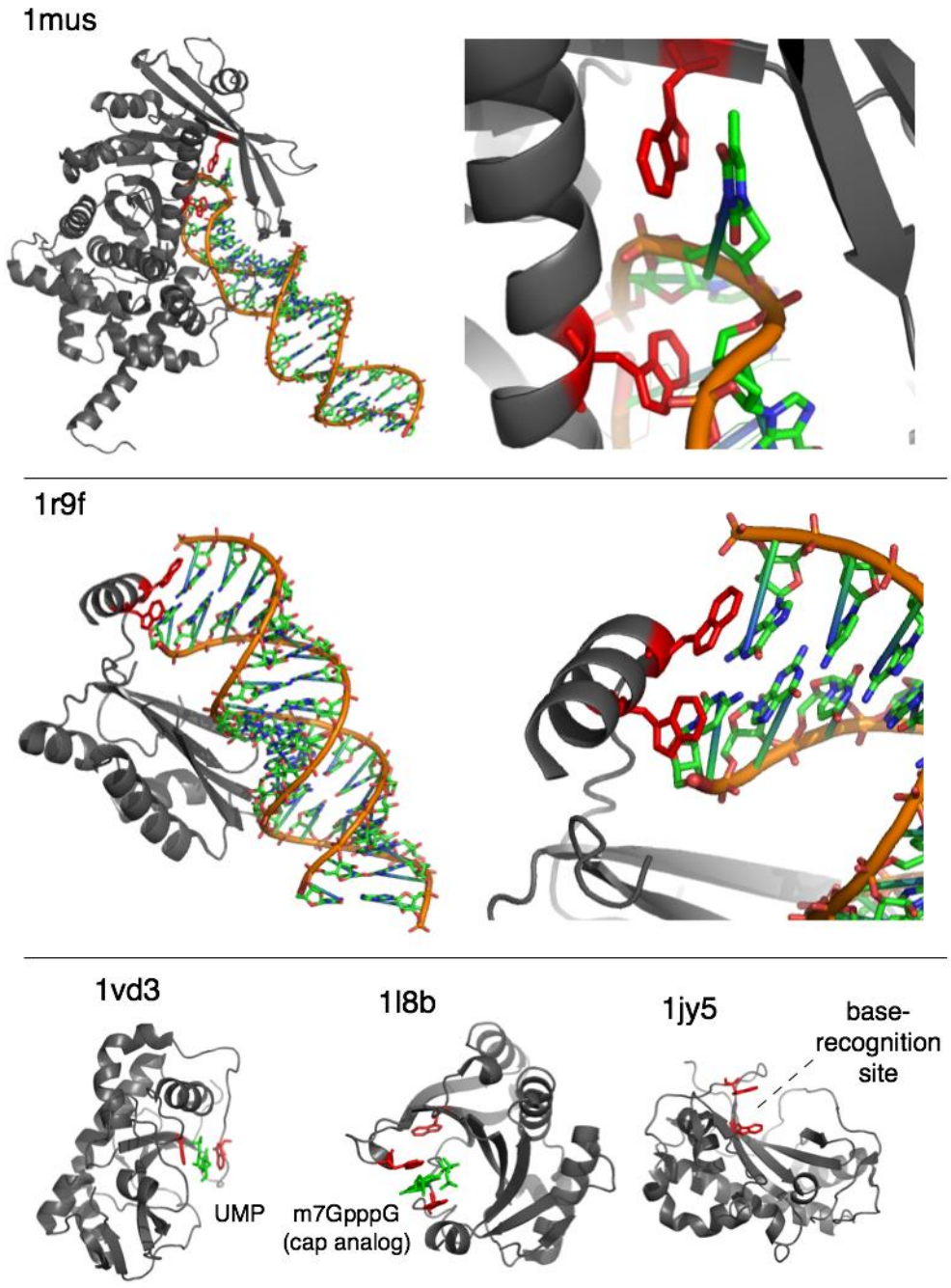


Figure 2.9: **Predicted binding-sites involved in DNA/nucleotide-recognition.**

the "flipped-out base" which is critical for hairpin formation, hairpin resolution and strand transfer in the process of transposition (Steiniger-White et al., 2004). In the structure, 1r9f, an siRNA involved in gene-silencing in response to viral infection, is bound by a viral suppressor of gene-silencing. The detected aromatic motif "brackets" the terminal base pairs at the end of the siRNA duplex, which contributes to the unique function of this protein (Ye et al., 2003). Lastly, three other correctly predicted nucleotide-binding sites (UMP-binding site, 5' cap-binding site, and an unbound base recognition site) are also shown (Figure 2.9, bottom panel).

Prediction of a novel binding site in PR-5d

While the presented results have focused largely on detected true positives that contain bound ligands (with the exception of the structural genomics proteins), most of the predicted binding sites are unbound, and have not been identified by previous studies. It is important to note, however, that because coplanar aromatic motifs can occur by chance, many of these are likely false-positives, and so careful examination of each prediction is necessary.

One prediction that was believed to be novel, testable and of potential biological importance, was a predicted binding site in tobacco pathogenesis-related family 5 protein (PR-5d, PDB ID 1aun), also known as neutral-osmotin like protein and thaumatin-like protein. The biological function of PR-5d is known to be involved in pathogen-resistance, but its mechanism is not clear (Koiwa et al., 1997; Koiwa et al., 1999; Osmond et al., 2001). The predicted binding site in PR-5d is interesting for a number of reasons. Three Trp residues (Trp 34, Trp 36, Trp 196) were predicted, and form a highly coplanar and accessible putative binding site on the surface of the protein (Figure 2.10A), which is highly reminiscent of the aromatic motifs in the binding sites of type A CBMs (insoluble-polysaccharide binding proteins) (Figure 2.7A). The putative binding site in PR-5d received a raw LDA score of 2.95, which placed it higher than 99.7% of all other proteins in the nrPDB, and even higher than 20/26 of the positive cases in the training set (Figure 2.7).

The binding activity of PR-5d towards insoluble cellulose was tested using a cellulose-pulldown assay of tobacco proteins. This assay purifies tobacco proteins that bind to insoluble-cellulose, which are then washed, run on a gel, and identified by mass spectrometry. A band at $\sim$23 kDa was excised and identified by mass spectrometry, which resulted in

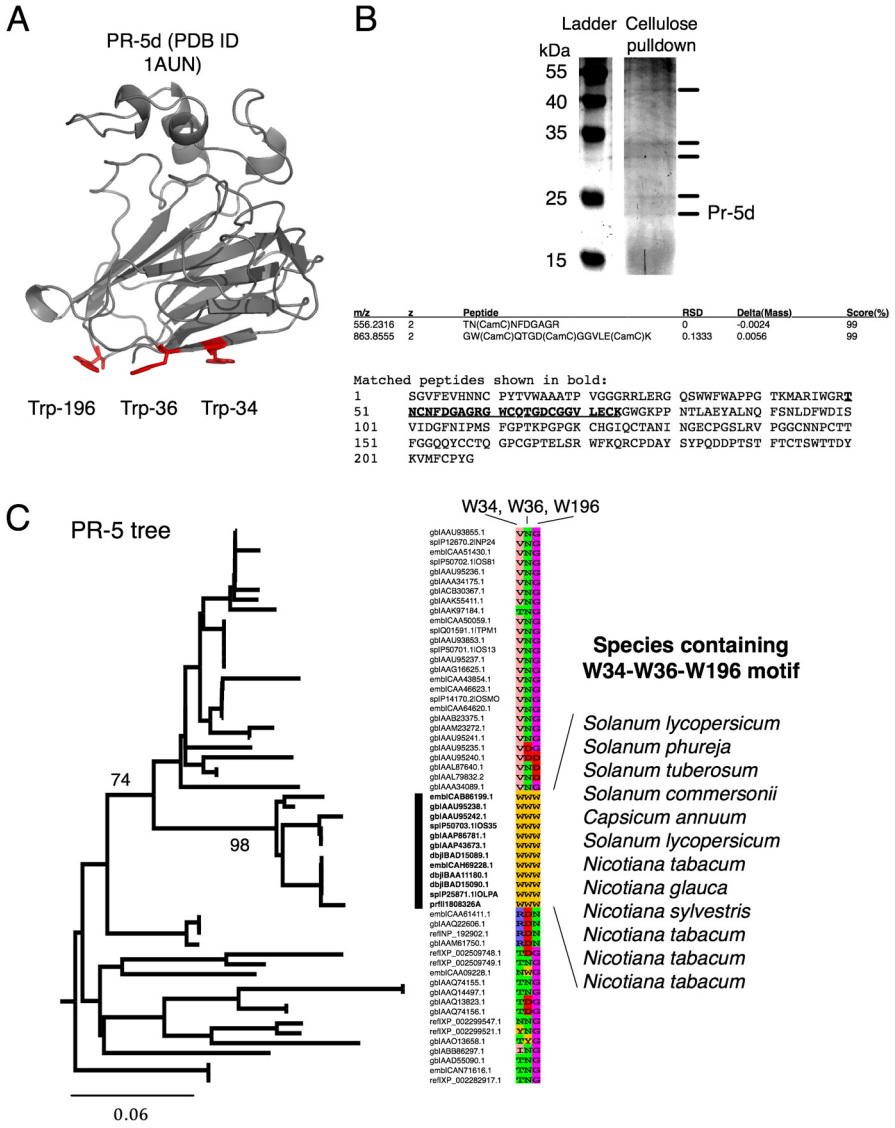


Figure 2.10: **Prediction, analysis, and testing of a putative binding-site in tobacco PR-5d.** (A) Predicted binding site in Pr-5d (PDB ID 1aun), highlighted in red. (B) A coomassie gel of purified cellulose-binding proteins from tobacco. The marked was excised and identified by mass spectrometry (MS). MS sequencing identified two peptides matching the PR-5d sequence. (C) A neighbor-joining phylogenetic tree of PR-5d and related proteins. Bootstrap values are indicated for two key clades. The residues in alignment positions 34, 36, and 196 (relative to the PR-5d sequence) have been mapped onto the tree, and demonstrate that the putative binding site (Trp 34, Trp 36, Trp 196) likely originated in an ancestral PR-5d protein in an ancient species of the Solanaceae family.

the identification of peptides matching PR-5d (Figure 2.10B). The identification of PR-5d in the cellulose-binding fraction demonstrates that PR-5d has insoluble-cellulose-binding activity and validates the computational prediction.

A BLAST search and phylogenetic analysis revealed that PR-5d is a member of a highly conserved clade of PR-5 proteins exclusive to the Solanaceae family of plants including tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), chili pepper (*Capsicum annuum*), and several species of tobacco (Figure 2.10C). This clade of PR-5d proteins, which is extremely well supported by a bootstrap value of 98/100, all share the Trp 34, Trp 36, Trp 196 motif, but it was not found elsewhere in the phylogeny (Figure 2.10C). Thus, there is clearly a selective pressure for conservation of this motif in these Solanaceae PR-5d proteins.

Proposed function of PR-5d

Solanaceae plants are common targets of the deadly plant pathogen, *Phytophthora* (from Greek phytón, "plant" and phthorá, "destruction"; "the plant-destroyer"). The evolved, putative binding site towards insoluble-polysaccharides in the plant-defense protein, PR-5d, may be an evolutionary adaptation towards defense against *Phytophthora*. Unlike fungi, which are commonly studied targets of PR-5 proteins, the distinguishing feature of *Phytophthora infestans* and other oomycetes is that they possess a cell wall made of insoluble cellulose (Zevenhuizen and Bartnicki-Garcia, 1969). Previous studies have indeed shown that PR-5 related proteins are active against *Phytophthora infestans* (Woloshuk et al., 1991). Future work mutating the WWW motif and observing the *in vitro* and *in vivo* effect on *Phytophthora* invasion and infection would help elucidate this intriguing story.

2.2.4 Summary

Glycan-binding sites in proteins are of critical importance in a wide variety of biological processes including host-pathogen interactions, immune responses, and carbohydrate metabolism. However, glycan-binding and other carbohydrate-binding proteins cannot be unified based on a common sequence pattern. Through this work, it has been shown that a recurring 3D surface pattern (coplanar aromatics) can be used to effectively distinguish many known and potentially novel GBPs from other proteins. This demonstrates the

usefulness of going beyond a sequence-based classification, and incorporating a recurring molecular trait that is a direct determinant of binding activity. By performing a large-scale screen of the PDB using the 3D pattern, numerous putative glycan-binding sites were found, as well as an unexpected number of different protein-ligand interactions that possess a similar aromatic-mediated binding mechanism (e.g., nucleotide-binding sites), and numerous structural genomics proteins of unknown function. One novel predicted binding site found in the PR-5d protein was validated experimentally using a cellulose pulldown assay. The predicted binding site in this protein is perfectly conserved in related PR-5d proteins, and yields new insights into its function. Ultimately, the results suggest that coplanar aromatic surface motifs are a highly useful structural signature for classifying existing and identifying novel GBPs, and other types of protein-ligand and protein-protein binding sites. The ability of the algorithm to identify numerous types of protein-ligand interactions highlights its generality and is thus not necessarily a weakness but a strength. Evolution has reinvented the coplanar aromatic motif by convergent evolution many times to bind glycans, nucleotides and many other types of ligands.

Chapter 3

Gene expression divergence

The following chapter includes published material from Doxey et al. (2007)[1].

The examples presented in Chapter 2 demonstrate how similar function-determining patterns can arise in unrelated structures by mutations of their protein surfaces. However, surface mutations cannot always account for the full breadth of functional diversity within protein families. In this section, I explore the functional diversity within a large gene family (β-1,3-glucanases) from *Arabidopsis thaliana* with approximately 50 members. This family shares significant overall sequence similarity and appears to have a conserved enzymatic function as indicated by a highly conserved active site. Thus, an understanding of how function has evolved and diversified within this family requires investigation of a different molecular determinant of function—gene expression. Through phylogenetic mapping of microarray data, it is revealed how gene expression patterns can evolve independently of protein sequence, and contribute to functional divergence. By examining expression divergence in the context of a phylogeny, and identifying expression patterns that have diverged or converged on similar patterns found in other unrelated proteins, it is possible to functionally annotate many additional family members whose functions are unknown.

[1]The included article is "Functional divergence in the *Arabidopsis* β-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states", *Molecular Biology and Evolution*, 2007, 24, 4, 1045-55, by permission of Oxford University Press.

3.1 Introduction

β-1,3-glucanases (glucan endo-1,3-glucosidases, β-1,3-Gs, E.C. 3.2.1.39) are a class of hydrolytic enzymes that catalyze the cleavage of 1,3-β-D-glucosidic linkages in β-1,3-glucans, and are found in bacteria, fungi, metazoa (Bachman and McClay, 1996) and viruses (Sun et al., 2000). β-1,3-glucanases form highly complex and diverse gene families in plants, where they are involved in a wide range of physiological and developmental processes (Jin et al., 1999; Leubner-Metzger and Meins, 1999). β-1,3-Gs have received a considerable amount of attention due to their role in plant pathogen-defense. As members of the PR-2 group of pathogenesis-related (PR) proteins, β-1,3-Gs are induced by pathogen infection and play an active antifungal role in hydrolyzing β-1,3-glucan, a major structural component of fungal cell walls (Leubner-Metzger and Meins, 1999). The degradation of β-1,3-glucan in cell walls is thought to contribute towards fungal cell wall destabilization as well as release of cell wall associated immune elicitors that further stimulate defense responses (Leubner-Metzger and Meins, 1999). Antifungal activity has been confirmed both *in vitro* (Sela-Buurlage et al., 1993) and *in vivo* (Jach et al., 1995) through genetic transformation studies. Another intriguing role of β-1,3-Gs associated with PR function is their involvement in cold response (Griffith and Yaish, 2004; Yaish et al., 2006). While research on β-1,3-Gs to date has focused primarily on their PR functions, β-1,3-Gs also play critical roles in normal developmental plant processes. Callose (plant β-1,3-glucan) functions as a transitory material in the cell-plate during cell division (Fulcher et al., 1976; Longly and Waterkeyn, 1977; Samuels et al., 1995), as a major component of pollen mother cell walls, pollen tubes, plasmodesmatal canals (Stone and Clarke, 1992), and seed-covering structures of several dicot species (Leubner-Metzger, 2003). Thus, in addition to their roles in pathogen-defense, β-1,3-Gs have been implicated in cell division, pollen development and tube growth, regulation of plasmodesmata signaling, cold-response, seed-germination and maturation (see Hoj and Fincher, 1995; Leubner-Metzger and Meins, 1999; Leubner-Metzger, 2003, for reviews).

The diversity of β-1,3-G function is mirrored by its large gene family size (Jin et al., 1999). Previous attempts to classify β-1,3-Gs have used isoelectric point, sequence similarity, and other sequence features. As a result, a number of distinct sequence classes have been defined in tobacco (Linthorst et al., 1990), barley (Hoj and Fincher, 1995), soybean (Jin et al., 1999), and *Arabidopsis* (Dong et al., 1991; Hird et al., 1993; Delp and Palva,

1999). The biological significance of β-1,3-G gene multiplicity and the relationship between sequence and function, however, remains unclear. Ultimately, while a phylogenetic analysis based on sequence alone provides a means to infer the duplication history of a gene family, it does not provide an understanding of how function has diversified. Furthermore, as is the case for β-1,3-Gs, while a common enzymatic activity (i.e., molecular function) can be inferred by the conservation of active site residues in most cases, the mechanisms underlying functional divergence in terms of biological process is unclear.

Divergence of gene expression patterns is one mechanism through which proteins may diverge in function within a homologous family (Li et al., 2005). Through acquisition or loss of regulatory cis-elements, the expression profiles of duplicate genes may diverge independently of coding sequence allowing genes to acquire new or modified functions. Therefore, the incorporation of expression data into a phylogenetic analysis is crucial in the construction of a proper model of functional divergence.

Arabidopsis thaliana is an ideal model organism for the study of expression divergence, an area of considerable interest in current evolutionary studies (see Li et al., 2005 for review) for a number of reasons (Blanc and Wolfe, 2004). *Arabidopsis* has undergone a number of rounds of whole genome duplication, providing a large number of polyploid genes. Full genomic information is also available, so it is possible to obtain all members of a particular gene family. Lastly, publicly available *Arabidopsis* microarray resources provide a means to classify and functionally annotate genes based on their expression patterns.

Here, phylogenetic and expression analyses are integrated to develop a model of functional divergence of the β-1,3-glucanase gene family in *Arabidopsis*. Microarray expression data are used to cluster genes into different expression categories, and functions are attached to the expression clusters using knowledge of co-expressed genes as well as existing literature. The expression clusters are subsequently mapped onto a phylogenetic tree, and ancestral expression states are inferred using parsimony. Finally, a random permutation test is performed to compare the observed parsimony score against a random distribution from which a p-value is calculated. The combined analysis of protein sequence, expression, and phylogeny is demonstrated as a powerful approach for investigating functional divergence in large gene families.

3.2 Methods

Sequence Retrieval and Phylogenetic Analysis

A BLASTp (Altschul et al., 1997) search using a representative β-1,3-G sequence (AGI locus identifier: At3g57270) was performed to screen the *Arabidopsis* protein database accessible at TAIR (The *Arabidopsis* Information Resource, http://www.arabidopsis.org). The top fifty hits had E-values $\leq$ 1e-32, while subsequent hits had E-values $\geq$ 0.084. All of the top fifty sequences were annotated as or similar to *glycosyl hydrolase family 17* by TAIR. A number of additional sequences are incorrectly annotated as *glycosyl hydrolase family 17* by virtue of sharing homology with the C-terminal domain (Henrissat and Davies, 2000), and were not included the data set. The set of identified glucanases was verified by PSI-BLAST (Altschul et al., 1997) using the NCBI non-redundant database restricted to *Arabidopsis* with an inclusion threshold of E $\leq$ 1e-30. No additional glucanases were identified. A multiple alignment (Appendix 1) of the sequences encoding the conserved glycosyl hydrolase family 17 domain was constructed with MUSCLE v. 3.52 (Edgar, 2004), and gaps and poorly aligned sections were removed using Seaview (Galtier et al., 1996) prior to phylogenetic analysis. Using ProtTest version 1.3 (Abascal et al., 2005), the best fitting model for amino acid evolution was determined to be WAG including invariant sites (I) and a gamma correction for among-site rate variation (G) according to the Akaike Information Criterion (AIC) and all other available statistical frameworks. Using this model and its associated parameters (gamma shape and proportion of invariant sites), a Bayesian phylogenetic tree was constructed using MrBayes version 3 (Ronquist and Huelsenbeck, 2003). Convergence was reached after 60,000 generations, sampling trees every 10 generations. The first 1,500 trees were discarded as "burn in" and a majority-rule consensus tree was built with posterior probabilities calculated for each clade. Using the same model, maximum likelihood analysis was performed with PHYML (Guindon and Gascuel, 2003) and bootstrap values determined from a population of 100 replicates. A neighbor-joining tree was also generated and bootstrapped (1,000 replicates) using ClustalX (Chenna et al., 2003). Conserved introns were identified by mapping intron positions onto the protein sequence alignment (Appendix 1). For prediction of GPI-anchor attachment sites, the GPI-SOM algorithm was used (Fankhauser and Maser, 2005).

Expression Analysis

Microarray data were retrieved from two online microarray data repositories: the Botany Array Resource (BAR), accessible at http://bbc.botany.utoronto.ca/ (Toufighi et al., 2005), and Genevestigator, accessible at https://www.genevestigator.ethz.ch (Zimmermann et al., 2004). Four datasets were downloaded from BAR including the AtGenExpress tissue, stress, hormone, and pathogen series comprised of microarray data from a wide variety of experimental sources. The corresponding AtGenExpress IDs are listed in Appendix 2. The probeset to AGI ID lookup was a table from TAIR labeled affy25k_array_elements-2006-01-06.txt. More information on these microarray experiments can be found online at http://www.weigelworld.org. The Digital-Northern tool within Genevestigator was used to retrieve additional microarray data concerning response to ozone, salicylic acid, ethylene, and a number of fungal pathogens (Genevestigator experiment numbers 13, 25, 85, 108, 113, 146, 147, 161). For experiments involving multiple time-points, the Genevestigator protocol was followed and fold-change values were averaged into a single representative value for each gene.

All expression data have been previously collected using the ATH1 22K Affymetrix GeneChip (Hennig et al., 2003). Only experiments using wild-type plants were included, and all microarray data were processed in the same manner. First, in order to reduce background noise, data points less than 50 arbitrary fluorescence units were set to a threshold value of 50. Expression values were calculated as the ratio of the average of replicate treatments relative to the average of the appropriate controls. The choice of controls and the calculations for the four AtGenExpress datasets and the Genevestigator dataset were identical to those specified in the BAR resource and Genevestigator, respectively. The ratio expression data were $\log_2$ transformed, and each of the five datasets was separately clustered using a two-way hierarchical procedure (centroid linkage, centered/Pearson correlation). In order to define clusters, a clustering threshold was set to a Pearson correlation coefficient of 0.5. Cluster version 3 (Eisen et al., 1998) was used for microarray data clustering and results were visualized with Treeview version 1.012. Both Cluster and Treeview are available via http://rana.lbl.gov/EisenSoftware.htm.

Genes outside the β-1,3-G family with expression profiles correlated to β-1,3-G genes were identified using ATTED (*Arabidopsis thaliana* trans-factor and cis-element prediction database, http://www.atted.bio.titech.ac.jp/) which uses microarray expression data

obtained from the public AtGenExpress project (RIKEN and Max Planck). The ATTED database uses Pearson correlation coefficients to determine co-expressed genes. Co-expressed gene pairs were defined as those having $r \geq 0.8$.

Ancestral state reconstruction and statistical analysis

Ancestral state reconstruction and reshuffling was performed with Mesquite version 1.06 (Maddison and Maddison, 2005). The expression classes were discretized by treating each cluster as a separate character state. Ancestral states were then reconstructed by parsimony using an unordered model in which all state changes are treated equally. If the resulting distribution of expression states on the phylogenetic tree is non-random and exhibits conserved, clade-specific patterns, one would expect the parsimony score (number of state changes in the tree) to be significantly lower than that of a random dataset. Therefore, to evaluate the statistical likelihood of an observed parsimony score, the data were randomly permutated by reshuffling the character matrix 10,000 times and calculating the parsimony score for each reshuffled matrix. From the resulting distribution of parsimony scores, a p-value was calculated as the fraction of random scores less than or equal to the observed score.

3.3 Results

Sequence characterization and phylogenetic analysis

Fifty β-1,3-G sequences were identified as members of the *Arabidopsis* β-1,3-G family. All sequences contain an N-terminal signal peptide and a glycosyl hydrolase family 17 (β-1,3-G) domain. A variable C-terminal domain, first characterized as the X8 domain (Henrissat and Davies, 2000), is present in just over half (27) of the sequences. Recently, the X8 domain has been defined as a new class (CBM43) of carbohydrate binding modules responsible for binding β-1,3-glucan (Barral et al., 2005). A C-terminal hydrophobic sequence, which may encode a transient transmembrane domain involved in GPI-anchor attachment (Henrissat and Davies, 2000; Borner et al., 2002) or possibly a vacuolar targeting peptide in other cases (Leubner-Metzger and Meins, 1999), was found in 25/50 of the sequences. The presence/absence of these features define five protein domain architectural groups (Figure 3.1).

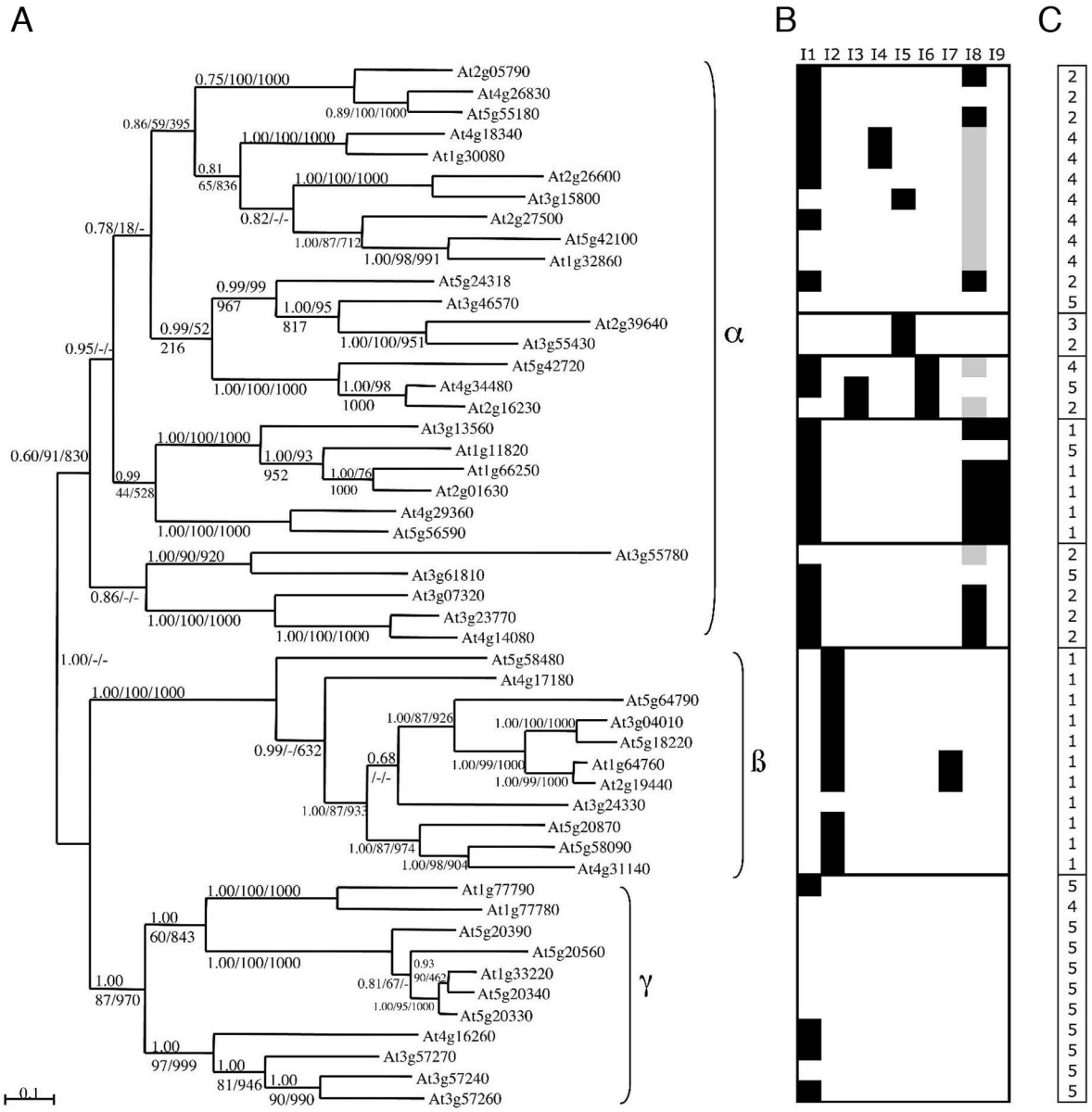
Figure 3.1: **Protein domain architectures observed in the *Arabidopsis* β-1,3-glucanase gene family.** The five domain architectural classes are based on the presence/absence of an N-terminal signal sequence (NTS), carbohydrate binding module family 43 (CBM43) and hydrophobic C-terminal sequence (CTS), in addition to the core glycosyl hydrolase family 17 domain (GH-17).

As shown in the multiple alignment (Appendix 1), the two catalytic glutamic residues, which correspond to E231 and E288 in barley isoenzyme GII (Varghese et al., 1994), are highly conserved (present in 49/50 and 46/50 glucanases, respectively). Other residues previously implicated in β-1,3-glucan binding and hydrolysis (Varghese et al., 1994) are also highly conserved (Appendix 1). Thus, β-1,3-glucan hydrolysis activity is likely a common

characteristic among nearly all putative β-1,3-Gs. Interestingly, several sequences with changes to highly conserved residues, including At3g55780 and At1g33220, also have weak expression levels and represent possible pseudogenes.

Maximum likelihood (ML), Bayesian, and neighbor joining (NJ) phylogenetic analysis (Figure 3.2A) partitioned the fifty β-1,3-G sequences into three major clades (denoted α, β, and γ). Generally, the tree topology is well supported by all three methods, with the exception of several higher order branches of clade α, which display lower ML and NJ bootstrap values. The best model chosen by Prottest was a WAG + I + G model under the AIC criterion, which produced a tree with likelihood -ln$L$ = 16410.06. In order to independently test the reliability of the tree topology, protein domain architecture and intron/exon structure were mapped onto the tree (Figure 3.2B). Conserved intron positions are located predominantly at the end of the N-terminal signal sequence, and at several different sites within the CBM43 domain (Appendix 1). The introns are highly clade-specific, with introns I2 and I7 present only in clade β, and I3, I4, I5, I6, I8, and I9 present only in clade α. The finer structure of the tree is also supported by the intron data, with a few minor exceptions. For example, the distribution of intron I5 would support a tree grouping At3g15800 closer to At2g39640 and At3g55430, though this is not supported by sequence data. Presence of the CBM43 domain and/or a hydrophobic C-terminus is also largely clade-dependent which reflects ancestral gains/losses of these sequence features (i.e., clade β lacks the CBM43 domain completely while all members of clade γ and over half the members of clade α contain the CBM43 domain). Since these features were not used in the construction of the tree, the clade-specific patterns of intron and protein domain conservation provides additional support for the proposed phylogeny.

Figure 3.2 *(following page)*: **Phylogenetic analysis of** *Arabidopsis* **β1,3-Gs.** (A) Majority-rule consensus tree generated by Bayesian inference of phylogeny. Bayesian posterior probabilities (/1), maximum likelihood support (/100) and neighbor-joining bootstrap values (/1000) are indicated above the clades where the clade is present in the respective tree. (B) Presence/absence of conserved introns and protein domain architecture mapped onto the phylogenetic tree. The locations of introns (labeled I1-I9) are shown in the multiple sequence alignment (Appendix 1). Black boxes indicate presence of introns, white indicates absence, and gray boxes indicate introns that are located in the C-terminal domain but do not align well with I8. (C) Protein domain architectural class as defined in Figure 3.1.


A
B
C
I1 I2 I3 I4 I5 I6 I7 I8 I9
At2g05790
At4g26830
At5g55180
At4g18340
At1g30080
At2g26600
At3g15800
At2g27500
At5g42100
At1g32860
At5g24318
At3g46570
At2g39640
At3g55430
At5g42720
At4g34480
At2g16230
At3g13560
At1g11820
At1g66250
At2g01630
At4g29360
At5g56590
At3g55780
At3g61810
At3g07320
At3g23770
At4g14080
At5g58480
At4g17180
At5g64790
At3g04010
At5g18220
At1g64760
At2g19440
At3g24330
At5g20870
At5g58090
At4g31140
At1g77790
At1g77780
At5g20390
At5g20560
At1g33220
At5g20340
At5g20330
At4g16260
At3g57270
At3g57240
At3g57260
α
ß
γ
0.1

Clustering of tissue and growth-stage expression data reveals distinct developmental classes

Presumably, the addition of gene expression data to phylogenetic analysis should provide a more complete model of functional divergence than can be achieved using sequence information alone. Clusters derived from microarray expression data tend to be significantly enriched for specific functional categories and thus can be used for inferring the function of unknown genes (D'haeseleer, 2005). For the purposes of expression clustering and functional assignment, the AtGenExpress developmental expression dataset (Schmid et al., 2005) was selected as a source of microarray data, which facilitates a classification of genes based on the patterns of gene expression in a range of tissues/organs and developmental stages. If the enzymatic function of the protein family is known (as is the case with β-1,3-Gs), spatial/temporal expression data for the corresponding gene should be particularly informative in determining its biological/physiological role.

Within the developmental dataset, the genes At1g11820 and At5g24318 were not present in the lookup table and At2g19440 and At1g64760 were recognized by the same probeset (267335_s_at) and were therefore excluded from subsequent analysis. At3g46570 and At1g77790 were also removed, as all expression values were beneath the noise floor threshold. The remaining 44 genes grouped into 13 expression clusters denoted A-M including 6 singleton clusters (Figure 3.3A). As demonstrated by the highly conserved developmental expression patterns (Figure 3.3A), a large portion of the β-1,3-G gene family appears to be under a set of several distinct developmental programs. Clustering of hormone, stress and pathogen response expression data was also performed, but a large number of genes formed singleton clusters (Figure 3.5) and, as discussed later, the clusters produced marginally significant or non-significant parsimony scores when mapped onto the phylogenetic tree. This may be attributed to the presence of highly divergent expression profiles or absence of phylogenetically relevant characteristics in the microarray data.

In addition to clustering gene expression profiles within the β-1,3-G family, identification of co-expressed genes outside of the β-1,3-G family may provide information useful for functional assignments. The ATTED database provides pairwise correlation of expression data for *Arabidopsis* genes using data obtained from the AtGenExpress project, and was used to identify genes significantly correlated (pairwise expression profiles with $r \geq 0.8$) with β-1,3-G genes (see Appendix 3). Using the known functions of correlated genes,

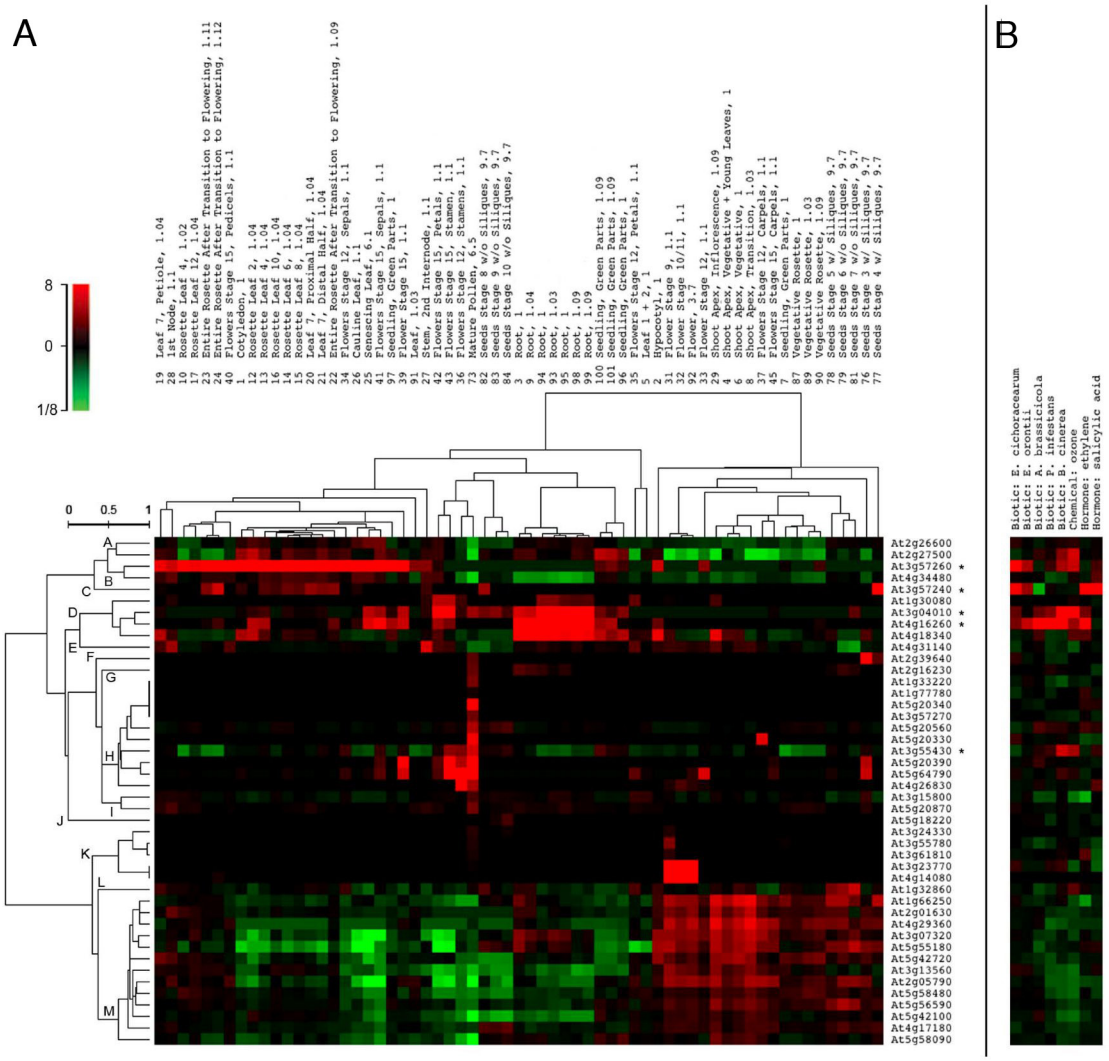


Figure 3.3: **Clustering of microarray expression data for genes encoding β-1,3-Gs.** (A) Developmental expression clusters (labeled A-M) were assigned using Pearson correlation coefficients with a centroid correlation threshold of 0.5. The experiments are labeled according to their AtGenExpress sample ID, tissue, Boyes growth stage (Boyes et al., 2001), and mutant genotype where applicable. Stress response expression data from fungal pathogen treatments and selected chemical stresses are also shown in (B) but were not used in clustering the developmental dataset. Ozone, salicylic acid, and ethylene response expression data were collected via Genevestigator (experiments 13, 113, and 17, respectively). Pathogen response microarray data sources are described in Table 3.1. Differential regulation is shown on a scale from 8-fold down regulated (1/8, green) to 8-fold up regulated (8, red).

tissue/developmental and stress response expression data, and existing literature, putative functions were assigned to the expression clusters.

β-1,3-glucanases in cell wall morphogenesis and cell division

The largest expression cluster (group M) contains 13 genes, and displays a somewhat non-specific expression pattern with abundant expression in a variety of tissues/organs including the flower, seeds, shoot apex and in some cases the root and hypocotyl (Figure 3.3A). The minimum expression correlation for a group M member with the centroid of the cluster is 0.62, despite the large group size and widespread expression throughout the plant. In general, group M genes display only a minimal response to most stresses and hormones (Appendix 2) and a somewhat negative response to biotic stress with the exception of the tumorigenic stressor *Agrobacterium tumefaciens* (Genevestigator experiment #8, data not shown), making a PR role improbable. Given the expression in multiple tissues and particularly the high relative expression in the shoot-apex, involvement of group M genes in a constitutive biological process such as cell division/cell-wall remodeling is likely. The deposition and removal of callose during cell plate formation in active meristems has been noted in previous studies (Fulcher et al., 1976; Longly and Waterkeyn, 1977; Samuels et al., 1995) and suggests a role for β-1,3-Gs in cytokinesis. β-1,3-Gs have also been implicated in the maintenance of callose-free plasmodesmata during cell differentiation and signaling (Rinne et al., 2001). In addition, recent computational and proteomic studies have revealed a large number of *Arabidopsis* β-1,3-Gs to be GPI-anchored, a feature that is strongly associated with cell wall remodeling (Borner et al., 2002; Borner et al., 2003; Elortza et al., 2003). As determined using the ATTED database, group M genes were found to be significantly co-expressed with a number of cell division and cell wall remodeling related genes including At3g02210 (COBRA cell expansion protein), At1g02730 (cellulose synthase family protein), At5g18580 (TONNEAU 2, a cell morphology control protein), At4g34160 (cyclin delta 3) and number of cell-division control and cell-division cycle family proteins (Appendix 3). *Arabidopsis* COBRA genes also encode GPI-anchored proteins involved in cell wall expansion and control of cell wall morphogenesis (Roudier et al., 2002). In the previous study by Borner et al. (2002), 9 of 13 of the genes in group M were predicted as GPI-anchored proteins, four of which were confirmed in a subsequent study (Borner et al., 2003). As predicted by the GPI-SOM algorithm, all GPI-attachment sites are located in the far C-terminal region following the hydrolytic domain and CBM43 domain when

present (Appendix 1). One interesting question regarding group M genes is why so many appear to have very similar developmental expression profiles. One possibility is that these genes have undergone expression divergence in categories that were not included in the microarray dataset. If expression patterns appear highly similar on a tissue/organ macro level, it is still possible that they are dissimilar on a smaller spatial or temporal scale, which may indeed be the case for genes involved in cell wall remodeling.

Flower-specific β-1,3-Gs and possible roles in microsporogenesis and pollen tube growth

Expression groups H and K include β-1,3-G genes highly specific to flower/reproductive organs (Figure 3.3A). While the roles of β-1,3-Gs in plant flowering are not entirely clear (Delp and Palva, 1999), several specific functions have been identified. In pollen development, before microspores can be released into the anther locule for pollen maturation, a thick callose wall surrounding the tetrad must be degraded. In *Arabidopsis*, this function is associated with two characterized "anther specific" *Arabidopsis* β-1,3-G genes (At4g14080 and At3g23770) (Hird et al., 1993), both found in expression group K along with three other somewhat weakly expressed genes. A second function of β-1,3-Gs in flower development involves dissolution of callose in the stylar matrix during pollen tube growth, a function proposed both for glucanases expressed in the style (Delp and Palva, 1999) as well as glucanases found in pollen grain itself (Huecas et al., 2001; Takeda et al., 2004). This possible function has been described for a previously identified *Arabidopsis* gene At5g20330 (BG4) found in group H, which was shown by previous RNA blots and *in situ* hybridization results to be expressed in the style and septum of the ovary (Delp and Palva, 1999). In the same study, the tandemly linked gene At5g20340 (BG5), also found in group H, was not preferentially expressed in the style or ovary, but displayed a flower-specific expression pattern (Delp and Palva, 1999). The expression data is consistent with both of these previous results (Figure 3.3A). Other genes highly expressed in the pollen and/or stamen are found in cluster H and may play similar roles in pollen tube growth. However, it must be noted that several of these genes are very weakly expressed and thus the small peak in pollen-specific expression observed for these genes is less convincing. A putative PR β-1,3-G, At3g57270 (BG1) (Dong et al., 1991), falls into this category. These genes may appear to be weakly expressed due to the lack of an appropriate experiment in the microarray dataset capable of inducing mRNA expression above background levels.

The proposed functions of expression groups H and K in microsporogenesis and pollen development are further substantiated by results obtained from the ATTED database (Appendix 3). At3g23770 is co-expressed with MS2 (r = 0.978), a male sterility protein expressed in the tapetum during the release of tetrad microspores, thus further supporting its role microsporogenesis (Aarts et al., 1997). Furthermore, two neighboring group H, pollen/stamen-specific genes (At5g20390 and At5g64790) expressed in the final stages of *Arabidopsis* development (Figure 3.3) are co-expressed with a tyrosine phosphatase known as AtPTEN1. This gene is also expressed exclusively in pollen grains during the late stage of pollen development and is an essential gene for pollen maturation (Gupta et al., 2002). Lastly, the expression of BG4 is significantly correlated with that of the INO (inner no outer) gene, which encodes a transcription factor essential in the determination of abaxial-adaxial patterning in ovule development (Villanueva et al., 1999). Future studies of these genes would be valuable for further understanding the role of β-1,3-Gs in microsporogenesis and pollen development.

Despite the flower-specific expression of the above-mentioned genes, several genes tend to show some expression in particular stages of seed development as well (e.g., At5g64790, At5g20390). These genes, along with some other genes including At2g39640, At5g18220 and At3g57240, may reflect the established role of β-1,3-Gs in seed development and after-ripening (Leubner-Metzger and Meins, 2000; Leubner-Metzger, 2003; Leubner-Metzger, 2005).

Root and leaf β-1,3-Gs

In addition to the pollen development and cell division related expression classes that account for the majority of β-1,3-Gs in *Arabidopsis*, expression clustering also revealed a class of β-1,3-Gs highly specific to the root (group D) and leaves (groups A-C). As illustrated in Appendix 2, genes found in these groups also displayed the most significant responses to fungal pathogens and therefore these groups are most likely to contain possible PR genes. Groups B and C contain two known PR-glucanases from *Arabidopsis*, At3g57260 and At3g57240, respectively, which are both highly expressed in leafy tissues. Genes in groups A-D were also generally more stress-responsive than other clusters (Appendix 2). This likely reflects common pathways involved in regulating stress-responses and responses to pathogen attack.

Identification of PR β-1,3-Gs using biotic-stress expression data

In order to more quantitatively determine which β-1,3-Gs are likely to be PR proteins, β-1,3-G genes exhibiting a greater than three-fold increase in expression in response to any of the five fungal phytopathogens for which data was available were identified and are shown in Table 3.1. Three genes had a greater than three-fold change in expression in more than one pathogen category, and two genes were responsive to one pathogen only. The genes identified using this procedure (marked by X's in Figures 3.3 and 3.4) include two known PR-glucanases (At3g57260, BG2, PR-2; At3g57240, BG3, PR-3) as well as three more potential PR genes (two root-specific and one pollen/stamen-specific). At4g16260 (a root-specific β-1,3-G) was highly expressed following treatment with ethylene (Figure 3.3B), displayed a significant expression response to the largest number of pathogens (4) and was most highly induced by pathogen stress of all the PR genes (22- and 53- fold induction by *A. brassicicola* and *P. infestans*, respectively). A previous study also reported this gene as being significantly induced by the bacterium *P. syringae* (Mahalingam et al., 2003). This gene represents an excellent target for future studies of *Arabidopsis* PR-proteins. Salicylic acid and ethylene, two major hormonal regulators of PR-responses (Leubner-Metzger and Meins, 1999), are associated with a greater than 3-fold induction in one (At3g57240) and two cases (At3g57240, At4g16260), respectively (Figure 3.3B). Treatment with ozone, another well-known elicitor of plant PR-proteins and defense responses (Langebartels et al., 2000), resulted in greater than 3-fold induction for 4 of the 5 genes. Ozone is known to induce mRNA levels of a PR β-1,3-glucanase in tobacco as well (Ernst et al., 1992).

Results obtained from the ATTED database provided additional support for the possible role of these genes in pathogen defense (Appendix 3). The known PR-glucanase (BG2) is significantly co-expressed with At2g14610 (pathogenesis-related protein 1, PR-1) (Buell and Somerville, 1997). Furthermore, At4g36010 (a PR-thaumatin protein) and At3g12500 (basic endochitinase, ATHCHIB, PR-3) are correlated with two previously uncharacterized, potential PR-glucanases, At3g04010 and At4g16260 respectively. Co-expression of β-1,3-Gs with other PR proteins has been shown to provide a synergistic antifungal effect *in vivo* (Jach et al., 1995).

Table 3.1: **β-1,3-G genes with greater than a 3-fold change in expression level following fungal pathogen treatment.**

AGI ID	*A. brassicicola*	*B. cinerea*	*E. cichoracearum*	*E. orontii*	*P. infestans*
At4g16260	22.48	10.15	2.38	4.35	53.44
At3g57260[a]	0.77	3.47	8.02	3.50	1.55
At3g57240[a]	0.26	0.89	6.32	3.16	1.14
At3g04010	2.06	10.52	2.05	1.08	2.59
At3g55430	1.30	4.81	0.62	0.91	1.56

Note.—Values indicate change in expression relative to controls. The corresponding experiments used by Genevestigator are *A. brassicicola* (experiment 161); *B. cinerea* (experiment 147); *E. cichoracearum* (experiments 85); *E. orontii* (experiment 146); and *P. infestans* (experiment 108).

[a] Previously characterized PR genes (Dong et al., 1991).

Reconstruction of ancestral expression states reveals functional divergence

In comparing expression data and sequence-based phylogeny, several related questions arise: do closely related β-1,3-G genes exhibit closely related expression profiles, or is evolution of sequence and expression uncoupled? If sequence and expression divergence are interrelated, which functional/expression classes are ancestral and which are derived? In order to investigate these questions, the functional classes obtained from expression clustering and analysis can be combined with the phylogenetic tree, and ancestral reconstruction of expression states can be used to build a model of functional divergence.

Using the program Mesquite (Maddison and Maddison, 2005), the expression clusters were mapped onto the phylogenetic tree as single states and the ancestral expression states were reconstructed by parsimony (Figure 3.4). The colors of the boxes at the terminal branches represent the expression/function classes of the corresponding genes. Genes with absent profiles were given null states. A change in color between two interconnected branches signifies a putative expression divergence event. As seen in Figure 3.4, a considerable number of neighboring genes, likely having arisen from relatively recent duplication events, share expression states (e.g., At3g23770 and At4g14080). Indeed, according to Blanc and Wolfe's (2004) study, these two genes are listed as recent duplicates due to a polyploid (whole genome duplication) event in *Arabidopsis*.

The parsimony reconstruction performed on the entire tree indicates the most likely hy-

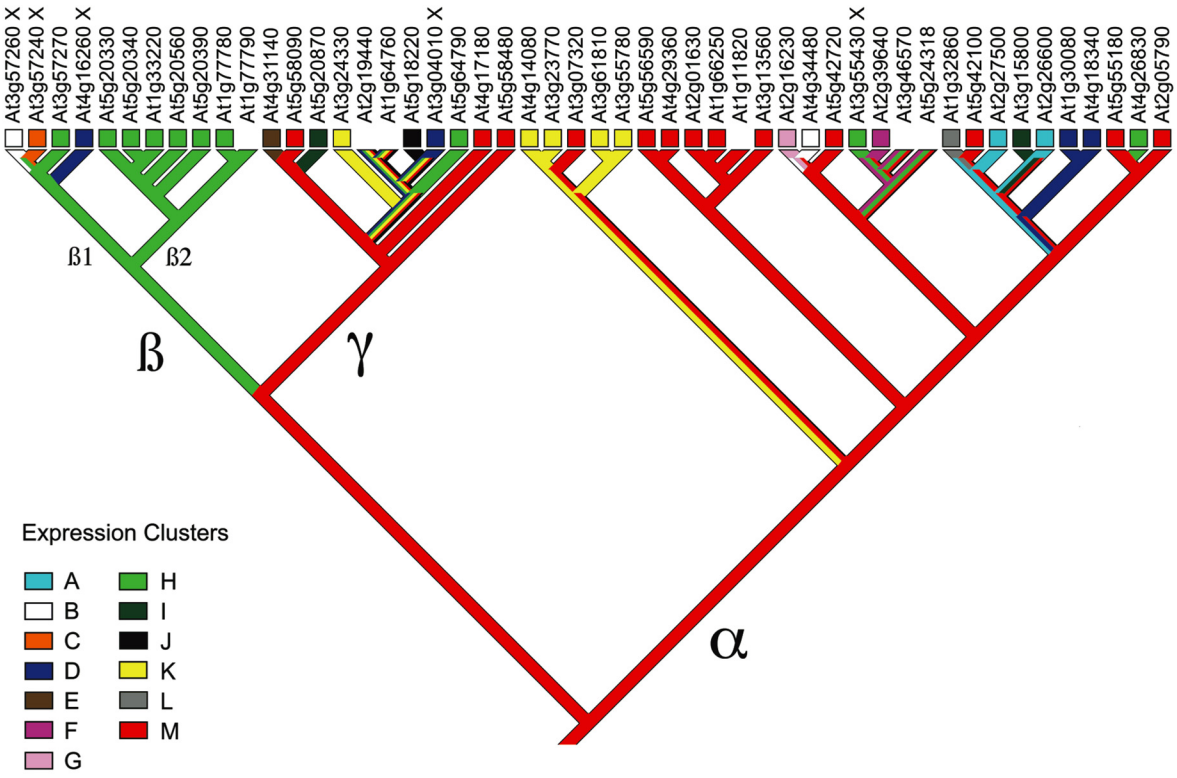


Figure 3.4: **Phylogenetic reconstruction of ancestral expression states using parsimony.** Colored boxes are shown at the terminal branches of genes included in the expression clustering. Genes with absent expression profiles (no box) are associated with null states in the reconstruction. The colors corresponding to each expression group are shown in the legend, and multi-colored branches are associated with ambiguous (multiple possible) states. PR-glucanases identified through fungal stress response expression analysis are marked with an X.

pothetical expression states for ancestral genes according to a maximum parsimony model. The root ancestral expression state, shared by the root of clades α, γ as well as the midpoint root of the entire tree, is the expression state M (Figure 3.4). Thus, according to parsimony reconstruction, the ancestral β-1,3-G gene most likely had a group M-like expression state and thus a cell division/cell wall remodeling like function, which accounts for the wide distribution of group M genes throughout the tree. A cell division/cell wall remodeling like function for the ancestral β-1,3-G genes also makes sense from an evolutionary perspective since cell division is more primitive than the other β-1,3-G functions and is consistent with

the functions of other related proteins and protein families. For example, the endo-β,1,3-G (eng1p) from yeast functions in degradation of the primary septum during cell separation (Martin-Cuadrado et al., 2003). Other plant glycosyl hydrolases such as xyloglucan endotransglycosylases (Campbell and Braam, 1999) and β-1,4-glucanases (Nicol and Hofte, 1998; Zuo et al., 2000) also have well documented roles in cell wall remodeling related functions such as cell wall biogenesis, expansion, and loosening.

According to the model, after several ancient rounds of duplication of a cell division related β-1,3-G gene, the gene family diverged in expression and thus function producing β-1,3-Gs involved in pollen development, pathogen resistance and other processes. A particularly interesting result is that, while some expression states are highly clade-specific, situations in which the same or similar expression profiles emerge independently at different locations in the tree are also quite common. At5g20390 and At5g64790, for example, are neighbors in the clustering tree as they have highly similar developmental expression profiles (r = 0.89) but are found in distant phylogenetic clades, β and γ, respectively. In cases such as these, the genes may have independently acquired or lost similar tissue/developmental cis-regulatory elements. Whatever the mechanism may be, it appears to be a common phenomenon in the evolution of the β-1,3-G gene family.

A major question remaining is, where do the PR β-1,3-Gs lie in the phylogenetic tree? As indicated by X's in Figure 3.4, two of the pathogen-induced genes are found within clades α and γ, while the major cluster of PR genes is located in clade β1, which includes the previously characterized PR genes (BG1, BG2, and BG3) as well as a suspected PR gene, At416260. The lack of the C-terminal domain and hydrophobic tail sequence in clade β indicates that it likely arose from an ancestral β-1,3-G gene which had its C-terminal region deleted. Removal of the C-terminal region including the GPI-anchoring site would have effectively directed the ancestral protein extracellularly. This represents a pivotal evolutionary event in which an ancestral β-1,3-G involved in host cell wall morphogenesis was suddenly free to hydrolyze β-1,3-glucan in the cell walls of potential microbial pathogens.

Since the expression state of group M appears to be tightly developmentally controlled and is not responsive to fungal stress, divergence from the ancestral expression state appears to be an additional requirement for evolution of PR-function. All five putative PR genes have diverged away from the ancestral state M, and are associated with four different non-ancestral expression states. This also reflects a considerable amount of expression divergence between PR genes themselves. Given the range of expression states

associated with PR genes, it is not surprising that PR β-1,3-Gs are also highly variable in their responses to hormones and chemicals (Appendix 2), and different species of fungal pathogens (Table 3.1). Possessing diversity with respect to these characteristics is a widely documented feature of PR β-1,3-Gs and is important for broadening the regulatory response to pathogen attack (Memelink et al., 1990; Ward et al., 1991; Leubner-Metzger, 2003). Thus, for a number of reasons, expression divergence appears to be a major factor in the maintenance of duplicated PR β-1,3-G genes in evolution.

Two other genes that display PR-like expression responses are found in clades α and γ, but unlike the PR genes in clade β, both contain a C-terminal domain. Furthermore, At3g04010 is a predicted GPI-anchored protein (Borner et al., 2002; Borner et al., 2003) and thus may be cell wall associated, which raises into question its ability to act as a PR protein. Whether these genes encode actual PR-proteins or simply display PR-like expression responses to biotic stress remains to be determined. Nevertheless, the multiple independent origins of PR-like expression patterns at different locations in the phylogenetic tree suggests that acquisition of PR-responsive regulatory elements is not an uncommon occurrence in the evolution of the *Arabidopsis* β-1,3-G gene family. This, paired with the considerable level of expression divergence among PR genes and the loss of the C-terminal region containing the hydrophobic C-terminus and GPI-anchoring site, provides a mechanism for evolution of β-1,3-Gs involved in pathogen-defense.

Statistical evaluation of ancestral state reconstruction

In order to statistically evaluate whether the pattern of expression divergence produced by clustering and ancestral state reconstruction is non-random, a permutation test was performed in which the expression states were randomly reshuffled and parsimony reconstruction was performed 10,000 times. The number of steps required in the random distribution ranged from 22 to 31, and the observed parsimony score of 22 steps is indicative of a non-random distribution with a permutation p-value of 0.0002. This supports a model of β-1,3-G evolution in which divergence of developmental expression patterns is partially coupled with sequence divergence. Only marginally significant or non-significant results were obtained using stress-, hormone-, or pathogen-response data, with p-values of 0.0429, 0.3165, and 1.0000, respectively (Figure 3.5). One possible reason for this is that β-1,3-G genes may have diverged faster in terms of stress, hormone, and pathogen response than

developmental response, and that the rate of expression divergence in these categories outpaced the rate of sequence divergence to a degree where the phylogenetic signal was lost. Alternately, it is possible that the non-developmental datasets were more variable due to the large number of microarray data sources and types of stress experiments, or there was an absence of phylogenetically relevant characteristics in the microarray data. In the case of the pathogen-response data, a p-value of 1.0000 (two-tailed p-value < 0.0001) is of interest as it suggests that similar pathogen-response expression profiles arise independently significantly more often than can be expected by chance. This may be a consequence of a positive selection pressure for increased pathogen response and is an interesting area for future analysis.

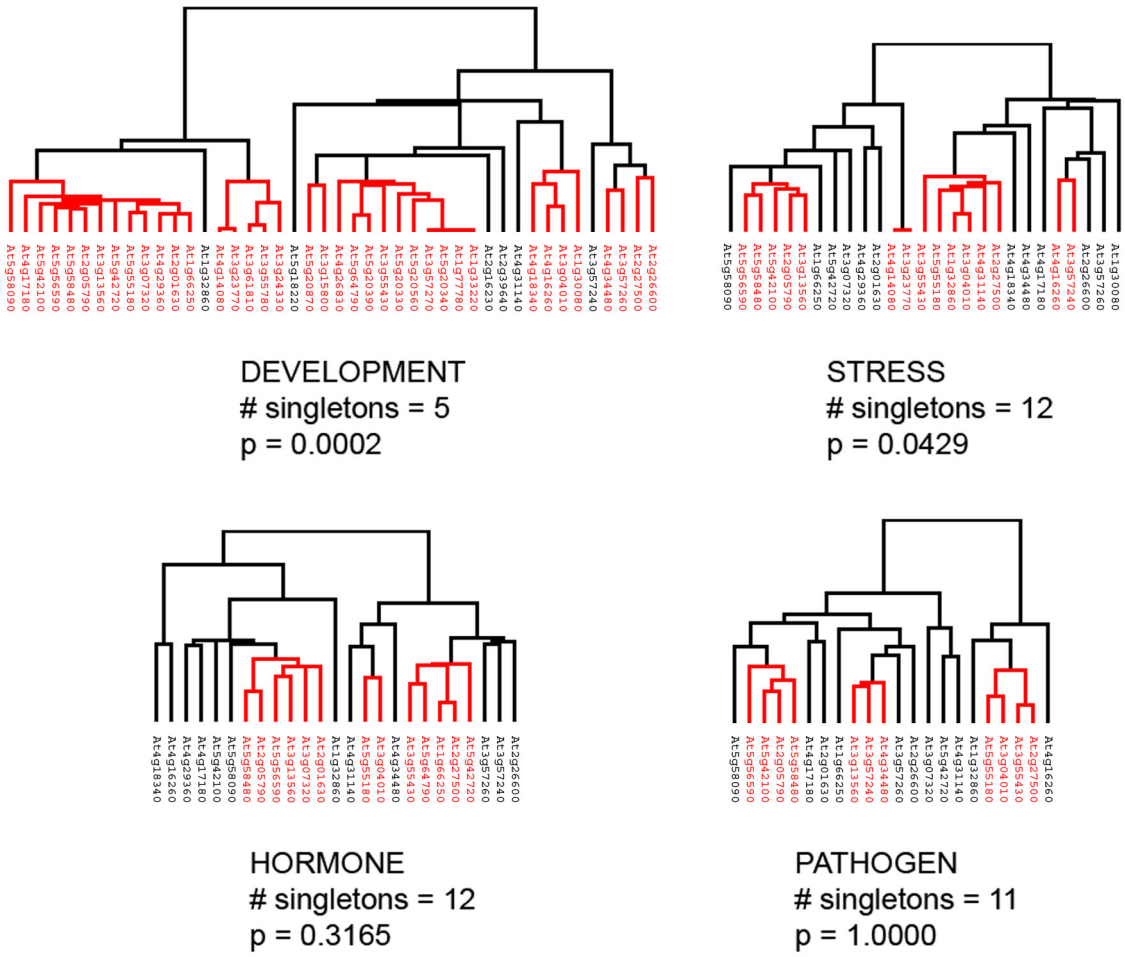


Figure 3.5: **Clusters obtained from hierarchical clustering of four types of gene expression data using a clustering threshold of 0.5 (Pearson correlation).** Permutation test statistics resulting from mapping and parsimony reconstruction of each set of clusters is also indicated.

3.4 Conclusion

The expression classes resulting from expression clustering and analysis are consistent with existing data on individual *Arabidopsis* genes, successfully reflect the known functions of β-1,3-Gs, and provide a means to putatively categorize uncharacterized β-1,3-Gs. Clustering of tissue/developmental expression data revealed β-1,3-Gs associated with cell wall remodeling as well as pollen development, while analysis of fungal stress expression data facilitated identification of known and several putative PR-glucanases. The ancestral function assigned through ancestral state reconstruction was cell division/cell wall remodeling and is consistent with the known functions of β-1,3-Gs in more primitive organisms as well as related gene families. Other interesting results include the independent origin of highly similar developmental expression profiles in more distantly related sequences and the variable degrees of expression divergence in different clades. The considerable level of expression divergence observed in the *Arabidopsis* β-1,3-G gene family provides a basis for assessing the evolution of β-1,3-G function. Furthermore, the acquisition of PR-like expression responses following divergence from the ancestral expression state combined with the loss of the C-terminal region and GPI-anchoring site (as seen in clade β) facilitates fungal stress-induced extracellular secretion, and thus evolution of β-1,3-Gs involved plant pathogen defense.

Ancestral state reconstruction paired with microarray data clustering and analysis provides a powerful method for investigating evolution of function in large gene families. It may be particularly useful in assessing functional diversity in large gene families where molecular function stays relatively conserved, but where biological function diversifies through *regulatory divergence*.

Conserved sequence motifs specific to a particular subclade (tree determinant residues) often define specific functions of that subclade. This is also likely true of functional gene expression traits, which can be identified in a similar way — by finding shifts in expression patterns occurring at different nodes in the phylogenetic tree, and attempting to understand these shifts in terms of biological function. By assessing ancestral shifts in gene expression, it is then possible to better understand the context in which new functions develop. For instance, loss of a developmental gene expression pattern, acquisition of stress and pathogen-response elements and extracellular secretion afforded by the loss of the GPI-anchor, provides an ideal regulatory pattern for the development of a subclade

of pathogen-defense genes. While dividing gene expression patterns into a set of discrete profiles is an imperfect and parameter-dependent approach, it is simple and easily testable by a parsimony model with few assumptions (equal state changes). Possible enhancements of the approach include: separately assessing individual expression clusters, experiments, and nodes for evidence of *functional* expression shifts; modeling of gene expression data in a more continuous fashion (reducing reliance on parameters and oversimplification resulting from clustering), and correlating expression shifts with changes in cis-regulatory sequence motifs.

Chapter 4

Structural repeats

In the previous two chapters, changes to the protein surface or its expression at the gene regulatory level, were investigated as determinants of function and functional change. These changes are masked when examining sequence similarity alone as an indicator of functional similarity.

In this chapter, a third feature of proteins (*internal repeats*) is investigated an element of protein evolution and function. Using the β-trefoil fold as an example, it is demonstrated that different members of the same protein fold or even family can possess general similarity while having been formed from separate repeat elements. A large-scale computational screen is performed to find cases in which distinct precursor repeats have independently generated β-trefoil structures through repetition. These cases provide insights into the evolutionary mechanisms by which protein structures themselves are assembled, and how functions can be amplified through repetition of supersecondary structural elements.

The results presented in section 4.1 demonstrate that sequence- or structure-based classification of β-trefoils into a single family in which each member is presumed to have arisen by duplication and divergence from a common ancestor, is an oversimplification and does not account for evolutionary changes that take place at the level of internal repeats. One case whose function is incorrectly annotated by existing methods (the PPOD family) is then studied in section 4.2. Through analysis of its internal repeats, conservation patterns, surface motifs and evolutionary history, its function is correctly inferred.

4.1 Repeat-mediated evolution of β-trefoils

4.1.1 Introduction

The default explanation for the existence of significant sequence and structural similarity between two proteins is that they have arisen by duplication and divergence from a common ancestor. This assumption is the basis of many approaches to homology-based functional inference. As an example from the previous section, *Arabidopsis* β-1,3-glucanases have all evolved from the same ancestral protein/fold (a TIM-barrel enzyme) and diverged in function through expression changes and other aspects. In this case, the 3D structure/fold is largely maintained due to structural and functional constraints.

Is it possible that different proteins with identical structures and even in some cases, significant sequence similarity, may not be related by *domain duplication*? If so, this would provide another limitation of approaches that evaluate overall similarity as an indicator of function and evolutionary relationship. To gain insights into this question, this work examines a highly common and symmetric superfold (the β-trefoil) (Orengo et al., 1994) composed of smaller repeated structural elements which may represent more basic units of protein evolution and function.

Despite considerable sequence diversity within the β-trefoil superfamily, all β-trefoil proteins share a common 3D structure, a threefold symmetrical capped β-barrel composed of three β-β-β-loop-β repeats (Murzin et al., 1992) (Figure 4.1). In addition to the well-known binding domains of the ricin type family of toxins (Hazes, 1996), and cytokine β-trefoils [IL1 (Priestle et al., 1989), FGFs (Ornitz and Itoh., 2001)], the β-trefoil superfamily also includes MIR domains (Ponting, 2000), agglutinins (Transue et al., 1997), STI-like domains [Kunitz inhibitors (Antuch et al., 1994) and *C. botulinum* neurotoxin C-terminal domains (Lacy et al., 1998)], actin-crosslinking proteins [fascins (Kureishy et al., 2002), hisactophilins (Habazettl et al., 1992)], LAG-1 DNA-binding domains (Kovall and Hendrickson, 2004), and AbfB domains (Miyanaga et al., 2004). Between families, there can be considerable sequence divergence; sequence identity can be below 5% in some cases, which falls in the range of the "midnight" zone of homology, where it becomes difficult to delineate evolutionary relationships between sequences (Rost, 1999).

Previous research on the evolutionary relationships within the β-trefoil superfamily revealed significant similarities between structurally aligned positions of highly divergent β-

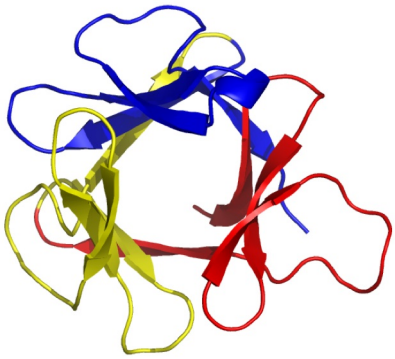


Figure 4.1: **A β-trefoil structure (hisactophilin, PDB ID 1HCD) divided into its three internal repeats.**

trefoils (Ponting and Russell, 2000). These detected relationships have led to the hypothesis that the β-trefoil superfamily arose from a common ancestor (Ponting and Russell, 2000).

The repetitive nature of the structure itself (Figure 4.1) suggests a mechanism by which the ancestral β-trefoil may have originated; that is, by triplication of a more primitive four-stranded element—a single repeat of approximately 40 residues (Rutenber et al., 1987; Ponting and Russell, 2000). Fingerprints of an ancestral triplication can be seen in the repeated sequence patterns of existing β-trefoil proteins. For example, a threefold repeated QXW galactose-binding motif is found in ricin type lectin family, which forms the binding domain of ricin type toxins (Hazes, 1996). In addition, MIR motifs are defined as a 40-residue segment, three of which form an intact β-trefoil domain (Ponting, 2000). In other β-trefoil families such as the eukaryotic cytokines (IL1s and FGFs), putative ancestral repeats have diverged considerably in sequence, beyond the point at which they can be recognized at the sequence level.

Increasingly sensitive computational approaches have been developed for detecting internal repeats in proteins based on multiple alignments, which can recognize ancestral repeats even in ancient protein families (Heger and Holm, 2000; Söding et al., 2006). But beyond detection of internal repeats, few studies have performed comparative analyses of repeats to study their ongoing role in the evolution of a protein fold/family. A recent study of the β-propeller fold identified β-propellers apparently assembled by "amplification" of individual blades (repeats) (Chaudhuri et al., 2008). This suggests that fold evolution from repeats, largely assumed to be an extremely ancient phenomenon, may be a recurring

evolutionary process. Do common globular folds such as β-trefoils also continually evolve from repeats and, if so, can these events be reconstructed experimentally?

This work explores the current database of β-trefoils in an attempt to characterize the extent that different β-trefoil domains have been constructed from separate *precursor repeats*. It was hypothesized that multiple, independent, origins of β-trefoil families/subfamilies via tandem repetition of unique $\beta\beta\beta$-loop-β supersecondary structural elements may underlie the extreme variation in observed sequence symmetry among different β-trefoils. Identified cases such as these would demonstrate that extant proteins with the β-trefoil fold do not necessarily arise by domain duplication from a common ancestor, but can form and reform from more basic units (structural repeats) of structure and evolution.

4.1.2 Methods

The basis of the computational approach used here is the identification of β-trefoil subfamilies containing greater *internal* similarity between repeats than *external* similarity to other subfamilies. Such cases would be analogous to genomic DNA repeats that undergo a process of concerted evolution[1]. In summary, repeats can be compared by:

1. Parsing β-trefoil structures and sequences into structural repeat units.
2. Measuring the sequence and structural similarity between repeats.
3. Clustering or comparing repeats based on these similarities.

Sequence dataset construction and analysis

All annotated β-trefoil domain sequences were retrieved from the NCBI using the Conserved Domain Database (CDD). All families annotated as β-trefoils by SCOP and PFAM with available structures in the PDB were included. See Table 4.1 for statistics on construction of the dataset. A script was written using the Perl programming language to parse these sequences and extract their β-trefoil regions according to the CDD information

[1]Concerted evolution is the tendency of duplicated genetic elements (e.g., genes, DNA repeats) to evolve in concert with each other. Concerted evolution of repetitive DNA sequences has been widely studied. In this case, repeats appear more similar within than between genomes/species (Elder and Turner, 1995).

included in the NCBI's GenPept file. All β-trefoil domains in each protein chain were extracted, which resulted in an initial dataset of 5,287 domain sequences.

Table 4.1: **Dataset construction and calculated sequence symmetries.**

Family	#Domains	# Domains after filtering[a]	Average sequence symmetry[b]	Representative structure used for alignment
AbfB	24	15	17.7	1WD3
Agglutinin	14	5	9.0	1JLX
CD Toxin	69	28	9.2	1SR4
Fascin	413	129	13.0	1DFC
FGF	775	140	10.3	1NUN
IL1	362	86	8.2	1MD6
STI / Kunitz	452	89	7.7	1WBA
LAG1	31	18	7.0	1TTU
MIR	1267	65	10.3	1T9F
Ricin	1604	518	14.3	1QXM
Toxin_R_Bind_C	89	15	7.3	3BTA

[a]Domain sequences were filtered by removing redundancy (see Methods).

[b]Sequence symmetry was calculated as the average pairwise percentage identity between the three repeats.

To remove redundancy, all domain sequences were grouped into clusters of highly similar sequences using the BLASTCLUST algorithm from the BLAST package with default parameters. The longest sequence from each cluster was selected as a representative and the remaining sequences were removed from the dataset.

β-trefoil sequences were then parsed into their individual repeats by aligning all sequences to their corresponding β-trefoil family HMM using the program HMMalign (hmmer.janelia.org), and dividing the sequences into three parts according to the repeat pattern evident within a representative structure (described in following section). A repeat prediction program (e.g., RADAR) was not used to select repeats because this could bias the detected symmetries to appear higher within sequences than between sequences. The representative structures used in repeat parsing were: 1WD3, 1QXM, 1DFC, 1NUN, 1T9F, 1SR4, 1MD6, 1JLX, 1TTU, 3BTA, and 1WBA (Table 4.1). Sequences that were truncated and/or contained insufficient data were excluded by only including sequences containing three repeats with length > 20 residues. The sequences from the structural dataset (described in next section), were also added, resulting in a final dataset consisting of 3,501

repeats from 1,167 β-trefoil domains.

Repeats were then clustered in two ways:

1. **Identification of connected components of a similarity graph**. An all-by-all BLASTp search was performed, and any two repeats with $E < 0.0001$ were connected. A Perl script was written to translate these results into a network that could be visualized with the program, Cytoscape (http://www.cytoscape.org). This approach will produce clusters analogous to those produced from a single-linkage clustering algorithm (Bejerano et al., 2004).

2. **Neighbor-joining**. A hierarchical clustering tree was constructed by the neighbor-joining algorithm (Saitou and Nei, 1987) using distances obtained from all-by-all global alignments. Due to the difficulty in statistically distinguishing higher-order clustering patterns from chance, this method was used primarily for data exploration.

Structural dataset and repeat comparison

The following procedures were conducted in collaboration with Aron Broom.

The structural dataset was built by retrieving all known β-trefoil structures from the Protein Data Bank (PDB), and dividing each β-trefoil into three separate β-trefoil repeat elements. The following PDB files were included in the dataset: 1ABR (2 domains), 1AVA, 1AVW, 1BAR, 1BFG, 1DFC (4 domains), 1DLL, 1DQG, 1EYL, 1FMM, 1GGP (2 domains), 1HCD, 1HWM (2 domains), 1IHK, 1IJT, 1ILR, 1J0S, 1JLX (2 domains), 1KNM, 1M2T (2 domains), 1MD6, 1N4K, 1NUN, 1Q1U, 1QQK, 1QXM (2 domains), 1R8N, 1RG8, 1S0E, 1SR4 (2 domains), 1T9F, 1TTU, 1UPS, 1V6W, 1VCL (2 domains), 1WBA, 1WD3, 1XHB, 1YBI (2 domains), 2AAI (2 domains), 2F2F (2 domains), 2I1B, 2IHO, 3BTA, and 8I1B.

Trefoil elements were defined as four consecutive β strands, where strands one and four belong to the β-trefoil barrel and strands two and three form part of the hairpin-triplet (Murzin et al., 1992). Trefoil elements were separated from each other at the turn between strand four of the element being considered, and strand one of subsequent element. An all-by-all structural similarity comparison was performed using the McLachlan algorithm (McLachlan, 1982) as implemented in ProFit (http://www.bioinf.org.uk/software/profit/)

and a pairwise similarity matrix was generated using the Q-score as the similarity metric (see equation 4.1 below).

Equation 4.1:

$$Q = \frac{N_{align}^2}{[(1+(RMSD/R_0)^2) \times N_1 N_2]}$$

In the above equation, N_1 and N_2 represent the number of residues in both structures being compared, $RMSD$ is the root mean squared deviation following structural alignment, N_{align} is the number of aligned residues, and R_0 is the distance cutoff (Å) for two atoms to be considered aligned. An R_0 value of 3.5Å was used.

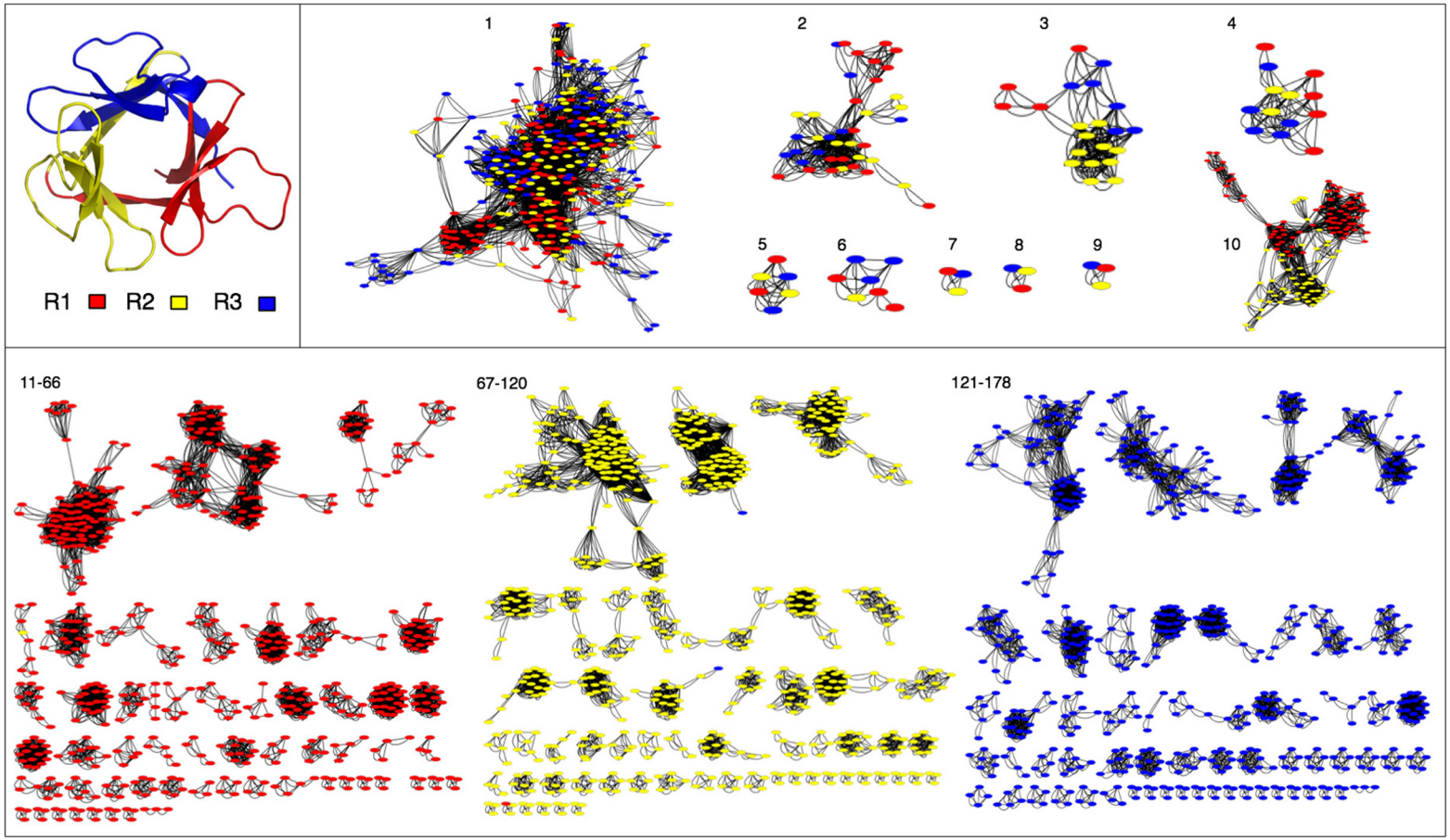
4.1.3 Results

Recurring repeat-mediated assembly of β-trefoils revealed by clustering

Given the size and diversity of the sequence repeat dataset, aligning all repeats and constructing a tree was not feasible. Thus, pairwise clustering methods were used to compare similarities between repeats. To obtain a global view of relationships between internal repeats, all-by-all sequence similarities were computed using BLAST, and any two repeats for which an evolutionary relationship could be detected (E < 1e-04) were connected. The resulting clusters can be visualized as a network as shown in Figure 4.2. The majority of the clusters (#11 to #178) contain only repeats of the same order within their parent domain (i.e., R1 from a particular domain grouping together with R1 repeats from other domains). This is expected and consistent with evolutionary relationships that have resulted from whole domain duplication, as this will preserve repeat order.

An unexpected number of clusters were observed that rather consist of all three constituent repeats (R1, R2, R3) [see clusters (#1 to #9) in Figure 4.2]. Furthermore, each of these clusters contains all three repeats from one or more complete β-trefoil domains (see Table 4.2 and Figure 4.3 for cluster composition and representative alignments, respectively). For the sequences within these nine clusters, there is significant sequence similarity between the three repeats within one or more β-trefoil domains, but considerably less or no detectable similarity to repeats from other clusters. These clusters therefore represent nine separate instances of evolution generating new β-trefoils by triplication of different precursor repeats.

Figure 4.2 *(following page)*: **Sequence-based clustering of internal repeats from the β-trefoil superfamily.** An example β-trefoil structure (PDB ID 1HCD) divided into its three constituent structural repeats (R1, R2, R3) is shown in the inset. All domains in the dataset were divided into these three internal repeats (see Methods), and any two repeats were connected if significant sequence similarity could be detected (BLAST E-value < 1e-04). All repeat clusters of size ≥3 were then visualized in a network form using Cytoscape. Ten clusters of interest are numbered, nine of which contain the first, second, and third repeat from one or more β-trefoil domains. Clusters 1-9 indicate separate evolutionary origins of a β-trefoil domain via repetition from a unique precursor element. Cluster 10, which includes a mixture of R1 and R2 but not R3, is evidence of an internal duplication. Clusters 11-66, 67-120, and 121-178 consist almost entirely of groupings of R1, R2 and R3 respectively.


R1
R2
R3
1
2
3
4
5
6
7
8
9
10
11-66
67-120
121-178

Clusters 1-9 range in size from three repeats, as in clusters #7 to #9 (each of which contain R1, R2 and R3 from an individual protein), to 376 repeats in cluster #1 (Figure 4.2). All repeats within these clusters are from parent domains of the same family, which are ricin-like (7 clusters), fascin-like (1 cluster), or of the α-L-arabinofuranosidase B (AbfB) family (1 cluster). The sequences associated with these clusters are of prokaryotic and eukaryotic origin, and are from families with known roles in carbohydrate-binding.

Table 4.2: **Cluster composition and representative sequences.**

Cluster	# Repeats in cluster	# (domains, repeats) where R1, R2, R3 are present	Family	Representative Sequence
1	376	64, 192	Ricin-like	gi|55229846
2	40	10, 30	Fascin-like	gi|86165939
3	21	4, 12	AbfB	gi|125714851
4	13	2, 6	Ricin-like	gi|111147635
5	6	2, 6	Ricin-like	gi|154159760
6	7	1, 3	Ricin-like	gi|55670423
7	3	1, 3	Ricin-like	gi|149242174
8	3	1, 3	Ricin-like	gi|119455865
9	3	1, 3	Ricin-like	gi|29611230

Sequence symmetry distributions

The observed distributions of sequence symmetries within families (i.e., average identity between internal repeats) are also consistent with the clustering results. That is, the ricin, fascin, and AbfB families, which according to the results have undergone recent fold-generating repeat events, also have the most highly skewed symmetry distributions compared to other families (Figure 4.4). Symmetries range from under 4% average identity between repeats to extremely high levels (66%) in β-trefoil domains that have arisen from recent repeat events. The abundance of apparent outliers with unusually high symmetries in the ricin and fascin families suggests that these recent events have occurred (or recurred) in certain subfamilies, while the repeats in other subfamilies have continued to diverge over time.

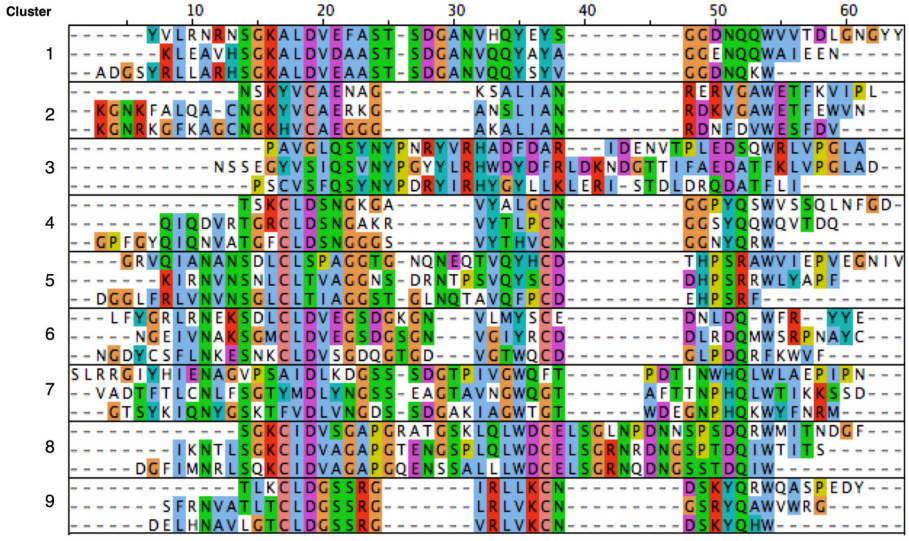


Figure 4.3: **Internal repeat alignments from representative sequences of clusters 1-9.** The alignment clearly shows unique internal similarities between internal repeats (greater internal than external similarity between β-trefoils), indicating of separate origins from distinct precursor repeats. Corresponding Genbank IDs are listed in Table 4.2

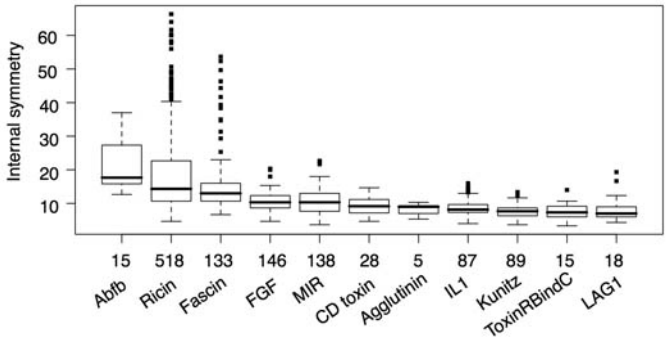


Figure 4.4: **Box-plot distributions of sequence symmetries for 11 β-trefoil families.** Sequence symmetry was calculated as the average percent identity between the three internal repeats for each β-trefoil domain. The rectangle shows the interquartile range (IQR); the line represents the median; the whiskers show the minimum and maximum unless there are outliers (1.5 x IQR), which are shown as points above or below the whiskers. The number of domains in each family is indicated on the x-axis.

When a simple BLAST search is not enough: an example

The most symmetric sequence overall (NCBI ID EAW37000) was a ricin domain from the cyanobacterium, *Lyngbya*. This sequence was analyzed in terms of its repeat pattern and similarity to other β-trefoils. The three repeats within this domain display a remarkable degree of sequence similarity (66% average sequence identity) (Figure 4.5A, bottom three repeats). The closest match to this domain in the NCBI database according to a BLAST search was another ricin type domain sequence (CAJ88326) (E = 2e-13) (Figure 4.5B). While the BLAST alignment occurs over the whole domain (Figure 4.5B), comparison of the individual repeats (Figure 4.5A) reveals that two domains are unlikely to be related by a whole-domain duplication. A more parsimonious explanation is that the proteins are composed of related, though clearly distinct repeat units (e.g., a nine-residue gap is common to all three repeats in CAJ88326). BLAST searches of the individual repeats from the EAW37000 sequence detected the other two internal repeats as the top ranked hits (E = 6e-12, 3e-08 from EAW37000 repeat 1 to repeats 2 and 3, respectively) while the following hits had E-values ≥ 0.7. This result suggests that EAW37000 has likely resulted from a unique triplication event.

Structural comparisons of repeats also provide evidence of internal repetition

Two of the smallest sequence clusters (#6 and #7) correspond to known protein structures (PDB IDs 1VCL, 2IHO), which allows for comparison of internal repeats at the structural level. These structures are both sugar-binding, hemolytic lectins from different species (Uchida et al., 2004; Grahn et al., 2007). These two proteins had the highest sequence symmetries among a dataset of 59 known β-trefoil structures (see Methods) and are clearly built from two related, but highly distinct, precursor peptides (see repeat alignment in Figure 4.6A). To determine whether evidence of separate fold-generating repeat events for these proteins could also be seen in their 3D structures, the repeat structural similarities were analyzed in these and 57 other β-trefoil structures from the PDB. Each individual repeat was structurally aligned with all other repeats, and the quality of alignment was quantified (see Methods). As shown in Figure 4.6B, there is a striking degree of internal structural symmetry within 1VCL and 2IHO, which rank first and second, respectively, out of all 59 β-trefoil structures in terms of average structural similarities between internal repeats. In addition, the two sets of repeats are significantly more alike internally than

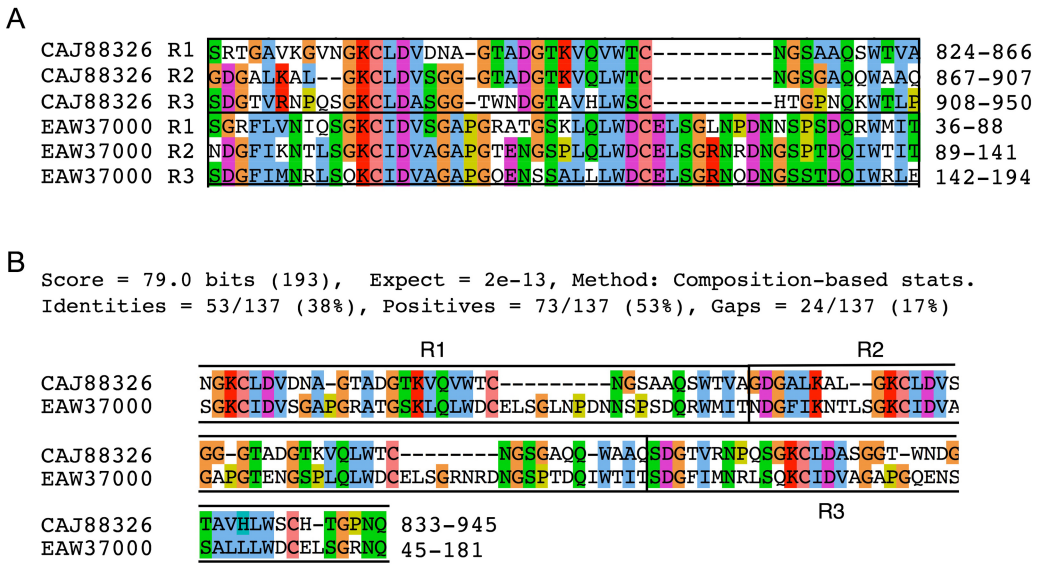


Figure 4.5: **Patterns of internal versus external sequence similarity for an identified highly symmetric sequence (NCBI accession # EAW37000).** (A) Three internal repeats within the EAW37000 sequence and its top BLAST match, CAJ88326. (B) BLAST alignment of EAW37000 and CAJ88326, with sequence similarity detected over the whole β-trefoil domain. The sequences are unlikely to be related via whole-domain duplication because both sequences contain unique sequence symmetry patterns (shown in A), which is indicative of two separate origins from unique precursor repeats. The internal repeats as shown in (A) are also indicated by vertical lines in (B).

they are to each other (Figure 4.6B) or to repeats in other β-trefoils (Figure 4.6C). Thus, the structure-based results are consistent with sequence clustering (Figure 4.2), and indicate that 2IHO and 1VCL have arisen from two separate fold-generating repeat events. The extent that the different ancestral repeats have maintained their unique structural characteristics in each structure is remarkable.

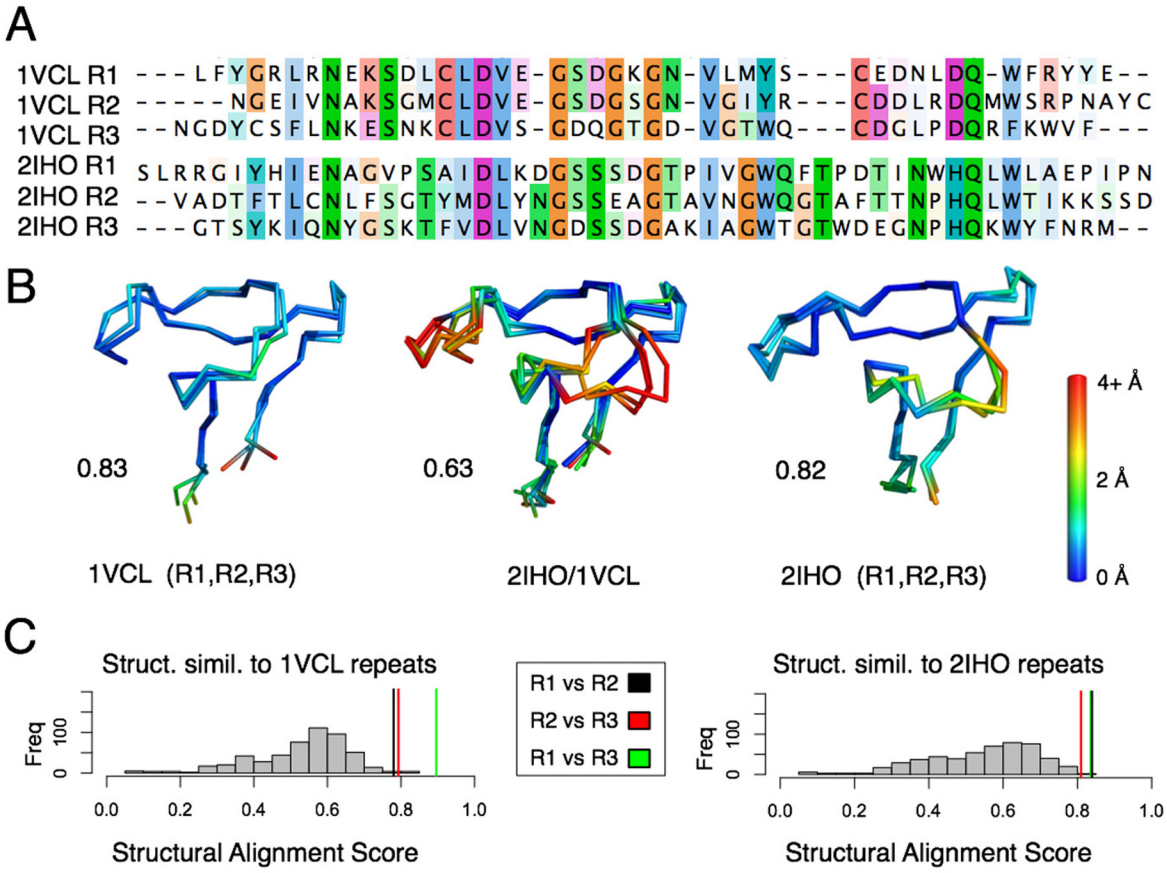


Figure 4.6: **Structural symmetry in 2IHO and 1VCL.** (A) Sequence alignment of internal repeats from 2IHO and 1VCL (domain 2), showing greater similarity within each sequence than between sequences. (B) Structural alignments of internal repeats within (left, right) and between (middle) both structures. The average structural alignment score between repeats is shown, with a score of 1.0 indicating identity. The alignments show a striking level of structural similarity for repeats within either structure, but a much lower similarity for repeats between structures, consistent with the sequence-based results. (C) Distribution of structural alignment scores between the three repeats in 1VCL (left) or 2IHO (right) and repeats from 58 other structures. Structural similarities between internal repeats in both structures are marked on the x-axis, and are significantly greater than the structural similarities with repeats from other structures.

4.1.4 Conclusion and Discussion

In summary, it has been shown through comparative analysis of sequence and structure that there have been multiple, independent evolutionary events that have generated β-trefoil domains by repetition of different peptide precursors. The uniqueness of these peptide building blocks can be seen not only at the sequence level, but in the 3D protein structures. These results ultimately demonstrate that smaller elements of protein structure than domains (supersecondary structural elements) can act as more fundamental units of duplication and inheritance in the evolution of a protein fold. This ongoing evolutionary process of repeat-mediated fold regeneration is likely an evolutionary mechanism for maintaining symmetry in the face of sequence divergence, and exists due to a selective pressure on symmetry-related protein functions.

Most common protein folds (Orengo et al., 1994) display some degree of internal structural symmetry. The internal repeats in these symmetric folds is thought to reflect their evolutionary origins by duplication and fusion of smaller, peptide elements (Söding and Lupas, 2003; Lupas et al., 2001; Andrade et al., 2001). It has been suggested that these peptides are relics of an "ancient peptide world" (Söding and Lupas, 2003). The cases of fold-generating repeats identified in this study are recent, ongoing examples of this process in action. This is consistent with a recent study of the β-propeller fold which identified β-propellers assembled by "amplification" of individual blades (repeats) (Chaudhuri et al., 2008). These results suggest that peptide "building blocks" of protein folds are not only limited to the distant evolutionary past, but continually used by evolutionary processes. The identified mechanism of fold-regeneration from single repeats may apply to numerous symmetrical fold families.

Why might some proteins continually evolve from repeated peptide elements? There are a number of potential functional and structural advantages of proteins that evolve in this way. If symmetry is important for either function or folding, the perfect or near-perfect symmetries in newly constructed β-trefoils may provide a selective advantage. Indeed, carbohydrate-binding β-trefoils are known to exhibit multivalent binding (i.e., multiple repeats binding to a common carbohydrate substrate) (Hazes, 1996; Boraston et al., 2004). This symmetry-related function is thus a selective pressure for duplication at the subdomain as opposed to the domain level. There may also be a link between symmetry and folding, as theoretical work has shown that funnelled energy landscapes, a requirement of

foldable proteins, are more easily achieved if the protein structure is symmetric (Wolynes, 1996). If symmetric protein structures evolve only by domain duplication, however, there is a tendency for symmetry to decrease over time as repeat sequences will naturally diverge from one another. Evolutionary conservation of symmetry will be less efficient if it is reliant on single-base substitutions. However, recurring formation of stable folds from repeats may, in considerably less evolutionary steps, regenerate symmetry, foldability, and create a repetitive and stable scaffold upon which old functions are amplified and new functions can evolve.

This work was part of a larger collaboration with Dr. Elizabeth Meiering (Waterloo) and Aron Broom (Waterloo), which, while not described in this thesis, resulted in the experimental confirmation of this evolutionary model through protein design. A 47-residue single repeat sequence was designed based on repeats from the most symmetrical sequence identified in the dataset (NCBI accession # AAV45265). For sequence design, a combination of consensus design, homology modeling, and free energy minimization was used. The repeat sequence was then tandemly triplicated to result in a final sequence of the designed protein (3-fold assembly of β-trefoil repeats, 3ABR). 3ABR had 73% identity to the template sequence and 40% identity to a β-trefoil of known structure (PDB ID, 2IHO). NMR and circular dichroism (CD) analyses confirmed that the reconstructed fully symmetric protein (3ABR) was well folded and very stable, and the CD spectrum most closely resembled that of another known β-trefoil structure (PDB ID 1AVW, STI family) according to the SELCON3 algorithm. The combined experimental results have demonstrated the physical viability of the identified recurring evolutionary process of fold-generation.

4.2 The PPODs from *Hydra*

In this section, one of the examples identified in the previous section (cluster 2 in Figure 4.2) is further analyzed to characterize its evolutionary history and function. The analysis suggests that it, like other identified β-trefoil families that have undergone internal repetition, likely plays a role in multivalent carbohydrate-binding. This family is an interesting example whose evolutionary relationships and function are incorrectly annotated according to overall domain similarity, but whose function and evolutionary origin can be revealed by more detailed analyses of internal repeats, motifs, and structural comparisons.

4.2.1 Introduction

Cnidaria, the sister group of the Bilateria, are one of the earliest diverging metazoan phyla (∼600 million years ago) (Ferrier and Minguillón, 2003), and are among the first to have such uniquely metazoan features as a simple body plan, a nervous system, and stem cells (Bosch, 2007). These features have made the cnidarian, *Hydra*, a prime model organism for studying the genetic and evolutionary basis of metazoan development (Wittlieb et al., 2006).

Numerous studies have examined the genetics of *Hydra* development through characterization of molecular and gene expression markers at different stages of *Hydra* morphogenesis. Studies have focused on morphology and development of the foot and particularly the basal disc, a site of substantial cell differentiation and regeneration (Hoffmeister and Shaller, 1985; Thomsen and Bosch, 2006). A combination of biochemical, histochemical, and molecular biological methods have led to the identification of two foot-specific putative peroxidases (PPOD1 and PPOD2) whose corresponding gene expression matched exactly the localization of the foot-specific peroxidase activity (Hoffmeister-Ullerich et al., 2002). Subsequent studies of PPODs have identified additional paralogs of the PPOD family in several species of *Hydra*, and revealed PPODs to be part of a multigene family. Homologs of PPODs have not been found in other metazoan genomes, however. The function and evolutionary origins of PPODs, which appear to be a "taxon-restricted" gene family, therefore, remain unclear.

In this work, a phylogenetic and structural analysis of the PPOD protein family from *Hydra* is performed in order to characterize its function and evolutionary history. Phyloge-

netic analysis suggests that the PPODs are likely a taxon-restricted protein family, which originated in an ancestral *Hydra* species by lateral transfer of a bacterial gene. Sequence analysis indicates that the gene family underwent a succession of internal duplication events that generated a unique internal repeat pattern, and possibly novel function in *Hydra*. An Evolutionary Trace like approach is applied to identify a conserved, repeated surface motif. This, combined with experimental information, identified a putative binding site and function of PPODs.

4.2.2 Methods

Sequence retrieval, domain parsing, and phylogenetic analysis

The first β-trefoil domain (residues 1-130) of PPOD1 from *H. vulgaris* was used as a query in a BLASTp search to retrieve related sequences (default parameters were used: Blosum62 matrix, Gap Existence = 11, Gap Extension = 1, Conditional compositional score matrix adjustment). All sequences that aligned to $\geq 90\%$ of the query and received E-values ≤ 0.01 were selected, redundancy was reduced using a 99% sequence identity threshold, and the sequences were aligned using MUSCLE (Edgar et al., 2004). Conserved regions of the alignment were then selected, and a Bayesian inference of phylogeny was performed using MrBayes (Ronquist and Huelsenbeck, 2003) with the WAG model plus gamma (4 rate categories). The corresponding domain structure of each full-length protein was also mapped onto the resulting tree. Domain annotations were retrieved from the NCBI Conserved Domain Database (CDD), with the exception of the PPOD-like domain, which was modeled manually using a profile Hidden Markov Model (HMM) approach. A profile HMM of PPOD-like β-trefoil domains was constructed using hmmbuild with the alignment generated above as input, and this HMM was used to identify the positions of all PPOD-like β-trefoil domains in each full-length sequence (default parameters, E-value cutoff = 0.01).

Repeat prediction

The program RADAR (Heger and Holm, 2000) was used for sequence-based prediction of internal repeats with default parameters.

Sequence logos

Sequences logos were generated for individual repeats as well as the whole PPOD sequence using the WebLogo server (Crooks et al., 2004).

Structural analysis

In the Evolutionary Trace (Lichtarge et al., 1996) based approach, evolutionary conservation (as determined by a multiple alignment in Jalview) was mapped onto the protein surface using PyMol.

Principal components analysis

Jalview was used to perform principal components analysis (PCA) based on the alignment of individual repeats from selected PPODs and a subfamily of related repeats from *Bacillus*. The components are generated by an eigenvector decomposition of the sum of BLOSUM scores matrix at each aligned position for each sequence pair. The method is further described in Casari et al. (1995). An in-house R script was developed for 3D visualization of the PCA results.

4.2.3 Results

Detection of internal PPOD repeats and fold recognition

As a starting point for the analysis, the sequence of PPOD1 (NCBI accession # AAZ31364) from *Hydra vulgaris* was analyzed to characterize its domain structure and select regions for subsequent database searching. This protein, a "long-form" PPOD, is composed of six internal repeats (or modules as described in Hoffmeister-Ullerich et al., 2002). A "short-form" PPOD (NCBI ID AAZ31368) composed of three internal repeats has also been identified. The automated annotation generated by the NCBI's Conserved Domain Database is shown in Appendix 4. Two fascin domains, which correspond to a known family of eukaryotic actin-bundling proteins (Kureishy et al., 2002), have been automatically annotated for the PPOD sequence based on overall detected sequence similarity.

To characterize the fundamental unit of duplication within PPODs, the RADAR program (Heger and Holm, 2000) was used to predict the internal repeat structure of PPOD1 and identify repeat borders. Consistent with Hoffmeister-Ullerich et al's (2002) analysis, RADAR predicted six repeats in the long form and three in the short form PPODs. The sequence alignment and amino acid conservation between repeats is shown in Figure 4.7A for PPOD1. The short form, PPOD2, is the most similar to the last three repeats of PPOD1, and thus appears to be the result of a truncation. The predicted repeats encompass the full-length PPOD1 sequence with the exception of the first ~20 residues, which encodes a signal peptide. PSI-BLAST searches of the signal peptide region did not identify significant similarity with other proteins outside of the *Hydra* PPOD family, which indicates that this signal sequence is also highly unique. It possesses characteristics of other secretory signal peptides from animals, however, as predicted by WolfPSORT (http://wolfpsort.org).

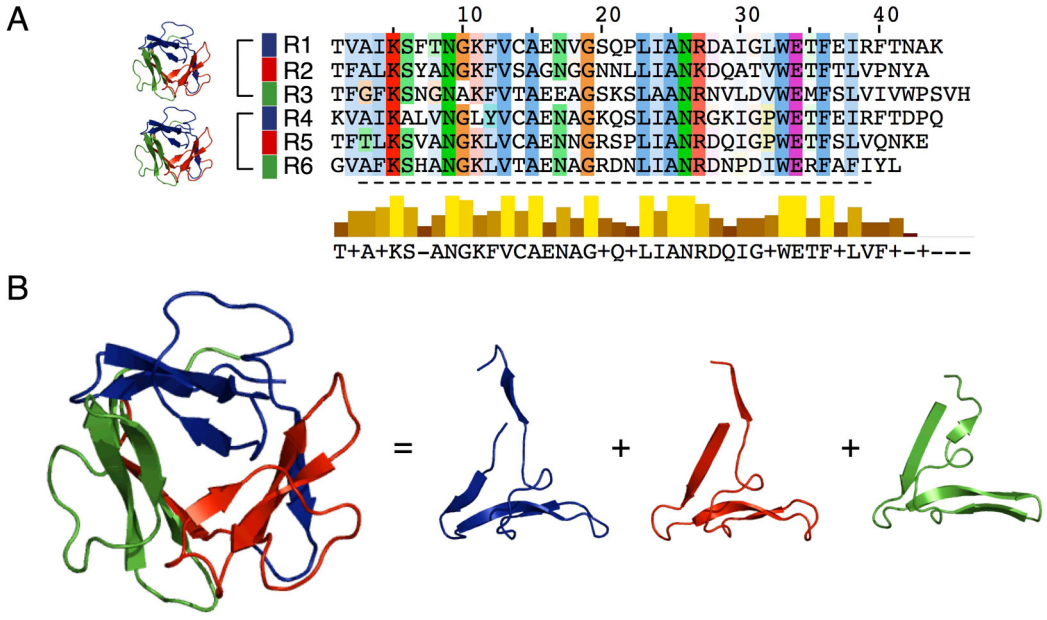


Figure 4.7: **Structural model and internal repeats of a representative PPOD domain.** (A) Six internal repeats of a long-form PPOD1 protein. Three repeats form a putative β-trefoil domain as shown in (B). Each repeat (colored uniquely) corresponds to a β-β-β-loop-β supersecondary structural element.

The PHYRE (Protein Homology/analogY Recognition Engine; Kelley and Sternberg, 2009) server was then used to predict the fold of PPOD1, and model of how the detected

internal repeats map to structural domains. All detected structures in the Protein Data Bank (PDB) matching regions of PPOD1 (estimated precision of 95%) had the β-trefoil fold (Murzin et al., 1992). The top three detected template structures had an estimated precision of 100% and were all β-trefoil domains from the crystal structure of human fascin (PDB ID 1DFC). In the majority of significant models, the first three repeats and last three repeats were predicted as two separate, contiguous β-trefoil domains. Shown in Figure 4.7B is a model of a single PPOD β-trefoil domain based on the last three repeats from PPOD1. As revealed by mapping the internal sequence repeats onto the structure, each repeat corresponds to a β-β-β-loop-β supersecondary structural element. Three of these structural elements assemble together to form a single β-trefoil fold (Murzin et al., 1992) (Figure 4.7B). PHYRE's predictions using the short form PPOD sequence or using the first three repeats from PPOD1 both produced the same overall result.

Phylogenetic analysis of PPOD and related proteins

Homologs of PPOD family proteins within genbank (nr protein database) were retrieved, and aligned using MUSCLE. Conserved regions of the alignment (full alignment shown in Appendix 5) were used to generate a Bayesian phylogenetic tree using MrBayes (Figure 4.8). The domain architecture of each corresponding protein, as defined by the NCBI's Conserved Domain Database, was also mapped onto the phylogeny. As shown by the resulting consensus tree (Figure 4.8), both domains from the two-domain PPODs are found in a single clade, separate from all other sequences, with perfect clade support [Bayesian posterior probability (BPP) = 1.00]. The ancestral two-domain PPOD gene has thus most likely resulted from a unique tandem duplication that possibly occurred within the *Hydra* lineage itself. Most of the PPODs share a common domain structure of two consecutive β-trefoil domains (6 repeats), with a few exceptions (e.g., NCBI ID XP_002164969).

The group of PPOD domains are further nested within a large group of bacterial sequences, and this is further outgrouped by several plant sequences. The clustering of PPODs within bacterial PPOD-like sequences is perfectly supported (BPP = 1.00), and was also found by other phylogenetic methods (neighbor-joining, parsimony and maximum-likelihood) (data not shown). This result is inconsistent with that expected from speciation, and suggests that the PPOD gene family instead originated from a bacteria-to-*Hydra* lateral gene transfer event. The notion that *Hydra* has obtained genes of bacterial origin

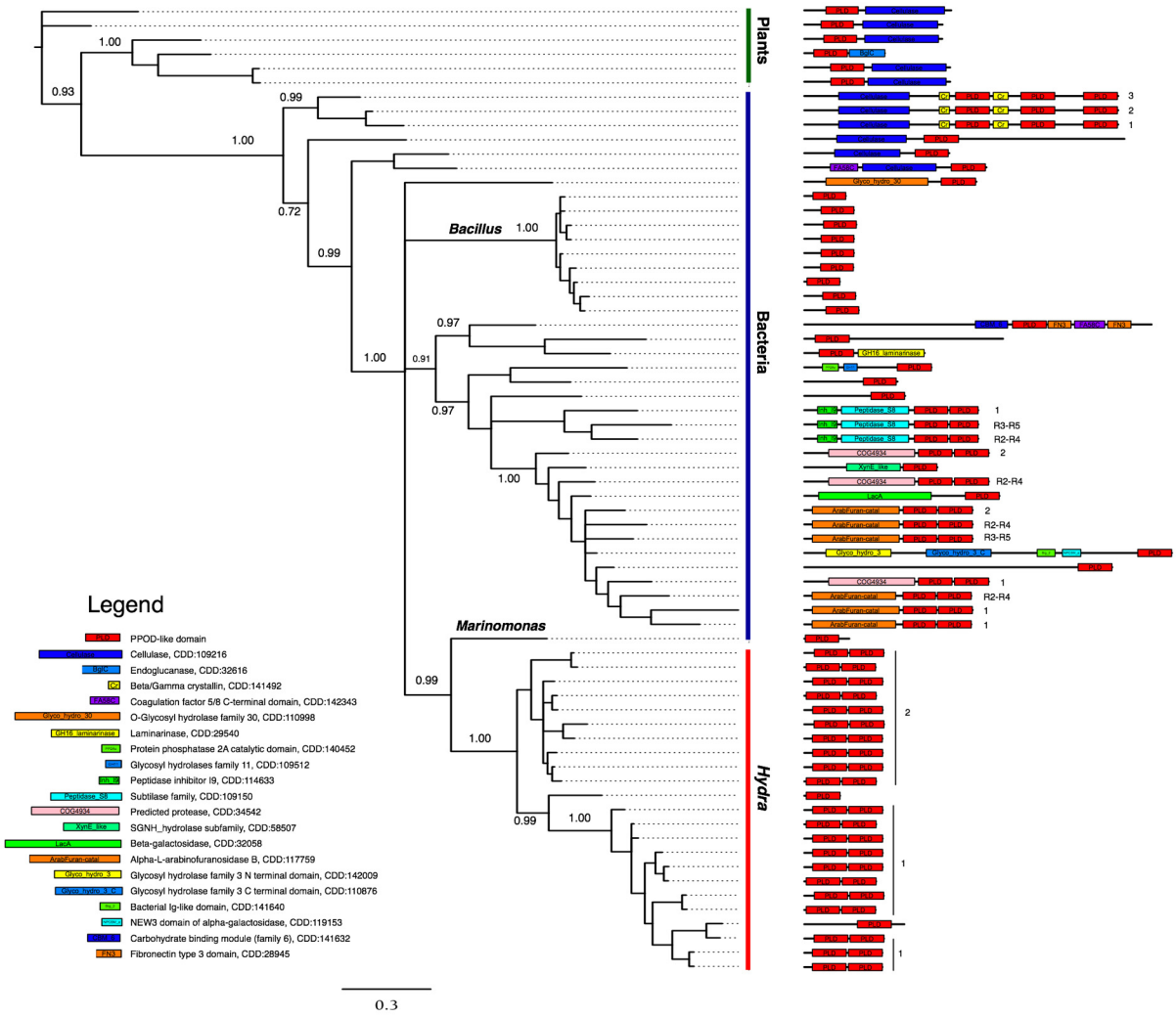


Figure 4.8: **Bayesian inference phylogenetic tree of PPODs and related domains, and domain architectures of corresponding proteins.** Posterior probabilities are indicated above the major clades, and the domain structure of the full-length protein associated with each PPOD-like domain is shown on the right. Where appropriate, numbers or repeat numbers (e.g., R2-R4) indicate which β-trefoil domain or consecutive set of three repeats was used in the alignment.

is consistent with its known close association with a diverse range of bacterial symbionts (Fraune and Bosch, 2007). A good candidate for the type of ancestral bacterial domain that could have been a progenitor of the PPOD gene family is the *Marinomonas* PPOD-like sequence (NCBI ID ZP_01074204). As shown in Figure 4.8, this protein is most closely associated with the PPOD group (this is extremely well supported, BPP = 0.99), it is from an aquatic species of bacteria, and it has a single domain structure that could have generated the ancestral two-domain PPOD through tandem duplication.

While the *Marinomonas* PPOD-like sequence encodes a single domain, most of the other PPOD-like domains from bacteria and plants are components of larger, multi-domain proteins, most of which have carbohydrate-related functions. For instance, the PPOD-like domains outside of the *Hydra* group are attached to such carbohydrate-related enzymes as cellulase, glycosyl hydrolase (GH) family 30 (glucosylceramidase), GH family 16 (laminarinase), GH family 11 (O-glycosyl hydrolase), GH family 42 (beta-galactosidase), GH family 3 (O-glycosyl hydrolase), and Arabinofuranosidase (Figure 4.8). In many of these cases, the PPOD-like domain has replaced the N- or C-terminally located binding domains normally associated with these enzymes, suggesting that it may act as a binding domain in these proteins, most likely targeting particular carbohydrates. There are a few cases in which PPOD-like domains are not fused to carbohydrate-related enzymes however (e.g., Peptidase S8), whereby a role in carbohydrate-binding appears less likely.

Concerted evolution of internal repeats

Individual repeats in PPODs and other repetitive β-trefoil domains (e.g., as in section 4.1) are examples of concerted evolution at the subdomain (internal repeat) level. To illustrate this, PPOD repeats were compared in greater detail to repeats within a related subfamily of domains from *Bacillus*. The domain family is labeled in the phylogenetic tree in Figure 4.8. This *Bacillus* family of fascin-like proteins is, like the PPODs, highly repetitive, contains detectable internal repeats, and has been maintained in *Bacillus* through several speciation events. The *Hydra* PPOD and *Bacillus* repeats are closely related between both families, with some repeats sharing more than 50% sequence similarity. However, as shown by the neighbor-joining tree in Figure 4.9A, the first major cluster contains all repeats from the *Hydra* PPOD family, while the second cluster contains all repeats from the *Bacillus* PPOD-like family; this topology was supported by a NJ bootstrap value of 786 out of

1000 replicates. According to the tree, the two domain families can be traced back to two different precursor repeats, which likely triplicated separately in each lineage.

For an alternative visualization of this result, an in-house sequence analysis module for R[2] was used to generate a 3D plot of repeats using principal components analysis (PCA) (Figure 4.9B). More closely related repeats are located closer together in 3D space, while distantly related repeats are located farther apart. Domains that are related through whole-domain duplications should produce a cluster of similar triangles; a cluster of points for repeat 1, and separate clusters for repeats 2 and 3, respectively. However, a domain that has resulted from a separate triplication event should produce a distinct triangle in 3D space because its three repeats are closest to one another, and not the repeats in other sequences. As shown in Figure 4.9B, the two repeat families produced two distinct clusters in 3D space. Ultimately, the domains in Figure 4.9 have either originated by whole-domain duplications of *Bacillus* fascin-like proteins or whole-domain duplications of the *Hydra* PPODs, and the families have a distinct symmetry pattern because the ancestral sequences were formed by separate triplicate repeats. While the higher order branching pattern of the tree in Figure 4.9A cannot be unambiguously resolved, an analysis of self-similarity for a representative single 6-repeat PPOD protein (NCBI ID AAZ31365) suggests that the ancestral 6-repeat gene evolved by a duplication of an ancestral 3-repeat protein (data not shown).

[2]This R module was designed in collaboration with Owen Woody (University of Waterloo).

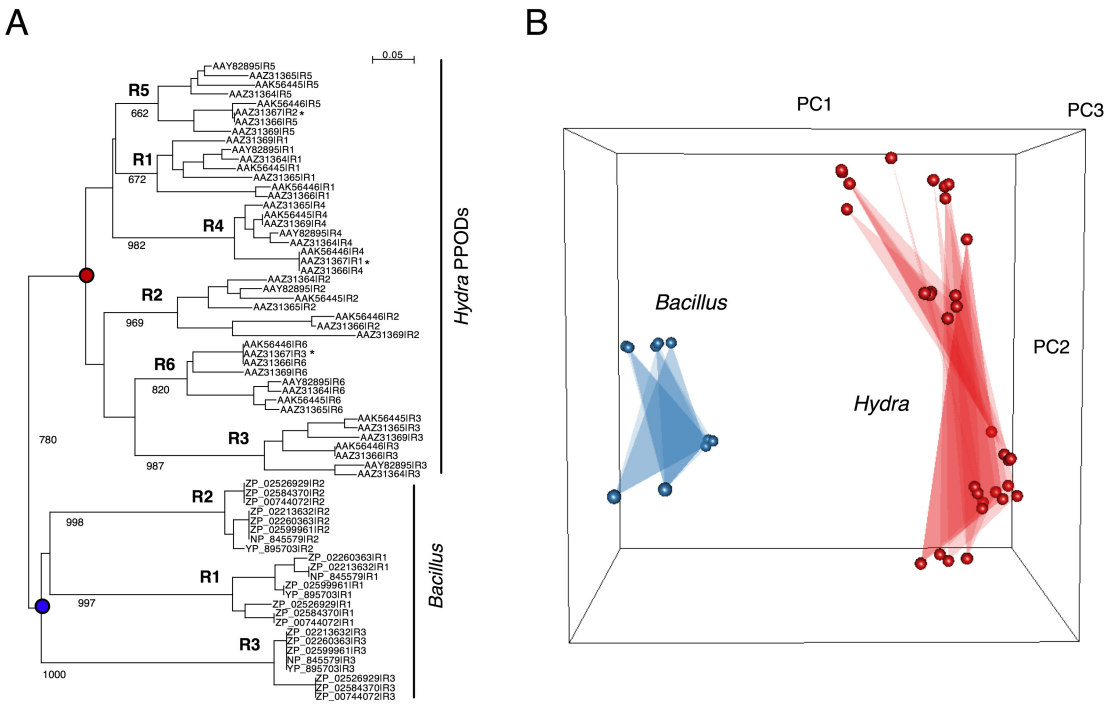


Figure 4.9: **Phylogenetic comparison of internal repeats in the *Hydra* PPOD family versus repeats in the *Bacillus* fascin-like protein family.** (A) Neighbor-joining tree of repeats R1-R6 from PPODs, and R1-R3 from *Bacillus* fascin domains. Bootstrap values for 1000 replicates are indicated. Asterisks refer to repeats from the short-form PPOD protein. (B) Principal components analysis of repeats based on sequence similarities.

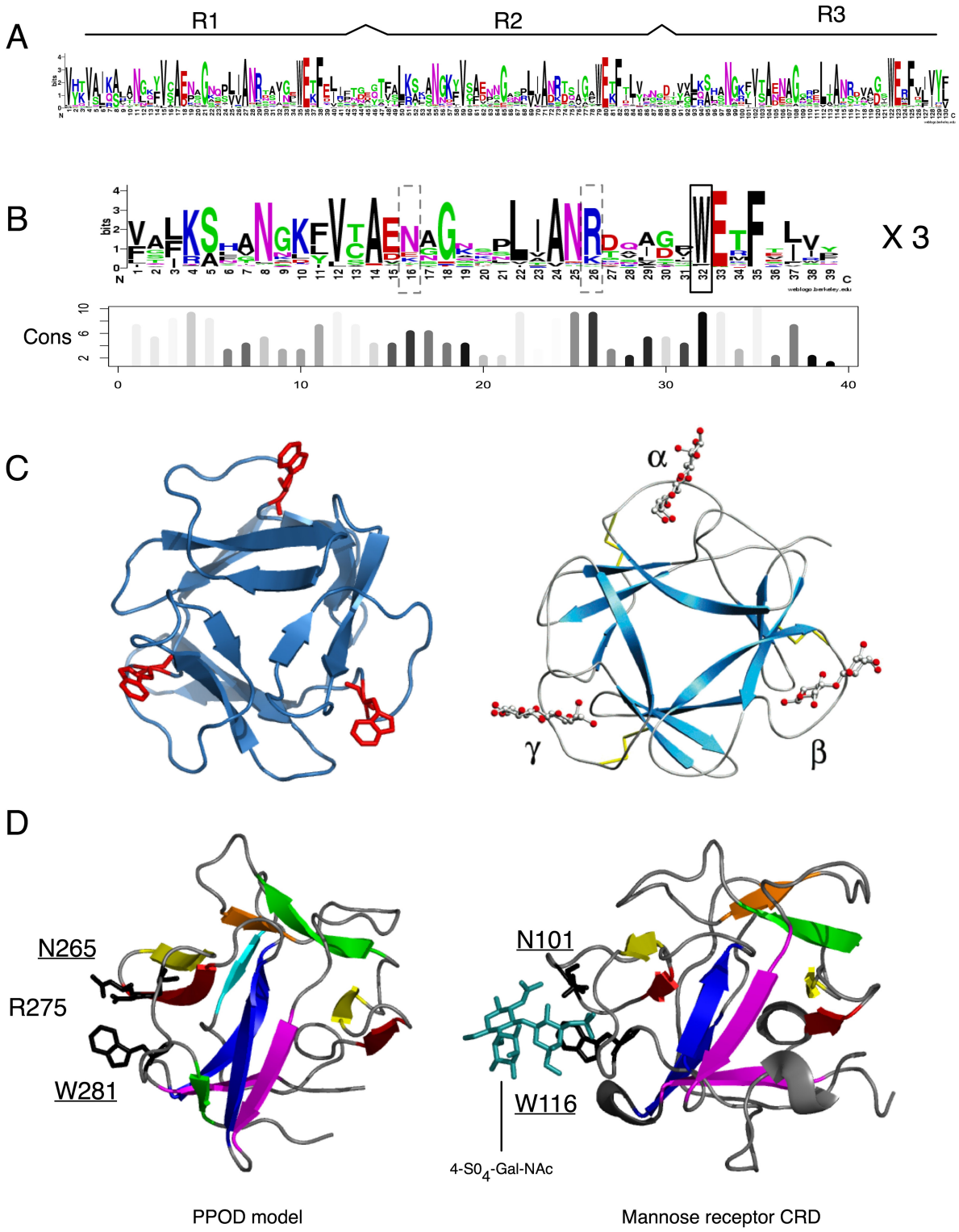
Structural modeling of PPODs and identification of the putative sugar-binding site

The sequence conservation patterns of PPODs were then examined in light of the structural model to identify conserved surface residues that may be of functional importance. As the profiles for each individual repeat (Figure 4.10A) are highly similar/redundant and indicate similar functional pressures, the repeats were combined into a single alignment and represented in the form of a sequence logo to obtain the overall sequence conservation pattern for the single internal repeat (Figure 4.10B). Below the logo the degree of conservation has been plotted and each bar has been coloured by solvent accessible surface area as inferred from the structural model. The residues that are both highly conserved and highly accessible represent likely functional sites. The top site that satisfies both of these conditions is the Trp residue found in the WEXF motif at the C-terminal end of each repeat (boxed in Figure 4.10B), which is interesting since Trp is a rare residue to be found on the protein surface but is the most overrepresented residue in carbohydrate-binding sites (Malik and Ahmad, 2007) where it plays a crucial role in binding to sugar rings. The importance of surface Trp motifs in sugar-binding was also demonstrated earlier in section 2.2.

The PPOD model (with the Trp residues highlighted as three putative sugar-binding sites) was then compared to the structure of the CBM13 domain from *Streptomyces lividans*. The CBM13 domain from *S. lividans* also adopts a β-trefoil fold, and exhibits multivalent binding to three lactose molecules, one molecule of lactose bound to each repeat (Notenboom et al., 2002). Strikingly, when the two structures are structurally aligned

---

Figure 4.10 *(following page)*: **Integrative sequence and structure-based modeling of sugar-binding sites in PPODs.** (A) A sequence logo of a single PPOD domain (three repeats). (B) A sequence logo generated by combining all repeats into a single alignment. Below the sequence logo, the conservation is plotted and each bar is colored according to solvent-accessibility (max = black; min = white). C) Left, a structural model of a single PPOD domain with the three putative sugar-binding Trp residues highlighted in red. Right, the structure of *S. lividans* CBM13 bound to three lactose molecules; adapted from Notenboom et al. (2002). (D) Left, PPOD model indicating the three putative sugar-binding residues described in the text. Right, structure of the CRD from the mannose receptor bound to sulfated 4-S04-Gal-NAc (PDB ID 1FWU). Identical residues between these two structures that are found in the same structural position are underlined.


A
R1
R2
R3
B
bits
X 3
Cons
0
10
20
30
40
C
α
γ
β
D
N265
R275
W281
N101
W116
4-S0$_4$-Gal-NAc
PPOD model
Mannose receptor CRD

(Figure 4.10C), the three lactose-binding sites in CBM13 match perfectly the positions of the three Trp residues in the PPOD model. Moreover, the lactose molecules bound by CBM13 are also in contact with a Trp and two Tyr residues (a common substitution for Trp in sugar-binding sites) in equivalent positions to the three Trp residues in the PPOD model.

The cysteine-rich, carbohydrate-recognition domain (CRD) of the mannose receptor is a β-trefoil lectin-domain known to bind sulfated sugars (Liu et al., 2000). While CRD does not apparently possess multivalent binding activity like *S. lividans*, it contains a single bound molecule of 4-S04-Gal-NAc. It was then compared to the corresponding site in PPODs by manual structural alignment. Again, the Trp residues in PPODs (i.e., W281 in Figure 4.10D) match a corresponding Trp residue (W116) in the CRD of the mannose receptor that is in direct contact with the bound sulfated sugar. Furthermore, another residue (N101) cited as an additional sugar-binding residue is shared in PPODs (N265) in the same structural position (this residue is also boxed in Figure 4.10B). Finally, a third residue that is unique and highly conserved in the PPOD repeats but is not found in CRD is R275. Given its conservation and proximity to W281 and N265, it is possible that this residue may make additional contacts with the sugar ligand.

4.2.4 Conclusion and Discussion

The previous annotation for PPODs within the NCBI as detected by overall sequence similarity is the fascin domain, which has a known role in actin-bundling in higher eukaryotes (Kureishy et al., 2002). The phylogenetic analysis done here demonstrates that PPODs have actually originated by lateral gene transfer from, most likely, a bacterial source. Lateral transfer from a bacterial source is consistent with known bacterial symbiotic relationships of *Hydra* (Fraune and Bosch, 2007). Moreover, structural analysis suggests that, while the PPOD family has a greater degree of overall sequence similarity with the fascin family, the likely functional motifs, which are well conserved in each repeat, are more similar to known binding sites from R-type lectin β-trefoil families with little to no detectable sequence similarity. It is possible that both binding sites have evolved independently in both lineages. Ultimately, this provides another example of a situation in which overall similarity poorly captures function, but where key motifs (i.e., a few residues) may represent a fundamental determinant of function.

NOTE: experimental confirmation

Böttger et al. (unpublished) have demonstrated that PPODs are not peroxidases, but carbohydrate-binding proteins with specificity towards sulfated sugars. Their experimental work has confirmed the structural analyses presented here.

Chapter 5

Recombination, mosaicism and extreme divergence

The following chapter contains published material from Doxey et al. (2008b)[1].

In the previous chapter, it was demonstrated that protein functional inference and classification based on domain similarities can be problematic when internal repeats are not properly accounted for. These smaller elements of structure can act as units of duplication instead of whole domains, and thus may in some cases represent even more basic units of protein function and evolution. In such cases, subdomains of a protein sequence can have independent evolutionary histories from the rest of the sequence (in which case it is called a *mosaic* sequence), and need to be assessed individually for an accurate characterization of that sequence's evolutionary history and function.

In this chapter, two additional examples of mosaic sequences are investigated that undergo recombination and evolve extremely rapidly. These cases represent a considerable challenge for homology-based bioinformatic approaches, because their evolutionary relationships are essentially scrambled by recombination and extreme sequence divergence.

In an analysis of the *Clostridium botulinum* neurotoxin (section 5.1), this problem is approached by analysis of the neurotoxin gene cluster as a whole. While individual sequence

[1]The included article is "Insights into the evolutionary origins of clostridial neurotoxins from analysis of the *Clostridium botulinum* strain A neurotoxin gene cluster", *BMC Evolutionary Biology*, 2008, 8, 316, from BioMed Central.

similarities within the neurotoxin gene cluster do not tell the full story, an evolutionary and functional picture emerges upon identification of recurring sequence, structural, and functional similarities identified by analysis of the neurotoxin as well as adjacent genes.

In the second section of this chapter, I explore the evolution of the flagellin hypervariable region (HVR) in detail, to uncover its history and patterns of functional divergence. Recombination, lateral gene transfer, mosaicism, repeats, and gene loss are observed. The results suggest that flagellin functions cannot necessarily be inferred based on phylogenetic analysis of the whole sequence or by species identification, because their function-determining HVRs can evolve independently of adjacent domains.

5.1 The Clostridium neurotoxin gene cluster

5.1.1 Introduction

Clostridial neurotoxins (CNTs) are the most poisonous biological toxins known and molecular agents of botulism and tetanus neuroparalytic diseases (Montecucco and Molgo, 2005). Due to their extreme toxicity and potential threat as bioterrorism agents, they are listed as Category A agents by the Centers for Disease Control and Prevention along with other deadly agents such as anthrax. Elucidating the mechanisms by which CNTs evolved is therefore of significant importance to our understanding of pathogen evolution and emerging diseases.

While considerable progress has been made in understanding CNT structure and function (Simpson, 1980; Blasi et al., 1993; Lacy et al., 1998; Hanson and Stevens, 2000; Swaminathan and Eswaramoorthy, 2000; Dong et al., 2006; Jin et al., 2006; Chai et al., 2006), like many toxins and virulence factors, the evolutionary origins of CNTs are unclear. CNTs are produced by four phylogenetically distinct groups (I-IV) of *C. botulinum*, and also by strains of *C. tetani*, *C. baratii*, and *C. butyricum* (Collins and East, 1998). As demonstrated by the scattered phyletic distribution of neurotoxin-producing clostridia (Collins and East, 1998) and the patterns of sequence similarity between different neurotoxin gene clusters (Jacobson et al., 2008), CNT genes appear to have undergone significant lateral transfer between different species of *Clostridium*. The occurrence of lateral transfer is also supported by the discovery of plasmid-encoded neurotoxin genes in numerous *C. botulinum*

strains (Marshall et al., 2007), as well as the existence of putative insertion sequences flanking the neurotoxin gene cluster (Dineen et al., 2003). Evidence of recombination has been found both between strains and within numerous genes in the neurotoxin gene cluster (East et al., 1996; Moriishi et al., 1996; Dineen et al., 2003; Hill et al., 2007; Hill et al., 2009). Mosaic sequences encoding the NTNH and neurotoxin genes have been reported (Moriishi et al., 1996; East et al., 1996), which is indicative of intragenic recombination events.

While CNTs have undergone frequent lateral transfer between species of *Clostridium*, no CNT homologs have been identified outside of the *Clostridium* genus. CNTs form an isolated protein family according to SCOP (Murzin et al., 1995) and PFAM (Finn et al., 2006) and have a unique structural architecture that complicates the identification of related proteins and potential ancestors. Its unique architecture may stem from a rich history of intragenic recombination events and/or the generation of novel domains through extreme sequence divergence/drift. While CNT domains have little detectable sequence similarity to proteins outside of the CNT family, there are however some structural and functional similarities to other domain families. The β-trefoil, a three-fold symmetrical structure that forms the C-terminal receptor binding domain (HCRc) and associated hemagglutinin-components, is common to interleukins, ricin-like lectins, and fibroblast-growth factors (Murzin et al., 1992). The adjacent HCRn domain, also involved in receptor binding, forms a jelly-roll like structure similar to laminin globular G domains (Lacy et al., 1998). The central translocase adopts a long α-helical structure containing α-helical bundles that resemble those found in translocase-like domains of other toxins (Wiener et al., 1997). Lastly, the N-terminal catalytic domain has been grouped under the zincin-like group of metalloproteases by SCOP and under the Peptidase MA clan by the MEROPS database (Rawlings et al., 2006). It contains a HEXXH zinc-binding motif found in other zinc endopeptidases, but has only weak structural similarity to other members of the Peptidase MA clan (Hanson and Stevens, 2000).

Diversity of domain and fold composition and extreme sequence divergence are common features of bacterial toxins (de Maagd et al., 2003). Rapid sequence evolution and recombination in toxin genes (see de Maagd et al., 2003) is largely a consequence of the evolutionary "arms race" between pathogen and host. Therefore, it is important to consider that evolutionarily related toxins may only share weak sequence similarity and may have undergone considerable structural rearrangements. Thus, insights into the origins of CNTs may be gained through analysis of adjacent sequences, which may possess shared

functions, and/or shared domains due to ancestral recombination events.

In this work, the neurotoxin gene cluster as a whole is examined for recurring links to common ancestral functions. Numerous links involving sequence, motif, and structural similarities for proteins within and outside of the neurotoxin gene cluster are identified. Multiple lines of evidence suggest that the neurotoxin and adjacent genes evolved from an ancestral collagenase-like gene cluster, linking CNTs to another major family of clostridial proteolytic toxins. Duplication, reshuffling and assembly of neighboring genes within the BoNT/A neurotoxin gene cluster may have led to the neurotoxin's unique architecture. The detected links provide novel insights into the evolutionary origins and ancestral function of the neurotoxin gene cluster.

5.1.2 Methods

Sequence dataset and database searches

Botulinum neurotoxins A-G (P10845, ABM73983, BAA08418, AAB24244, CAA43999, 1904210A, CAA52275), and NTNH/A (YP_001253341) sequences were retrieved from NCBI. The flagellin and collagenase sequences used in the alignment of the HEXXH-containing segment were *Clostridium haemolyticum* flagellin [FliA(H)], BAB87738; *Pseudoalteromonas tunicata* flagellin, ZP_01132756; *Azoarcus* sp. flagellin, YP_934037; *Desulfuromonas acetoxidans* flagellin, ZP_01312630; and *Burkholderia pseudomallei* collagenase, ZP_01765667. Accession numbers for flagellin and collagenase sequences are BAB87738, ZP_01132756, YP_934037, ZP_01312630, and ZP_01765667. The following default parameters were used in all PSI-BLAST (Altschul et al., 1997) searches unless specified otherwise: Blosum62 matrix, Gap existence: 11, Gap Extension: 1, E-value cutoff = 0.005, with conditional compositional matrix score adjustment. The SSEARCH (Pearson and Lipman, 1988) program from the FASTA package (version 3.515) was used to search the *C. botulinum* protein database, and was obtained via the SANGER website (ftp://ftp.sanger.ac.uk/pub/pathogens/cb/). SSEARCH was run with default parameters, except for the -z 11 flag, which computes the regression by reshuffling the target sequence library (removing the influence of homologous sequences present within the genome). For searching additional *Clostridium* species, the following genomic datasets (protein sequences) were retrieved from the NCBI FTP server: *C. acetobutylicum* ATCC

824, *C. beijerinckii* NCIMB 8052, *C. botulinum* A ATCC 3502, *C. botulinum* A ATCC 19397, *C. botulinum* A Hall, *C. botulinum* A3 Loch Maree, *C. botulinum* B1 Okra, *C. botulinum* F Langeland, *C. difficile* 630, *C. kluyveri* DSM 555, *C. noyvi* NT, *C. perfringens* 13, *C. perfringens* ATCC 131245, *C. perfringens* SM101, *C. phytofermentans* ISDg, *C. tetani* E88, *C. thermocellum* ATCC 27405.

Construction of sequence similarity heat map

A perl program was written to generate a 2D sequence similarity matrix based on all-against-all Smith-Waterman alignment scores using 3615 sequences from the *C. botulinum* protein database. Proteins were ranked by E-values computed by the SSEARCH program with default parameters. The matrix consists of query sequences on the Y-axis, target database proteins on the X-axis, and data values correspond to percentile ranks. This approach was used to detect distant pairwise similarities within gene clusters that may reflect ancient gene duplication blocks. The matrix was visualized using Treeview version 1.1.1 (http://rana.lbl.gov/EisenSoftware.htm).

Permutation testing

The PRSS component of the FASTA package (Pearson and Lipman, 1998) was used for sequence reshuffling and the permutation test. The permutation reshuffling test calculates the optimal Smith-Waterman alignments of the first query sequence with N reshuffled versions of the second query sequence. The alignment score of the unshuffled sequences is compared to the distribution of scores obtained using the reshuffled query sequence, which is fit to an extreme value distribution. From this distribution, the probability that the observed alignment score could have resulted from a random sequence of the same composition is estimated. Default parameters were used and 1000 reshuffled sequences were used to generate the random distribution of alignment scores.

To detect potential compositional bias, the composition of CNTs and CBO0798 was analyzed relative to all protein sequences in *C. botulinum* strain A as a reference. One amino acid type, asparagine, was found to be significantly elevated in both CBO0798 and CNT sequences ($Z > 2$ standard deviations). To verify that PSI-BLAST hits from CBO0798 to CNTs sequences was not due to composition, all asparagine residues were removed from

CBO0798 and the top-scoring alignment detected via PSI-BLAST (*C. butyricum* BoNT/E), and permutation reshuffling tests were repeated using the altered sequences.

5.1.3 Results and Discussion

Ancient gene duplications within the BoNT/A neurotoxin gene cluster

A comprehensive analysis of pairwise sequence similarities was performed for all proteins encoded within the *C. botulinum* (strain Hall A, ATCC 3502) genome (Sebaihia et al., 2001), in an attempt to identify distant homologs of CNTs and possible sequence remnants of the evolutionary process by which CNTs originated. This initial analysis was limited to a single genome for a more sensitive detection of pairwise homologies using a restricted database, however subsequent searches were also performed using all available clostridial genomes. For the 3615 proteins encoded within *C. botulinum* (strain Hall A, ATCC 3502) (Sebaihia et al., 2001), a "heat map" of pairwise sequence similarity was constructed (see Methods) (Figure 5.1). For each pairwise alignment, the E-value and percentile rank relative to all other pairwise alignments was calculated using SSEARCH (Pearson and Lipman, 1998). When compared by percentile rank, the neurotoxin gene cluster stood out as a hot spot of local pairwise sequence similarities. The neurotoxin gene cluster can be seen as a distinct cluster of high-scoring pairs in the centre of the heat map region in Figure 5.1A. Based on both the percentile ranks and E-values for the pairwise alignments corresponding to these genes (Figure 5.1B), there are clear sequence similarities between multiple sequences within this region, including BoNT/A, non-toxic non-hemagglutinin (NTNH), the adjacent hemagglutinin (HA) components and the adjacent CBO0798 gene encoding a flagellin protein (NCBI accession YP_001253335). BoNT/A and NTNH produced the top-scoring alignments with each other out of 3615 proteins in *C. botulinum* strain A (E = 1e-22, 9e-24), an expected result given previously identified sequence similarities between BoNTs and NTNH (Arndt et al., 2005) as well as their virtually identical domain architecture according to the NCBI's conserved domain database annotation (e.g., for NCBI IDs ABP48106 and BAA90660). Surprisingly, the next highest match in both cases corresponds to the CBO0798 flagellin gene located immediately upstream of the neurotoxin gene cluster (Figure 5.2A). The associated E-values were 0.041 and 0.42 for BoNT/A and NTNH, respectively (Figure 5.1B). CBO0798 aligned with NTNH and BONT/A in two different CNT regions (I and II) (Figure 5.2B). Additional searches using the sequences

of CNTs from other strains also identified CBO0798 as the most consistent top ranked hit out of all *C. botulinum* strain A proteins with sequence identities between CNTs and CBO0798 ranging from 20-24%, and the strongest alignments involving region II of CNTs.

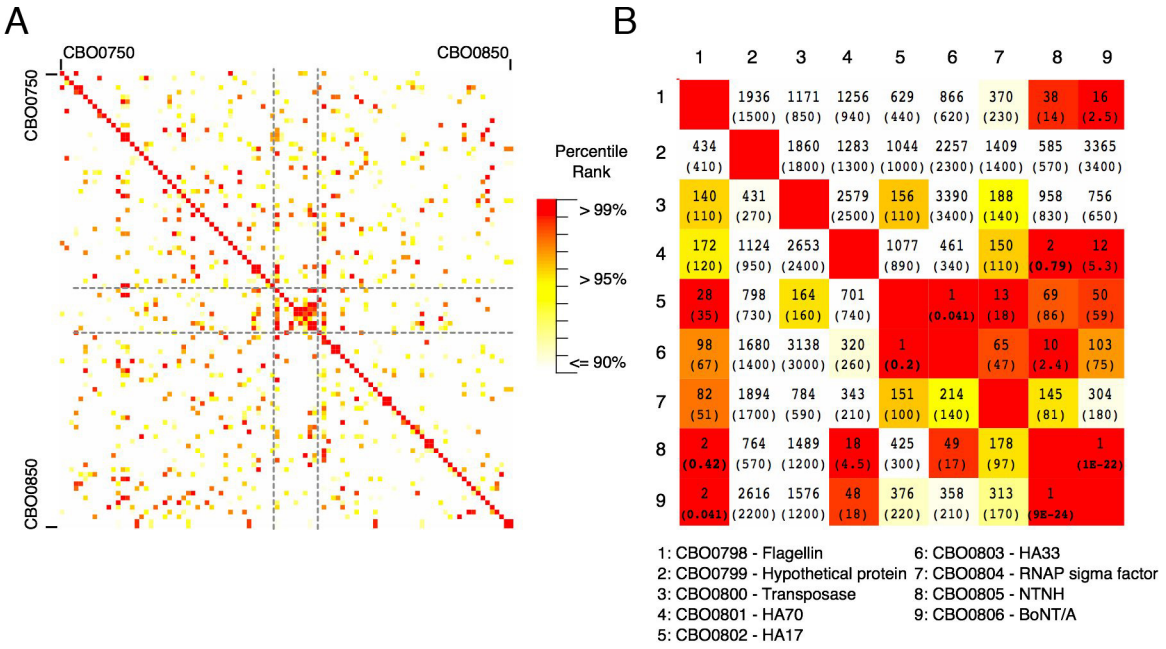


Figure 5.1: **Protein sequence similarity heat map surrounding the BoNT/A neurotoxin gene cluster.** Protein sequence similarity heat map surrounding the BoNT/A neurotoxin gene cluster. Sequence similarity scores, E-values and percentile ranks were calculated for all pairwise combinations of putative proteins encoded in the *C. botulinum* strain A genome. (A) A heat map of the percentile ranks for pairwise alignments involving 100 genes surrounding the neurotoxin gene cluster (described in Methods). A duplication "hot spot" corresponding to the neurotoxin gene cluster is apparent in the central region. (B) Similarity ranks and E-values (in brackets) for pairwise protein sequence alignments in the neurotoxin gene cluster, corresponding to BoNT/A, NTNH, CBO0798, associated hemagglutinin components and other neighboring genes. E-values < 1 are in boldface.

In addition to the detected similarities between CBO0798 vs. BoNT/A and NTNH vs. BoNT/A, sequence similarities were also detected between the β-trefoil hemagglutinin components (HA33 and HA17). HA33 and HA17 were identified as reciprocal top ranked matches (E = 0.041, 0.2), and a weak similarity was detected between HA33 and the C-terminal (beta-trefoil) regions of NTNH (ranked 10th, E = 2.4). Sequence similarity was also found between the hemagglutinin components HA70 and residues 39-474 of NTNH

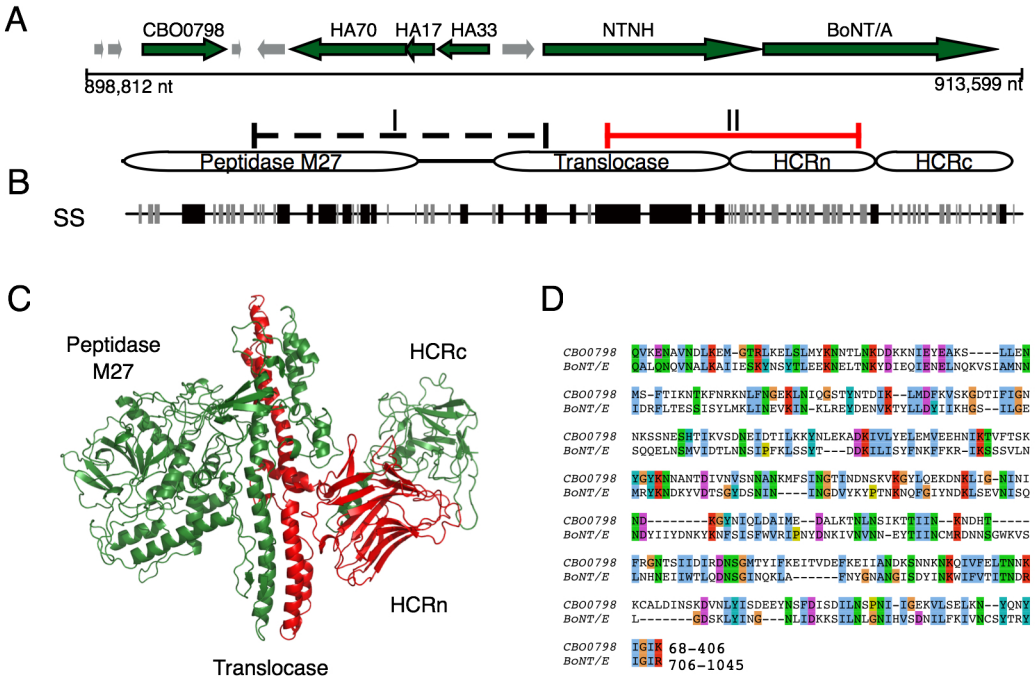


Figure 5.2: **Genomic location of flagellin CBO0798 and regions of sequence similarity with CNTs.** (A) Genomic context of the neurotoxin gene cluster for *C. botulinum* A. str. Hall. (B) Domain structure of CNTs and regions of detected similarity with NTNH (region I) and BoNT/A (region 2) according to SSEARCH. The CNT schematic in (B) is based on a multiple alignment of CNT-related sequences. CNT secondary structure is indicated below the schematic with black lines representing α-helices and grey lines representing β-sheets. (C) The structure of BoNT/A (PDB ID 3BTA) with region II highlighted in red. (D) A Smith-Waterman alignment of region II from *C. butyricum* BoNT/E and CBO0798.

(ranked #2 out of all pairwise alignments with HA70 as the query, E = 0.72). Though the E-values calculated above are not all statistically significant, the high-ranking scores relative to the 3615 *C. botulinum* proteins suggest that multiple genes within the BoNT/A neurotoxin gene cluster are likely distant homologs that have undergone extensive sequence divergence.

Sequence similarity to the upstream flagellin gene

To identify other clostridial sequences homologous to CBO0798, a PSI-BLAST (Altschul et al., 1997) search was conducted starting with the CBO0798 sequence (default parameters, results restricted to Clostridia). All homologs identified in the first iteration were members of the flagellin family. The second iteration identified additional flagellins, followed by the type E botulinum toxin (BoNT/E) from *C. butyricum* with an E-value of 0.05 [23% sequence identity over residues 88-406 of flagellin and 727-1045 (region II) of BoNT/E]. To check for the influence of composition on the alignment, two permutation reshuffling tests were performed, which calculate the probability that random sequences of the same composition could result in similar alignment scores. The permutation reshuffling tests detected significant sequence similarity between the two proteins with ($p = 0.0024$) and without ($p = 0.011$) statistically overrepresented amino acids included (see Methods).

According to the sequence alignments produced by PSI-BLAST and SSEARCH, the region of CNTs with the strongest detected similarity to CBO0798 (region II) includes most of the translocase domain as well as the HCRn domain (Figure 5.2B-D). Region I was also detected by SSEARCH, spanning the peptidase and "belt" region, though without definitive statistical significance ($E = 0.41$). The translocase, an extended α-helical domain, has a general structural similarity to the central helical regions of known flagellin structures (see PDB IDs 1io1, 2zbi, 2d4x). The β-domains of flagellin are highly variable however, and it is this variable region of flagellin that shares similarity with the HCRn domain of CNTs. As a structure is not available for the variable region of CBO0798, 3D-PSSM (Fischer et al., 1999) was used to predict the fold of CBO0798's central region. The structure for the CNT's HCRn jelly-roll domain was the top ranked structural match for this region (E=0.34), additionally supporting homology between the two proteins.

CBO0798 is annotated in the NCBI database as a member of the flagellar hook associated protein 3 (FlgL) family. This flagellin gene has been mentioned in previous CNT studies due to its close proximity to the neurotoxin gene cluster (Dineen et al., 2003) and its existence in numerous *C. botulinum* type A strains and associated plasmids (Smith et al., 2007). Flagellins are also known to have key roles in the virulence of bacterial pathogens (Ramos et al., 2004), have been shown by mass spectrometry studies to interact with CNT components (Hines et al., 2005), and possess previously unreported common structural features with CNTs (i.e., both contain a central region composed of extended

α-helices followed by β-rich domains (Lacy et al., 1998; Samatey et al., 2001)). These additional functional and structural links further support a potential evolutionary relationship between CBO0798 and CNTs.

Collagenase-like domains in the flagellin hypervariable region

Comparative sequence analysis of CBO0798 was performed by aligning CBO0798 to other flagellins from *Clostridium* species. According to the alignment, CBO0798 has a highly divergent central region containing a unique insert (residues 135-360), and this insert region comprises a large portion of CBO0798's alignments with CNTs. The existence of a unique central region within CBO0798 is not surprising, since flagellins are known to contain conserved regions at the N- and C-terminus but have a hypervariable central region that is structurally exposed on the flagellar surface (Ramos et al., 2004). As the structurally exposed region of the flagellar filament, the hypervariable region can interact with the host cell and is thus critical to flagellin-mediated virulence (Ramos et al., 2004). Interestingly, it is the variable region of CBO0798 that is central to the CBO0798-CNT alignments and that was predicted by 3D-PSSM to possess a jelly-roll fold similar to HCRn.

To characterize the origins of the insert, I examined similarly located inserts identified within the hypervariable region of a small number of additional flagellins from *Clostridium* species. While the sequences within the hypervariable region are highly divergent from one another as expected, one insert in particular [the insert of FliA(H) from *C. haemolyticum*] was identified to be both the largest insert and the only insert region with detected homology to other proteins using PSI-BLAST. FliA(H) is a relatively close homolog of CBO0798, as FliA(H) was the only flagellin detected using CBO0798's C-terminal region (residues 114-452) as a BLAST query sequence (E = 0.076). A PSI-BLAST search revealed that the hypervariable region of FliA(H) possesses significant similarity to microbial collagenases (E = 8e-04, iteration 2) and to the hypervariable regions of several flagellins from non-clostridial species (Figure 5.3). Remarkably, both the detected microbial collagenases and collagenase-like regions within the identified flagellins contain a HEXXH motif, the critical catalytic residues responsible for the CNT's zinc-endopeptidase activity. The alignment of CBO0798 with collagenase-containing flagellins and alignment of the HEXXH-containing segments from these flagellins, BoNT/B, and a representative microbial collagenase are shown in Figure 5.3.

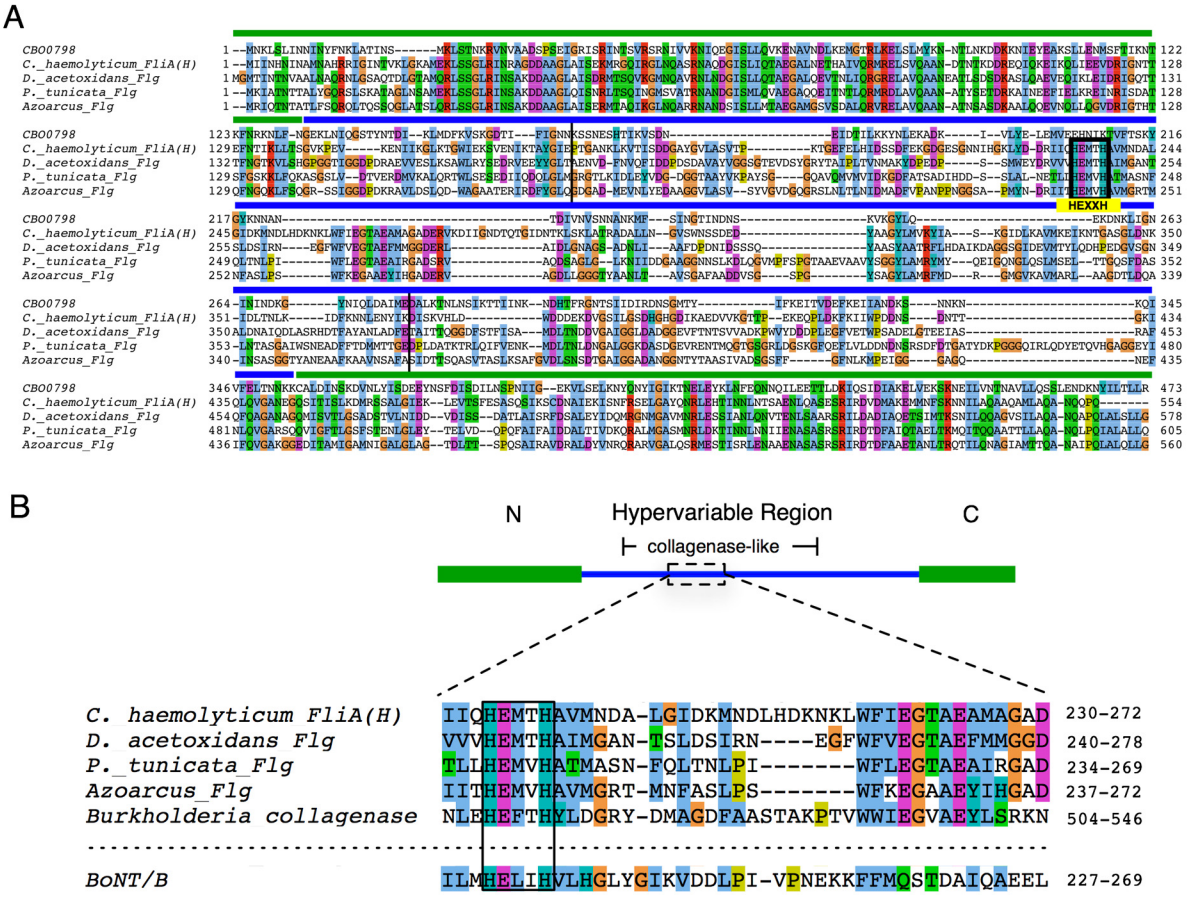


Figure 5.3: **Collagenase-like sequences within the flagellin hypervariable region.** A multiple alignment of CBO0798 and collagenase-containing flagellins identified by PSI-BLAST. The alignment was generated using MUSCLE with default parameters. The conserved N- and C-terminal regions are indicated by green bars above the alignment, and the hypervariable region is indicated by a blue bar. The first and second black vertical bars in the alignment correspond to the collagenase-containing region identified by a PSI-BLAST search using C. haemolyticum FliA(H) as the query. B) A schematic of a representative collagenase-containing flagellin [based on the FliA(H) sequence] showing the N- and C-terminal regions, hypervariable region, location of the collagenase and HEXXH-motif containing segments. An alignment of similar HEXXH-containing segments from BoNT/B, a microbial collagenase, and the collagenase-containing flagellins are shown below the schematic. Accession numbers are provided in the Methods.

The identified link to collagenase sequences by analysis of the flagellin hypervariable region is a striking result given the strong similarities between collagenases and the CNT's Peptidase M27 domain. Both collagenases (Peptidase M9s) and Peptidase M27s are zinc-endopeptidases and are grouped under the same peptidase family (thermolysin-like Peptidase MA clan) by the MEROPS database (Rawlings et al., 2006). As exotoxins, collagenases play a major role in clostridial toxicity by degrading collagenous host tissues (Hatheway, 1990; Harrington, 1996). For instance, *C. perfringens*, a species responsible for clostridial myonecrosis (gas gangrene), produces a tissue-degrading collagenase known as kappa-toxin (Harrington, 1996). Collagenases are therefore an excellent candidate evolutionary precursor of CNTs as both collagenases and CNTs function as clostridial toxins, and both share the same fundamental proteolytic mechanism.

As the hypervariable region encodes the outer exposed portion of the flagellin filament, it would be ideally situated to interact with (and potentially degrade) host cell wall components such as collagen. There may therefore exist a novel class of virulent flagellins with collagenase activity. Future experimental verification of this predicted activity would be valuable, and could potentially lead to a new avenue of research on bacterial virulence mediated in part by flagellins.

Additional evidence of collagenase-related functions within the neurotoxin gene cluster

Several additional links to collagenases and collagen-related domains were detected for other sequences present within the BoNT/A neurotoxin gene cluster. All sequenced *Clostridium* genomes were screened for potential homologs of each of the BoNT/A neurotoxin gene cluster components. In a dataset of over 55000 sequences, a search of BoNT/A detected flagellin as the third top ranked hit outside of the CNT family (E = 0.23). While HA33 expectedly displayed similarities with other ricin-like components (e.g., a ricin-domain from a *C. acetobutylicum* cellulase, NP_347343, E = 0.019), HA70 displayed the strongest similarity to *C. perfringens* enterotoxin (YP_697710, E = 0.0042) followed by *C. tetani* collagenase (NP_783761, E = 0.22). A HEXXH binding motif was also identified within this collagenase sequence. A PSI-BLAST search of flagellin CBO0798 restricted to the *Clostridium* genus also detected collagen-adhesion proteins with alignments spanning the hypervariable region after three iterations (E = 0.017, ZP_02635881). This result is consistent with the

analysis linking CBO0798 with flagellins containing collagenase-like hypervariable regions.

Another key result was obtained when examining sequence and structural similarities between the HCRn domain and the full NCBI nr database, including eukaryotic sequences. After two iterations starting with BoNT/A's HCRn domain, PSI-BLAST detected a region of chicken type XII collagen (AAA48635, E = 0.03). The detected sequence similarity occurred with collagen's thrombospondin N-terminal like domains. Recently, the structure of this family of domains has been determined for the NC4 domain of collagen X (Lepännen et al., 2007). The fold of NC4 (PDB ID 2UUR) is remarkably similar to that of HCRn (Figure 5.4). To determine the extent of structural similarity between these two domains, I analyzed structural neighbors of the NC4 domain using the VAST structural alignment algorithm (Gibrat et al., 1996). Sorted by VAST E-value, the two most structurally similar domains to PDB ID 2UUR were its identified fold family (the thrombospondin N-terminal domain) followed by the HCRn domain of the tetanus neurotoxin (E = 10e-9.9). Ranked by sequence similarity based on structural alignments, the tetanus HCRn domain ranked first out of all known structures in the Protein Data Bank (%ID = 16.8).

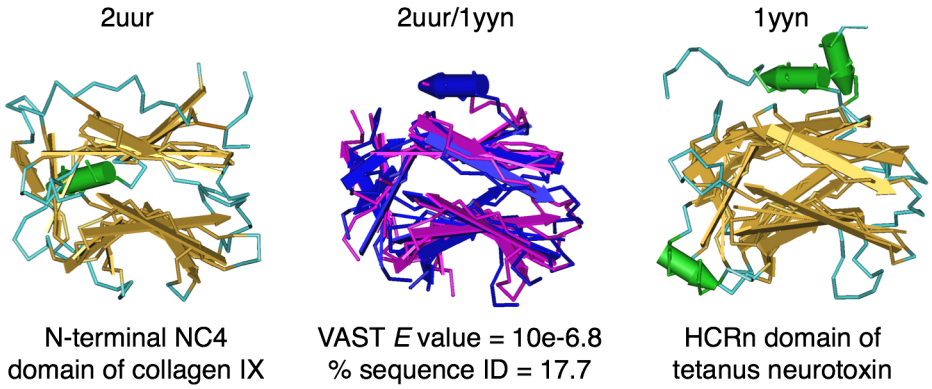


Figure 5.4: **Structural similarity between HCRn and the NC4 domain of collagen IX.** A structural superposition of the human collagen IX NC4 domain (2UUR) and the TeNT HCRn domain (1YYN) was performed using the VAST alignment algorithm (http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html). In the structural alignment, 2UUR and 1YYN are colored pink and blue, respectively.

As the detected similarities between the HCRn domain and the collagen NC4 domain occur across kingdoms, this may represent an instance of structural mimicry rather than a direct evolutionary relationship. Given the multiple identified links to collagenases, and

that structural mimicry of collagen has been proposed as a mechanism for other collagenase enzymes (De Souza et al., 1996), the link between HCRn and the collagen NC4 domain may be indicative of a similar mechanism. A role in collagen-binding is entirely possible for CNTs as previous studies have shown that expression of TeNT enhances adhesion of epithelial cells to collagen, laminin, and fibronectin (Proux-Gillardeaux et al., 2005). While the observed similarities support the hypothesis of convergent evolution and structural mimicry, the possibility that HCRn was transferred to *Clostridium* from a eukaryotic source cannot be completely ruled out. This scenario has been demonstrated recently for the *Clostridium* glyceraldehydes-3-phosphate dehydrogenase gene (Takishita and Inagaki, 2009).

5.1.4 Conclusions

A rich history of recombination events paired with rapid rates of mutation may have scrambled evolutionary relationships between sequences in the neurotoxin gene cluster and other proteins. Thus, the neurotoxin and other genes within the neurotoxin gene cluster appear highly unique and have an unclear evolutionary origin. The approach taken in this study was to independently assess multiple sequences within the neurotoxin gene cluster for evidence of common ancestral functions/sequences.

Multiple independent links to collagenase-related sequences were identified by analysis of the neurotoxin gene cluster, including the detected similarities involving the upstream flagellin gene (CBO0798) in the BoNT/A neurotoxin gene cluster, distant BLAST hits to collagenase-related domains, and detected structural similarities to the collagen NC4 domain. As microbial collagenases are phylogenetically widespread compared to CNTs, they represent a protein family that is likely ancestral to CNTs. Given this and the multiple detected links to collagenase-related sequences, it is proposed that an ancestral function of the neurotoxin gene cluster may have been related to collagen binding and degradation, a hypothesis that places CNT sequence, structure, and function within the broader context of other clostridial toxins and the evolution of clostridial pathogenesis.

While the links may be weak or even borderline significant individually, as a whole they represent a recurring pattern that suggests a common ancestral function for the neurotoxin gene cluster as a whole. The identified recurring functional links may come about by considerable sequence divergence of the neurotoxin gene cluster from an ancestral collagenase-like

gene cluster or through multiple recombination events of related sequences throughout the neurotoxin gene cluster, or most likely, through a combination of these two processes.

5.2 Tracing the evolution of the flagellin hypervariable region

Sequences encoding the bacterial flagellin protein are an extremely difficult case to analyze by standard homology-based methods. Flagellins possess three distinct regions, an N-terminal domain, C-terminal domain, and middle hypervariable region (HVR) which evolves at an extremely fast rate. Extreme rates of mutation make it difficult to align different HVRs. Furthermore, flagellins are known to undergo recombination, which must be accounted for to infer correct evolutionary relationships between sequences. Because the HVR is a surface-exposed region, it also has important functional characteristics. Thus, characterization of evolutionary relationships between HVRs is important for a proper characterization of flagellin function.

Here, large-scale patterns of flagellin HVR evolution are investigated by analyzing similarities at the protein sequence level and mapping these HVR similarities onto a phylogenetic tree based on the adjacent, conserved N-terminal domain. HVRs are also clustered into "types", which are reconstructed as discrete characters on the tree, thus allowing for visualization and detection of phylogenetic incongruencies that result from recombination and lateral gene transfer. The analysis reveals three recurring mechanisms underlying the evolution of flagellin HVRs: duplication and divergence of the flagellin_IN domain; frequent deletion/reduction of the HVR, which has occurred independently in numerous flagellin lineages; and intragenic recombination between flagellins and lateral transfer. The characterization of these patterns in a phylogenetic context provides a more detailed and accurate picture of flagellin evolution and function.

5.2.1 Introduction

The flagellin protein is the main structural component of the bacterial flagellar filament, known for its role in bacterial motility and virulence. All motile bacteria (including Spirochaetes in which flagella are referred to as axial filaments), depend on the flagellum for motion, which involves the action of a membrane-bound flagellar motor that drives rotation of an external filament and thus propels the bacterial cell through its external environment. In addition, the extracellular location of the flagellar filament and its potential

for developing adhesin-like properties makes it a key determinant of virulence in bacterial pathogens and a major immune elicitor (Smith and Ozkinsky, 2002).

Vast numbers of individual flagellin monomers, each of which consists of an N-terminal domain, C-terminal domain and a central hypervariable region (HVR), assemble together to form the hollow flagellar filament. The nature of the flagellin structure and how it assembles to form the filament, is intimately tied to the variable degrees of sequence conservation in these three domains. The N- and C-terminal domains interact with and stabilize each other and are found internally within the filament. The N- and C-terminal domains are thus highly constrained, and accept few mutations throughout evolution. The most conserved residues in these regions are involved in maintaining the structural integrity of the core filament and inter-subunit structural interactions [e.g., the universally conserved residue, Gly426 in *Salmonella typhimurium* flagellin (Samatey et al., 2001; Beatson et al., 2006)].

The flagellin HVR, however, located on the outer, solvent-accessible side of the filament, is relatively free from the structural constraints that are imposed on the N- and C-terminal domains. As the surface-exposed portion of the flagellar filament, the HVR is ideally positioned to interact with the external environment (Ramos et al., 2004). Indeed, flagellins have been found to play roles in host-cell adherence by pathogenic bacteria as well as biofilm formation (Yao et al., 1994; Gardel and Mekalanos, 1996; McGee et al., 1996; Rabaan et al., 2001). For instance, *Pseudomonas aeruginosa* flagellin has been found to bind mucin on epithelial cell surfaces (Arora et al., 1996; Lillehoj et al., 2002). Recent studies have also shown that many flagellins are glycosylated, which contributes to virulence by mediating host-cell binding and/or host immune evasion (Arora et al., 2005; Guerry et al., 2006; Ichinose et al., 2008).

As a unique component of bacteria, bacterial flagellin is not surprisingly a potent elicitor of host immune responses (Smith and Ozinsky, 2002). The O:H serotyping of pathogenic strains of *E. coli*, for instance, is based on the cell surface lipopolysaccharide O and the flagellar H antigen encoded by the *fliC* gene (Reid et al., 1999). The flagellin HVR is directly involved in immunogenicity, as the HVR forms the surface-exposed antigenic portion of the flagellin protein (Reid et al., 1999; Salazar-Gonzalez and McSorley, 2005).

The mammalian immune system has developed a strategy to recognize flagellin despite the rapid sequence evolution of its HVR. Key studies have shown that the mammalian pathogen-recognition protein, toll-like receptor 5 (TLR5), is dedicated to recognition of

bacterial flagellin (Hayashi et al., 2001) through interaction with the conserved, N-terminal domain (Smith and Ozinsky, 2002). However, major human pathogens including *Campylobacter jejuni*, *Helicobacter pylori*, and *Bartonella bacilliformis*, appear to have countered this strategy and are not recognized by TLR5 due to key mutations in the N-terminal domain (Andersen-Nissen et al., 2005). This example demonstrates the ongoing evolutionary arms race between flagellin and pathogen-recognition receptors of the host immune system.

Despite recognition of conserved segments in the N-terminal domain of flagellin by TLR5, it is the HVR of flagellin where this evolutionary arms race appears to take place most frequently. Rapid sequence evolution of the flagellin HVR is thought to provide antigenic variation and promote evasion of host immune responses (Smith et al., 2003; Mortimer et al., 2007). Two hypotheses have been put forth regarding the evolution of the flagellin HVR: "unconstrained evolution" by genetic drift; and lateral gene transfer and recombination (Reid et al., 1999). The latter is supported by numerous studies demonstrating flagellin recombination and mosaicism within and even between species (Frankel et al., 1989; Harrington et al., 1997; Smith et al., 1990; Smith and Selander, 1990; Wang et al., 2003). However, as most flagellin HVRs are too divergent to properly align, phylogenetic studies of flagellin HVRs have been largely restricted to closely related and thus *alignable* flagellins within closely related species.

In the previous section (5.1) on the neurotoxin gene cluster of the human pathogen, *C. botulinum*, the analysis identified a highly unique and interesting HVR possessing significant sequence similarity to microbial collagenases in flagellins from *C. haemolyticum* and three distantly related species (Doxey et al., 2008b). The unusual phyletic distribution of this HVR type and its distant similarities to other proteins suggested that: 1) the collagenase-like HVR may be an example of intragenic lateral gene transfer and recombination; 2) protein sequence similarities may be useful for investigating patterns of HVR evolution that are too divergent to study at the nucleotide level. In this work, large-scale patterns of flagellin HVR evolution are analyzed at the amino acid level. HVRs are clustered into related types, and each type is reconstructed on the N-terminal domain and C-terminal domain trees, which allows for the detection and visualization of phylogenetic incongruencies. These phylogenetic incongruencies cannot be accounted for by sequence drift alone, and outline a rich history of intragenic recombination, mosaicism and lateral transfer.

5.2.2 Methods

Domain parsing

The NCBI Conserved Domain Database (CDD) was used to retrieve 2,806 protein sequences of bacterial flagellin containing both the N-terminal domain (Flagellin_N, pfam id 00669) and the C-terminal domain (Flagellin_C, pfam id 00700). To remove sequence redundancy, the program CD-HIT (Li and Godzik, 2006) was used with default parameters, and resulted in a final non-redundant dataset of 1,224 unique flagellin sequences. The corresponding HMMs obtained from the PFAM database were then used to determine the regions encoding the flagellin N- and C-terminal domains within these sequences with the program HMMsearch from the HMMER package version 2.3.2 (default parameters). The hypervariable region (HVR) was defined as the sequence located between the identified N- and C-terminal domains, and was extracted in each sequence for subsequent analysis.

Phylogenetic tree construction

The N-terminal and C-terminal domains from flagellin sequences were aligned separately to their respective HMMs using the program HMMalign with default parameters. Poorly aligned sequences containing large gapped regions were removed from the analysis. Only the alignment positions corresponding to match states output by the HMM were used as sites in phylogenetic analysis. A maximum likelihood tree was then constructed using PhyML (version 3.0) with the following parameters: WAG model; 1 rate category; gamma distribution parameter = 2; branch lengths optimized. The approximate likelihood ratio test option was used to compute a measure of clade support (aLRT, SH-like branch support) (Anisimova and Gascuel, 2006).

Sequence analysis of HVRs and clustering

A combination of Perl scripts, R, and the program, Gablam (Davey et al., 2006) was used was used to conduct all-by-all BLASTp searches between HVRs, record alignment statistics, and generate a similarity matrix for visualization. The heat map was coloured according to the bitscore produced by BLAST.

The flagellin D3 domain and flagellin_IN domain were searched for in all HVRs using hmmsearch from the HMMER package, using HMMs from the HMMER 2.0 PFAM database. Searches were conducted with default parameters.

Due to extreme sequence dissimilarity, most HVRs could not be aligned. The multiple alignments that were attempted were far too "noisy" to use in phylogenetic analysis. Thus, as an alternative, simplified approach, the HVRs were clustered into related types based on all-by-all BLAST results. Each HVR was searched against all other HVRs using BLASTp (version 2.2.16). Default parameters were used and composition-based statistics were used to reduce compositional bias. If BLAST detected significant similarity (E-value $<$ 1e-10) between two HVRs, the two HVRs were considered related and therefore were grouped together in the same cluster. This process was repeated using all HVRs as queries. This resulted in a set of separate clusters, whereby each cluster contains sequences that are directly or indirectly related, while no or considerably less similarity exists between sequences from different clusters.

To account for the possibility of evolutionary relationships between flagellin HVRs and domains found outside of the flagellin family (for instance, due to insertions of unrelated domains into the flagellin HVR), PSI-BLAST searches were also performed using a representative from each cluster, performed against the full NCBI non-redundant protein database.

Reconstruction of HVRs according to cluster type

Each HVR cluster type was designated as a unique character state (A, B, C, etc.). As the aim of this study was to examine large-scale patterns of HVR evolution, it was acceptable to discretize and thus collapse clusters of HVRs into types, and instead examine the broader phylogenetic patterns of the clusters themselves. Moreover, the resolution of the analysis can be varied by increasing or decreasing the BLAST threshold used in the initial identification of clusters.

The HVR cluster types were then mapped onto the flagellin N- or C-terminal domain phylogenetic tree as character states and reconstructed using maximum likelihood. For the sake of simplicity, the default transition model (equal transition rates between states) was used. Reconstruction of ancestral states was performed using the implementation of DISCRETE within the APE package for R.

By examining the reconstructed likelihoods of presence/absence of each HVR type throughout the tree, it is then possible to determine if each HVR type has arisen once (monophyly) or multiple times (polyphyly) throughout the tree. In the former case, the most recent common ancestral (MRCA) node of all the sequences possessing the HVR type, should have a high likelihood of also possessing that HVR type, while in the latter case the MRCA node should have a very high likelihood of not possessing that HVR. This value was used as a simple but effective measure of homoplasy[2] for each state.

5.2.3 Results

Types of flagellin domain architectures

2,806 complete flagellin sequences containing both the flagellin N and C terminal domains were retrieved from the NCBI using the Conserved Domain Database (CDD). As shown in Figure 5.5, three common domain annotations are found in these sequences, as well as one unique set of three sequences with an extremely long HVR. The largest group (2,207 sequences) consists of an HVR that does not correspond to existing PFAM or CDD models. The second largest group (333 sequences) contains one or more copies of the "flagellin hook IN" or flagellin_IN domain (PFAM label), named because of a conserved central Ile-Asn pair. Two copies of this domain (as shown in the representative architecture) are most common, but some flagellin HVRs contain more than two copies or a single copy. The flagellin_IN domain is not only found in this domain architecture, but is also found between the N- and C-terminal regions of the flagellin hook associated protein 2 (HAP2 or FliD), hook associated protein 3 (FlgL) (Liu and Ochman, 2007), and in several other flagellin architectures. This is an important characteristic to mention as the flagellin_IN domain may therefore be an ancestral trait present in the ancestor of flagellins. While the function of the flagellin_IN domain is unclear, it does not appear to be necessary for structure formation and likely plays a flagella-specific functional role (Maruyama et al., 2008). A crystal structure of *Sphingomonas* sp. A1 flagellin containing the flagellin_IN domain has recently been determined (Murayama et al., 2008). The third largest group (263 sequences), which contains flagellins from Enterobacteriaceae according to NCBI annotations, possesses

[2]A homoplastic trait is one that has arisen independently in separate evolutionary lineages due to convergent or parallel evolution.

a unique HVR commonly referred to as the flagellin D3 domain. This domain, which has a known crystal structure (PDB ID 1UCU) is shown in Figure 5.5 (left) along with the N-terminal and C-terminal domains. No specific function has been assigned for the D3 domain, but it is known to be the major determinant of flagellar antigenicity in *Salmonella* (Sebestyén et al., 2008).

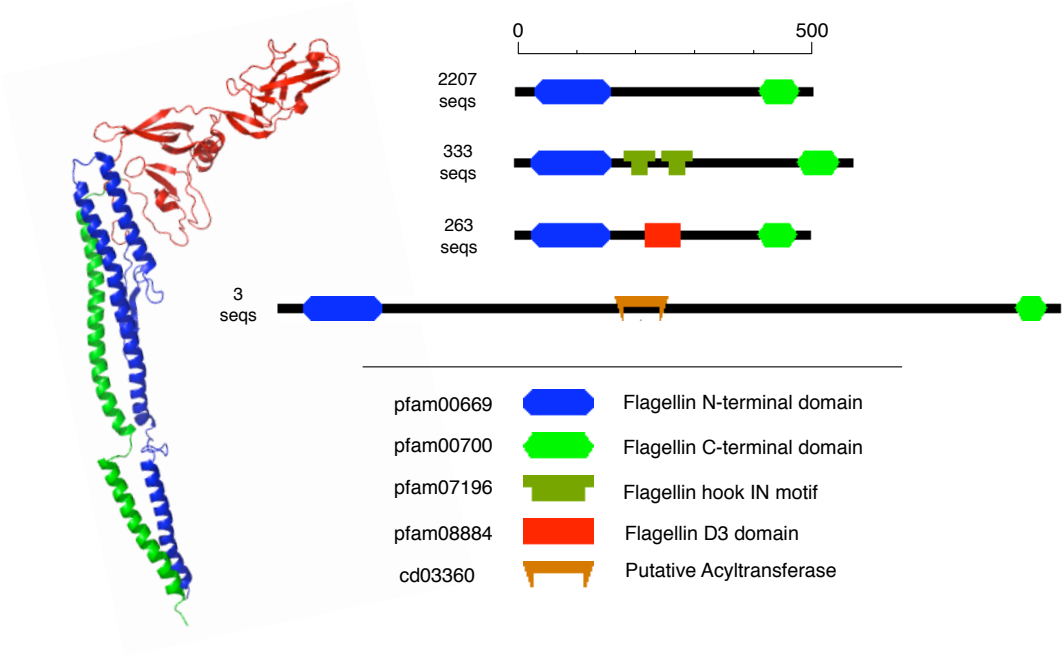


Figure 5.5: **Flagellin domain architectures investigated in this study.** An example flagellin structure (PDB ID 1UCU) is shown on the left that contains a flagellin D3 domain (red).

Phylogeny of the flagellin N-terminal domain

The full dataset of flagellin sequences was reduced to a set of 1,175 non-redundant sequences using CD-HIT, and aligned using HMMER. A maximum-likelihood phylogenetic tree (Figure 5.6) was constructed using a highly conserved region of the flagellin N-terminal domain (100 sites). A separate tree was also constructed based on a highly conserved region (72 sites) of the flagellin C-terminal domain. Separate analyses of the HVR (described in later sections) were performed by mapping HVR traits onto each of these trees, but results are presented only based on the N-terminal tree for consistency.

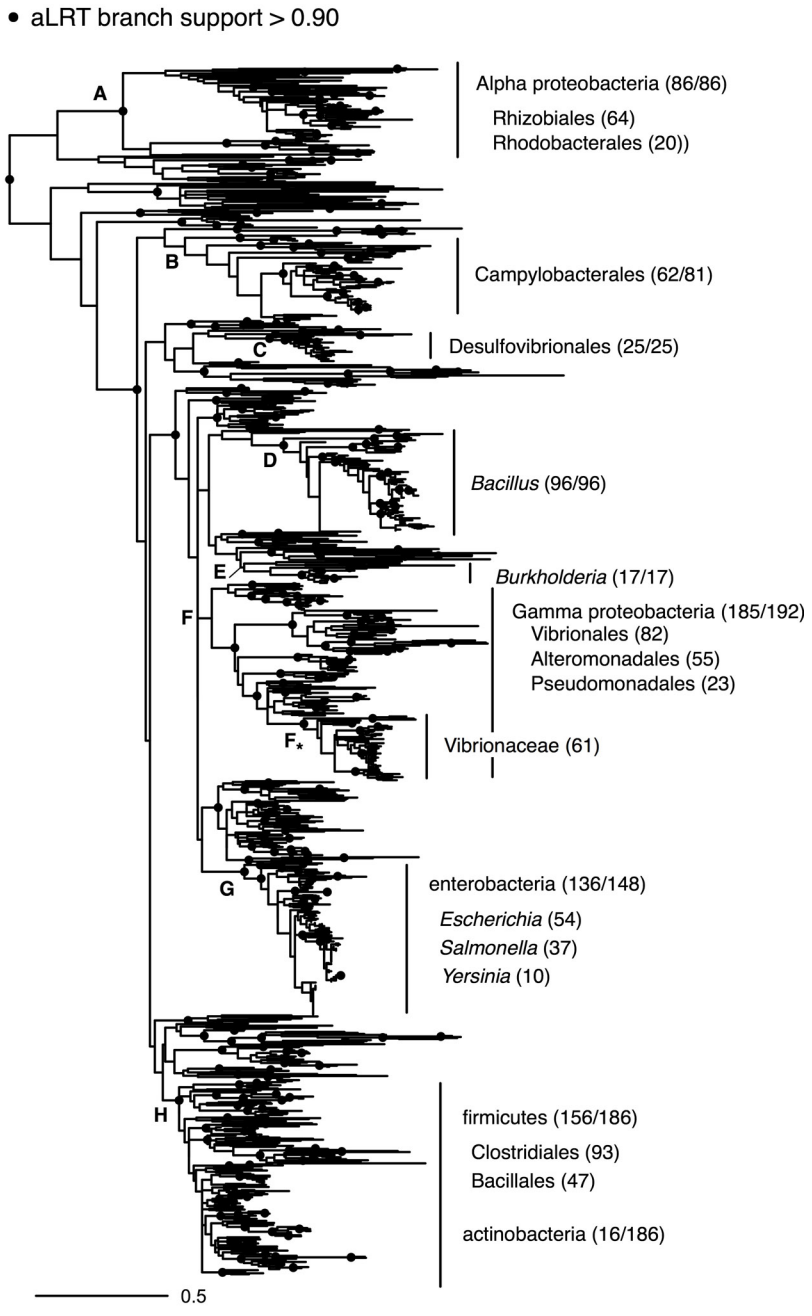


Figure 5.6: **Phylogeny of the flagellin N-terminal domain.**

The N-terminal domain tree is shown in Figure 5.6. Major clades are indicated along with a summary of the species of bacteria associated with each clade. Most of these are supported by aLRT (SH-like test) values of greater than 0.90. Eight large clades have been labelled "A" to "H", and each of these clades contain flagellins predominantly from a major phyletic group of bacteria. For instance, clade G corresponds to the earlier described group of Enterbacteriales flagellins containing the D3 domain. The most abundant genera in this clade are *Escherichia* (54), *Salmonella* (37), and *Yersinia* (10). Clade D on the other hand, contains flagellins exclusively from the genus *Bacillus*. Other notable clades include clade B which contains flagellins from Campylobacterales bacteria including the human pathogens *C. jejuni* and *H. pylori*, clade F* which contains a group of Vibrionaceae specific flagellins (clade F contains a broader selection of Gamma proteobacteria), and clade H containing predominantly flagellins from the firmicutes phylum (e.g., *Clostridium* and *Bacillus*). Clades that are unlabeled contain flagellins from a wider array of phyla that cannot be as easily summarized.

Evolution of the flagellin HVR: frequent HVR deletion, and duplication of the flagellin_IN domain

Once the flagellin N-terminal domain tree was constructed, the next question was how the HVR, the important virulence-related and antigenic determinant of flagellins, has evolved in the context of this tree. Do flagellins with most closely related N-terminal domains, also have the most similar HVRs or is the pattern more complex?

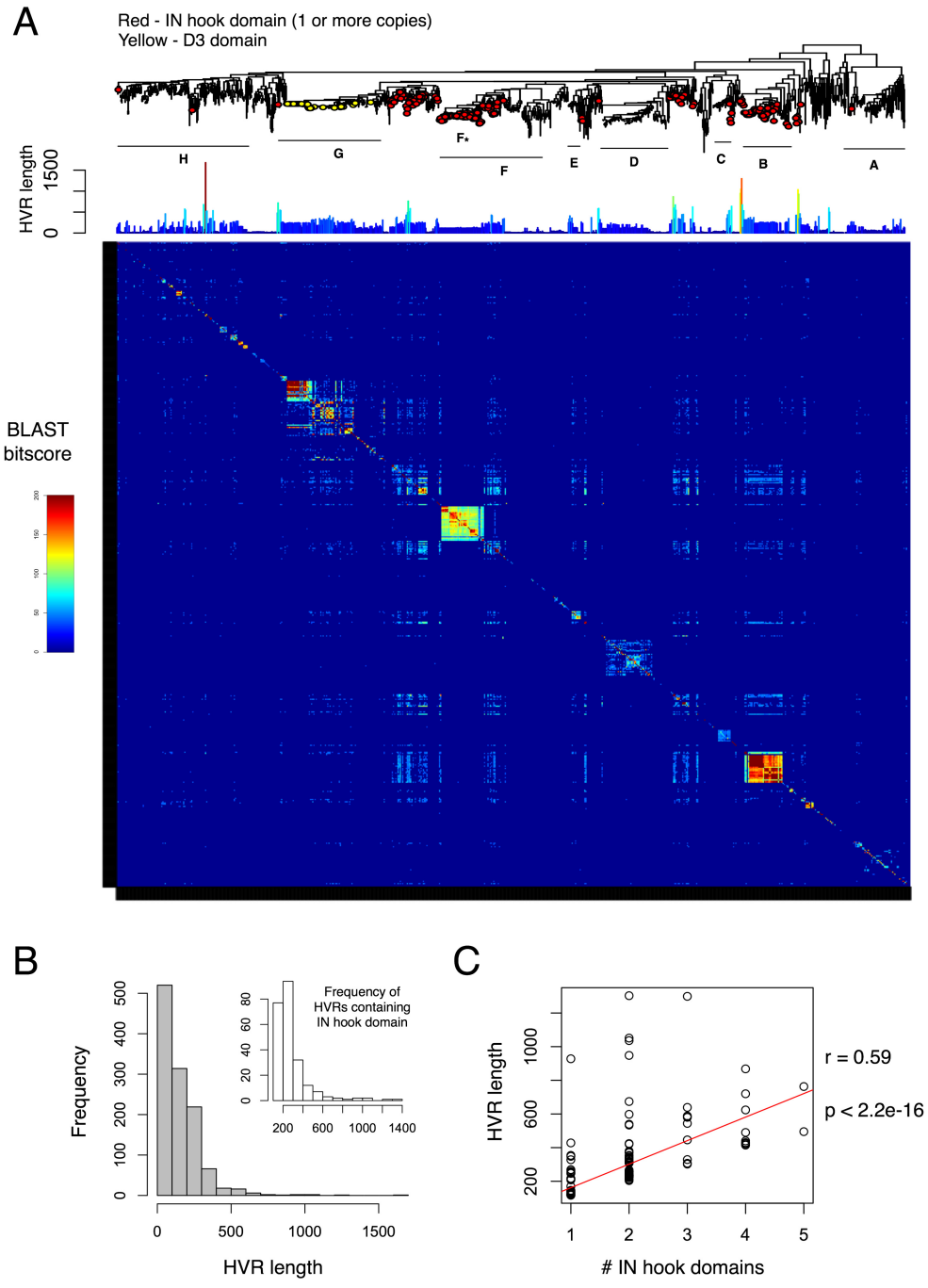
To address this question, all-by-all BLAST similarities were computed between each individual flagellin HVR, and pairwise HVR sequence similarities were visualized in the form of a heatmap, colored by BLAST bitscore (Figure 5.7A). Above the heat map is a plot of HVR sequence length and the flagellin N-terminal domain tree, on which both plots have been mapped. The distribution of HVR lengths is surprisingly variable (Figure 5.7B), ranging from 0 to 1,681 residues. In the heat map, numerous "blocks" of sequence similarity are apparent along the diagonal, which represent related groups of flagellins whose HVRs are the most similar to other members within the same clade. These represent monophyletic clusters of highly similar HVRs, and thus have arisen in a particular ancestral flagellin gene and been largely retained in its descendants. Several monophyletic "blocks" are notable, including subgroups of clade G (Enterobacteriales), clade F* (Vibrionaceae), clade D in

which the similarities appear more distant but are nonetheless monophyletic, and clade B (Campylobacterales) whose HVR is extremely well conserved.

In addition to these monophyletic patterns, an interesting, somewhat scattered pattern of more distant HVR-vs-HVR similarities is evident from the heat map. This similarity pattern appears as a series of segmented horizontal and vertical blocks (colored light blue due to weaker BLAST bitscores). Not surprisingly, this segmented pattern corresponds quite well to the pattern of HVR lengths that are indicated above the heat map. The flagellin HVR has been deleted or largely reduced in multiple lineages, and this coincides with loss of sequence similarity between all other HVRs, which appears as long horizontal and vertical stretches of dark blue. This pattern of HVR loss/reduction appears to have occurred frequently and independently throughout flagellin evolution, and suggests the existence of a selective pressure towards removal of the flagellin HVR in certain lineages.

To determine whether the recurring similarity pattern and uneven distribution of HVR lengths correlates with the presence of certain domains in the flagellin HVR, the D3 domain and flagellin_IN domain were searched for in each HVR using their respective HMMs, and mapped onto the tree. The D3 domain maps specifically to cluster G (Enterobacteriales), and thus does not by itself explain the length distribution or similarity pattern. However, the flagellin_IN domain is distributed throughout the phylogeny, and coincides well with the segmented pattern of HVR-vs-HVR similarity illustrated in the heat map. Furthermore, the HVR lengths, which are quite varied even when restricted only to HVRs possessing a flagellin_IN domain (Figure 5.7B), correlates significantly with the number of detected flagellin_IN domains in each HVR [pearson correlation coefficient (r) = 0.59, p < 2.2e-16] (Figure 5.7C). Thus, duplication and sequence divergence of the flagellin_IN domain appears to be a dominant mechanism underlying the evolution of a large proportion of flagellin HVRs. Deletion of the flagellin HVR appears to be a second common mechanism.

Figure 5.7 *(following page)*: **Mapping of HVR-vs-HVR similarities onto the flagellin N-terminal domain phylogeny.** (A) A heat map of HVR-vs-HVR bitscores resulting from all-vs-all BLAST alignments. The N-terminal domain phylogeny and HVR lengths have been mapped onto the heat map (above). Presence/absence of the flagellin IN hook domain and D3 domain as determined using a profile HMM search is also indicated by red and yellow circles on the phylogeny, respectively. (B) The HVR length distributions for all flagellins and IN hook-containing flagellins (inset). (C) A significant, positive relationship between HVR length and number of detected IN hook domains.


A
Red - IN hook domain (1 or more copies)
Yellow - D3 domain
H
G
F*
F
E
D
C
B
A
HVR length
0
1500
BLAST
bitscore
B
Frequency
HVR length
Frequency of
HVRs containing
IN hook domain
C
HVR length
# IN hook domains
r = 0.59
p < 2.2e-16

Phylogenetic incongruence between HVR types and the flagellin N-terminal domain

If HVRs have evolved congruently with the adjacent N-terminal region, having not undergone any recombination, then clusters of similar HVRs should map onto the flagellin N-terminal domain tree in a monophyletic pattern. Deviations from this may indicate more unusual evolutionary scenarios such as homoplasy (e.g., convergent evolution of similar HVR types in different flagellins), or recombination of the HVR between flagellins. The collagenase-like HVR (examined previously in section 5.1) was hypothesized to be one such example of an HVR with a scattered phylogenetic distribution.

To investigate this, a simple clustering-based approach was used to group HVRs into related "types", and these types were mapped onto the phylogenetic tree. For clustering, any two HVRs possessing significant sequence similarity (BLAST E-value $<$ 1e-10) were grouped together, which resulted in 235 singleton clusters (HVRs too small or divergent to cluster with other HVRs), 62 clusters of size two to four, and 32 larger clusters containing four or more HVRs (largest cluster contained 61 sequences). Cluster membership (NCBI IDs) and cluster-tree mapping information is included in Appendix 6. These 32 clusters containing four or more HVRs were analyzed in greater detail by mapping them onto the N-terminal phylogenetic tree as shown in Figure 5.8. This allowed us to visualize the phylogenetic concordance (or discordance) between the cluster assignments (HVR "types") and the phylogenetic tree of the N-terminal region.

To compute a quantitative measure of the degree of phylogenetic congruence between the HVR types and the tree, the following approach was used. The cluster assignments (HVR types) were reconstructed as discrete states by maximum likelihood on the phylogenetic tree. In the reconstruction, HVR type was considered a character with 32 distinct states (one for each cluster ID), and sequences with HVRs not within these categories were assigned an unknown (i.e., "?") state. An equal state change model (equal rates of change between character states) was assumed. For each HVR type, the most recent common ancestor (MRCA) node of all sequences possessing that HVR type was then assessed for the probability that it also shares the same HVR type. If this probability is high, the HVR type was likely present in the MRCA, is thus likely to be monophyletic. HVR types for which this value (denoted p_{mrca} in Figure 5.8) is low, is indicative of homoplasy / recombination.

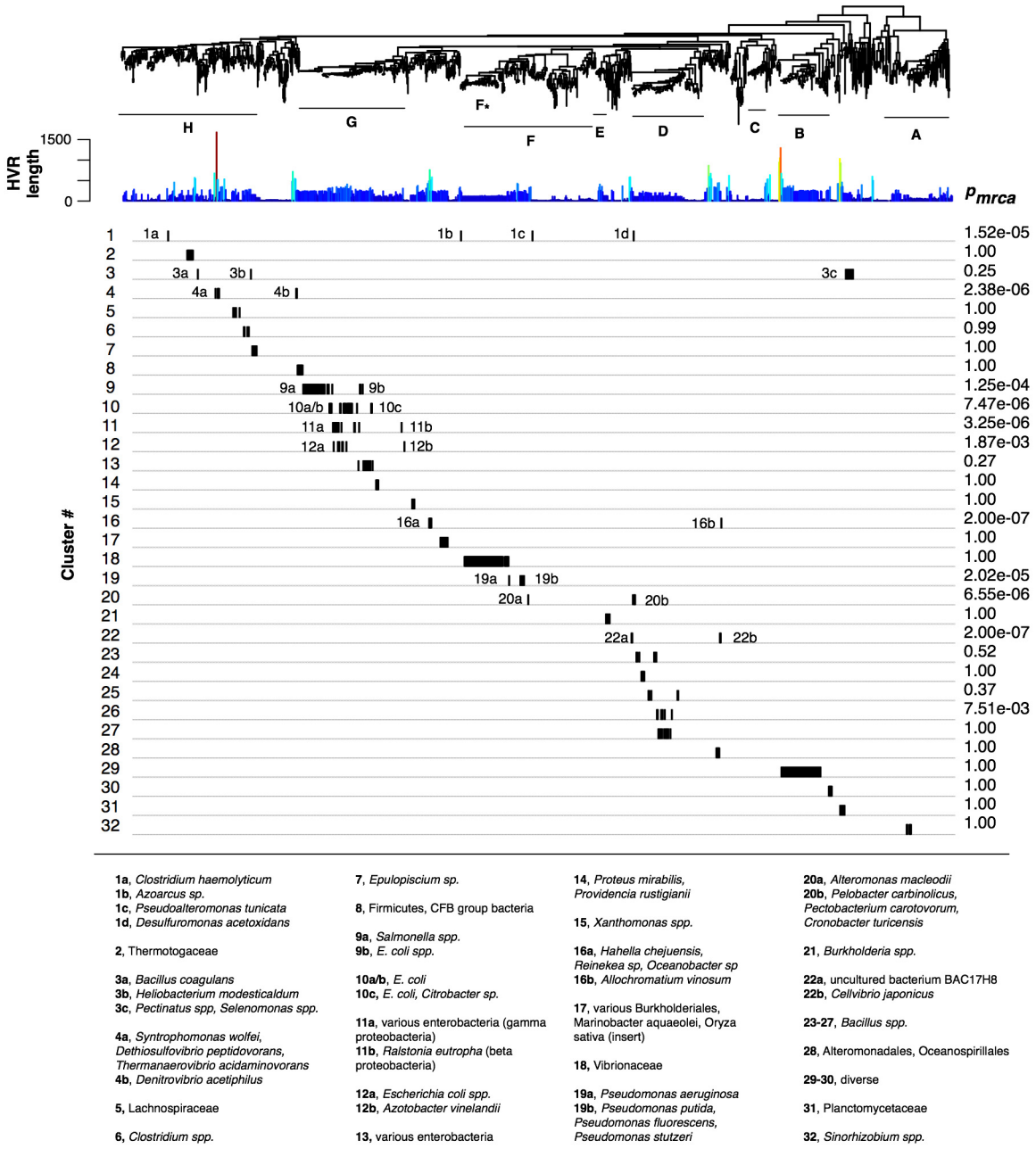


Figure 5.8: **Mapping of flagellin HVR types on a phylogenetic tree of the N-terminal domain.**

As shown in Figure 5.8, there are many cases of large-scale phylogenetic incongruence between HVR cluster types and the N-terminal domain phylogenetic tree, which suggests intragenic recombination of the flagellin HVR or convergent/parallel evolution of similar HVR types in different flagellin lineages. Consistent with the earlier hypothesis, the collagenase-like HVR (cluster 1) was found in four phylogenetically distant flagellin lineages. The probability that the MRCA of these sequences also contained this HVR type was found to be extremely low (p_{mrca} = 1.52e-05). Ultimately, the N-terminal domains of these four flagellins are distantly related according to the tree, while their HVRs are actually most closely related to each other out of all the HVRs. This interesting example is extremely unlikely to have occurred by simple mutational drift of the flagellin HVR.

In addition to this example, there are 10 other HVR types with p_{mrca} values < 0.01 (8 of these are < 0.001), suggesting that ancestral intragenic recombination events are widespread in flagellin evolution. The scattered distributions of these cases are shown in Figure 5.8. For instance, a shared HVR type was found in two groups of distantly related flagellins in cluster 16, labelled "16a" and "16b", respectively.

Interestingly, four of these apparently recombination-prone HVR-types are associated with the Enterobacteria-specific clade G (Figure 5.8). One identified example, cluster 9, suggests a recombination and lateral gene transfer event took place between a group of *Salmonella* and *E. coli* flagellins (Figure 5.8). A previous study of these particular *E. coli* flagellins (Tominaga, 2004) also suggested that they have been laterally transferred from *Salmonella*. Cluster 12 also has a scattered phylogenetic distribution, occurring in flagellins from *E. coli* and *Azotobacter vinelandii*. In addition, an unusual *Ralstonia eutropha* flagellin (11b) from cluster 11 possesses an HVR type found in several Enterobacteriaceae species (11a). Identification of flagellin recombination events within the Enterobacteriaceae is consistent with several previous analyses. Flagellin recombination has been demonstrated in *Salmonella* (Frankel et al., 1989; Smith et al., 1990) and *E. coli* (Reid et al., 1999; Wang et al., 2003).

Flagellin mosaics resulting from HVR recombination

Scattered leaves of the N-terminal domain tree that possess a similar HVR type represent putative *mosaics*: genes that have resulted from intragenic recombination of HVR sequences from different lineages and possibly, lateral gene transfer (Smith, 1992). By

aligning the full length sequences corresponding to these clusters, clear mosaic patterns were identified. Sequence alignments demonstrating five examples of mosaic flagellin proteins are shown in Figure 5.9. In each alignment, a flagellin mosaic (bolded and underlined, second sequence in alignment) has been aligned to a sequence (A) containing a very similar N- and C-terminal domain (first sequence in alignment) but a dissimilar HVR. It has also been aligned to one or more sequences sharing the same HVR type but possessing less similarity with the N- and C-terminal domain (these sequences are in the rows below the mosaic sequence).

The collagenase-like HVRs associated with cluster 1 will be described in greater detail as this group was part of the motivation for this study, and also has interesting functional characteristics. As shown in the upper left alignment of Figure 5.9, the *C. haemolyticum* flagellin N-terminal and C-terminal regions are most similar to those in a *C. botulinum* flagellin sequence. The *C. haemolyticum* flagellin HVR region, however, is most similar to HVRs found in three other flagellin sequences that are more distantly related to *C. haemolyticum* flagellin according to their N- and C-terminal domain sequences. Ultimately, the phylogenetic and sequence analysis suggests that the collagenase-like HVR originated independently in these four flagellin sequences, either by separate insertions of a similar collagenase-like sequence, or through intragenic recombination. It is also worth noting that sequence similarities between collagenase-like HVRs in these four flagellins are greater than their similarities to microbial collagenases (data not shown), which suggests that the flagellin collagenase-like HVRs form a related subfamily of collagenase-like sequences.

Regarding the function of the collagenase-like HVRs, it is possible that they may be involved in defense against bacteriophages. Not only do some phages contain collagen-like sequences in their tail spike proteins, but some are known invade bacteria through flagella (Samuel et al., 1999). Thus, it is possible that collagenase-like HVRs cleave phage tail-spike proteins (Huiskonen et al., 2007), and form a first line of defense against invading phages.

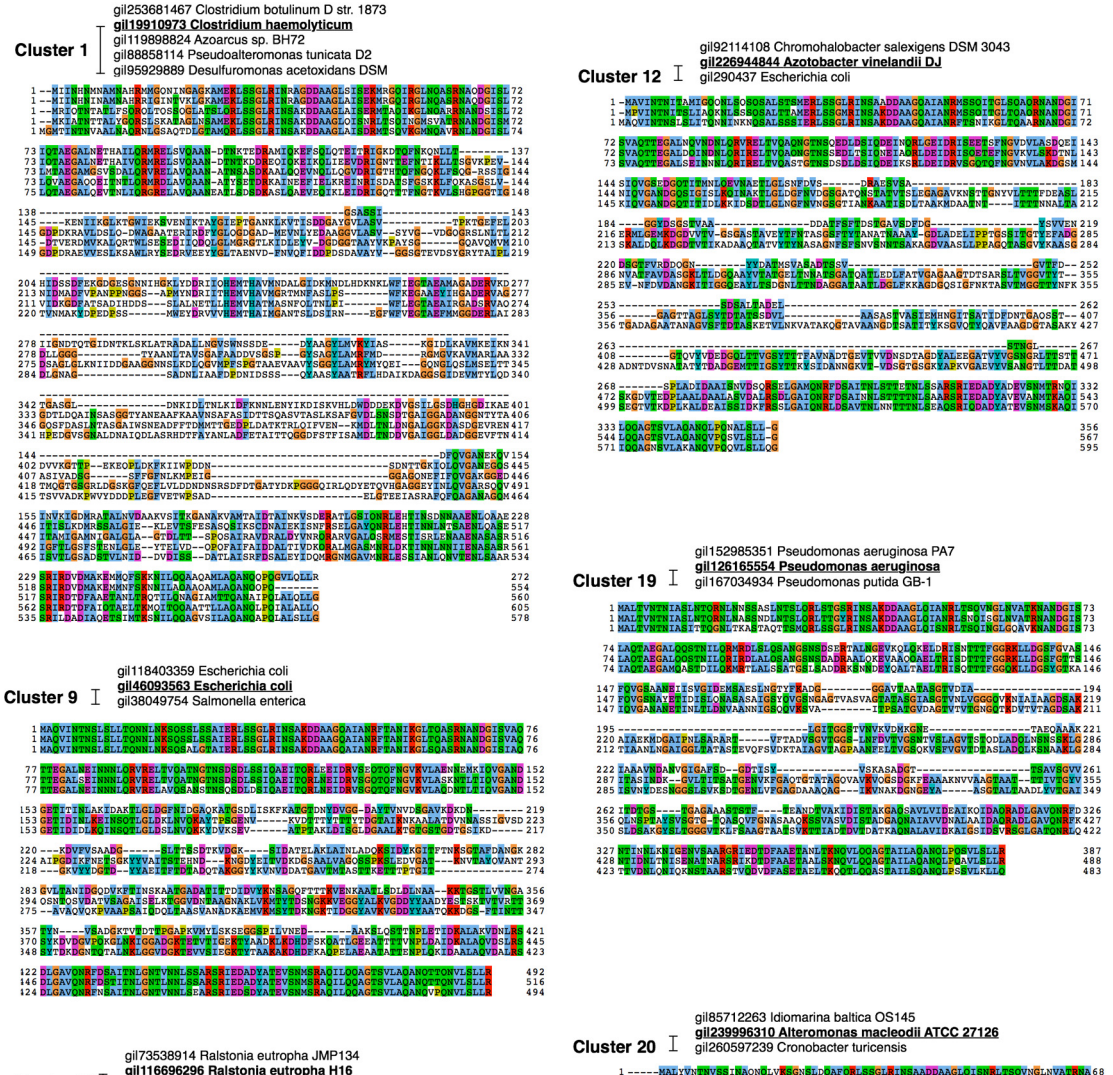


Figure 5.9: **Sequence alignments indicating flagellin mosaic structures resulting from putative recombination events.**

5.2.4 Conclusion

A number of mechanisms appear to be at work in the ongoing evolution of flagellin HVRs. Most notably, there is evidence of widespread intragenic recombination and lateral transfer of HVR types between flagellins. While evolutionary patterns within clusters have not been assessed, large-scale patterns resulting from potentially ancestral recombination events have been revealed. Due to extreme sequence divergence, these more distant patterns are difficult to examine at the nucleotide level, but can be more easily studied at the protein sequence level.

There is also evidence of a strong selective pressure in numerous lineages towards loss/reduction of the HVR. It is possible that flagellins without HVRs play more of a structural role in the assembly of the flagellar filament, and are used predominantly for motility, while flagellins containing longer, more specialized HVRs play additional roles such as host-cell binding by pathogenic bacteria and biofilm formation. It makes sense that a flagellin HVR will likely be lost over time unless constrained by a function (e.g., adhesin-like binding). HVR-mediated binding to host cells and other extracellular substrates is one such function that has been documented for a number of flagellins.

Lastly, a recurring evolutionary mechanism for building flagellin HVRs appears to be duplication of the flagellin_IN hook domain. This is perhaps the most widely conserved domain found in the flagellin HVR, and undoubtedly plays a critical role that is currently unclear. Future studies that reveal the role of the flagellin_IN domain will provide important insights into the function and evolution of bacterial flagellins.

Ultimately, mosaic sequences such as bacterial flagellin represent challenges to standard bioinformatics methods that rely on overall sequence similarity to infer evolutionary relationships or function. By dividing these sequences into smaller units of function and evolutionary history, and analyzing their phylogenetic distributions, it is possible to more accurately elucidate patterns of evolutionary change and function.

Chapter 6

Mimicry, pathogens, and virulence factors

A recurring idea of this thesis is that sequence similarity alone provides an inadequate representation of functional similarity. For instance, two sequences may be closely related, where one has antifreeze properties and the other does not. Or, two flagellins may be related, but have different functions through mutations in their HVRs (e.g., an HVR-less flagellin functioning only in motility versus a flagellin with an adhesin-like HVR that functions in host-cell binding).

In this chapter, I extend this idea beyond individual proteins and protein families, to entire genomes. The idea I explore in this chapter is a recurring theme in this thesis, that of molecular mimicry. The surfaces of antifreeze proteins structurally mimic the repetitive surface of ice; the aromatic side-chains of carbohydrate-binding proteins mimic the sugar rings in glycan chains; and neurotoxin domains appear to be mimicking host proteins. Thus, mimicry seems to be a powerful recurring feature, beyond sequence similarity, with the capability of distinguishing function.

Just like related proteins may differ in terms of their antifreeze activity based on the degree to which they mimic ice, related bacteria may differ in terms of the degree to which they encode mimics of host proteins. Subtle differences in the proteomes of related bacteria may better distinguish their virulence than a simple measure of sequence (or *genome*) similarity/relatedness. Moreover, virulence factors as a group cannot be unified as a sequence family. Host-mimicry patterns and the abundance of these patterns in

pathogenic versus non-pathogenic genomes, a genomic context like approach, may be a better predictor of novel virulence factors. In this chapter, I test a simple idea — the ability of overrepresented host-protein similarities to distinguish virulence factors and mechanisms in bacterial pathogens of humans.

6.1 Introduction

Numerous studies have shown that pathogens encode virulence factors that functionally or structurally mimic host proteins (Stebbins and Galan, 2001). Through mimicry, virulence factors can facilitate evasion of host immune detection, trigger host autoimmune disease (Albert and Inman, 1999), or interact and interfere with host proteins (Finlay and Cossart, 1997).

A well known example of host-mimicry by bacterial virulence factors is the *Yersinia* YopH virulence factor, which mimics human tyrosine phosphatases. YopH mimicry of tyrosine phosphatases reflects its role in dephosphorylation of human p130cas and the focal adhesion kinase (FAK), which allows it to disrupt macrophage defense (Stebbins and Galan, 2001). Another well established example of mimicry is the the *Legionella* virulence factor, RalF, which mimics human guanine-exchange factors (GEFs) and recruits AFP-ribosylation factor (Arf) to inhibit phagosome-lysosome fusion and form a replication vacuole (Nagai et al., 2002). Section 5.1 on clostridial neurotoxins includes another potential example of host-mimicry by a virulence factor. That is, the clostridial neurotoxin's HCRn domain shares similarities with the mammalian collagen NC4 domain, which, given the numerous functional links to collagen-binding and degradation for other genes in the gene cluster, may also reflect an existing or ancestral role in collagen mimicry. Bacterial proteins capable of catalyzing enzymatic reactions native to the host species is an additional example of host-mimicry. The pathogen *Helicobacter pylori* encodes an enzyme similar to human fucosyltransferases capable of producing the lewis X trisaccharide which mimics human sugars and allows the pathogen to escape immune recognition (Sun et al., 2007). For a thorough review on mimicry by bacterial virulence factors see Stebbins and Galan (2001).

While most studies of this type have been done on a case-by-case basis, it is possible that new virulences factors involved in host protein mimicry/modulation may be detectable

on a genomic scale by comparison of protein sequence similarities between human, bacterial pathogens and non-pathogens. In otherwords, rather than attempting to classify virulence factors based on similarities between each other, it may be possible to unify many existing virulence factors and predict novel cases by examining their similarities to host-proteins. These similarities may result from unique coevolutionary pressures that exist in human pathogens that are absent in their close relatives. Thus, in this chapter, a large-scale comparative analysis of host-protein similarities is performed for 154 non-redundant, microbial genomes, including 80 pathogenic and 74 non-pathogenic species. Human and bacterial proteins associated with enriched mimicry patterns in pathogenic versus non-pathogenic bacteria are analyzed for their potential involvement in virulence-related functions. The results suggest that the approach effectively predicts virulence factors with many structures and functions, and can be used to infer novel virulence mechanisms.

6.2 Methods

Protein sequence datasets for human as well as 187 bacterial genomes were retrieved from the NCBI [RefSeq human protein database build 36 (37,742 proteins)] and the Comprehensive Microbial Resource at TIGR/JVCI (http://cmr.jcvi.org). The dataset was constructed by retrieving all genomes with manually curated annotations regarding pathogenicity (human pathogen or non-pathogen) using the Genome Properties system of the Comprehensive Microbial Resource (Peterson et al., 2001). The crux of the analysis is that microbial genomes are scanned for unique similarities to host (human) proteins rather than similarities to known virulence factors. This approach has several potential advantages: identification of potential host targets, added functional information regarding host proteins, and ability to identify novel virulence factors that are not similar to existing, known virulence factors.

An all-by-all BLAST analysis was conducted using BLASTp v. 2.2.16, in which each human protein was used as an individual query in a separate BLAST search of each individual organism's protein database. Default BLAST parameters were used with composition-based statistics to correct for potential compositional bias. A BLAST E-value cutoff of 1e-06 was used to identify putative matches, from which a presence/absence matrix was constructed. To remove genome/species redundancy, one of any two bacterial genomes of

the same species were removed if the Pearson correlation coefficient between their presence/absence profiles was greater than 0.9. Nine genomes were removed based on the similarity matrix to human proteins, while eleven (an additional two) genomes were removed for analysis based on plant extracellular matrix (ECM) protein similarity data. An enrichment factor (EF) (described in text) was then calculated for each human protein, as described below.

For each human protein, the top match in each bacterial genome was retrieved via BLASTp, and an E-value threshold of 1e-06 was used to define a "hit". An enrichment factor (EF) was then calculated for each human protein by comparing the number of detected hits in pathogen versus non-pathogen genomes, using equation 6.1.

Equation 6.1:

$$EF_k = \log \frac{(H_p+1)/N_p}{(H_{np}+1)/N_{np}}$$

H_p, H_{np} — # pathogens and non-pathogens containing a putative mimic of protein k

N_p, N_{np} — # pathogen and non-pathogen genomes in dataset

Following calculation of enrichment factors, the top 1% of human proteins according to EF were selected as well as the top matches in bacterial genomes. A list of NCBI GI accession numbers for the top 1% group of human proteins is included in Appendix 7. The top pathogen protein matches were identified by compiling a dataset containing all pathogen proteins (236,304 total proteins), and performing BLAST searches with each human protein as a query.

To analyze the extent that the pathogen proteins associated with the top 1% are potential virulence factors, their similarities to the VFDB database were compared with similarities of random pathogen proteins to the VFDB. Two analyses were performed: one with 934 unique pathogen proteins associated with the top 1% group; and a second with 138 unique pathogen proteins that were top matches to human proteins in the top 1% group. Hits were again defined as BLAST matches with E $<$ 1e-06. The background frequency of matches between pathogen proteins and the VFDB was 36,163 out of 236,304 total pathogen proteins (0.153%). A binomial exact test was performed to estimate the significance of the observed result.

For analysis of overrepresented go terms, the top 1% of human proteins were analyzed in three ways, with no filtering, filtering to proteins possessing unique matches to pathogen

proteins, and filtering to proteins possessing unique top reciprocal matches to pathogen proteins. With the no-filtering option, multiple matches between a single pathogen protein and numerous human proteins is permitted. As this does not reduce the bias that may result from the occurrence of similar sequences in the same GO category, the filtered analyses were performed to account for this. For analysis of overrepresented GO terms, Funcassociate (Berriz et al., 2003) was used with the following parameters (Simulations = 10000; P-value cutoff = variable (0.05 to 0.2); Mode = unordered; Overrepresented only; Evidence codes: all).

The R statistical programming package (v. 2.8.1) was used to analyze and visualize the presence/absence profiles for the top 1% of human proteins (according to EF). The hclust function was used for hierarchical clustering of presence/absence profiles using the following default parameters: cluster method = "complete"; distance-metric = Euclidean. To organize the presence/absence profiles into a more manageable set for manual analysis, the hierarchical clustering tree was divided into 80 subtrees. Most (but not all) of the subtrees consist of related human proteins (e.g., paralogs).

6.3 Results

The EFs for all human proteins are normally distributed and centered around a mean of -0.06. More detectable similarities to human proteins were actually detected in non-pathogens, which may reflect a general trend towards genome reduction in intracellular parasitic pathogens (Moran, 2002). However, the human and pathogen proteins associated with the highest EFs (top 1%) were extracted and analyzed in further detail. In this subset ("top 1% group"), there is evidence of significantly elevated virulence-related functions as described in the following sections.

To quantitatively determine whether the approach detected virulence factors better than can be expected by chance, the number of matches between the pathogen proteins associated with the top 1% group and known virulence factors in the Virulence Factors of Pathogenic Bacteria Database (VFDB) (Chen et al., 2005) were counted. Out of 934 unique pathogen proteins detected, 210 had hits in the VFDB (2301 sequences) with E $<$ 1e-06. This was significantly greater (p = 4.5e-09, binomial test) than that expected by randomly selecting proteins from the genomes of known pathogens. When limiting the

queries to only include unique top matches, of which there were 138, 39 of these had hits in the VFDB (p = 7.5e-05, binomial test). Thus, the top 1% is significantly enriched in proteins involved in bacterial pathogenesis.

A statistical analysis of overrepresented gene ontology (GO) terms (Berriz et al., 2003) identified several functions of human proteins in the top 1% group that were significantly enriched. These included extracellular matrix constituent (e.g., collagen), lipid and fatty acid metabolism, and interferon regulation (Table 6.1, Appendix 8). The enriched GO terms reflect well-established mechanisms of bacterial pathogenesis. For instance, binding of bacterial adhesins to host ECM proteins, pathogen recognition via immune receptors, and modification of host lipids are established virulence mechanisms of many bacterial pathogens (Finlay and Cossart, 1997). When the analysis was further restricted to highly unique human-bacteria protein similarities (i.e., top reciprocal matches), overrepresented functions included "peptidoglycan receptor activity" (p = 0.002) and "ceramide kinase activity" (p = 0.026), which have key roles in recognition and internalization of pathogens by mammalian cells (Liu et al., 2001; Gulbins et al., 2004).

Table 6.1: **A reduced list of overrepresented biological functions in human proteins associated with the top 1% of enrichment factors. For the full list see Appendix 8.**

N[a]	X[b]	LOD	P	Padj	GO ID	Description
4	4	2.72	7.94E-08	0	GO:0004102	choline & O-acetyltransferase activity
4	4	2.72	7.94E-08	0	GO:0032648	regulation of interferon-beta production
26	40	2.06	5.98E-37	0	GO:0005581	collagen
11	19	1.91	1.80E-15	0	GO:0032012	regulation of ARF protein signal transduction
4	7	1.88	2.67E-06	0.003	GO:0032647	regulation of interferon-alpha production
5	10	1.77	3.11E-07	0.001	GO:0032365	intracellular lipid transport
4	9	1.68	9.36E-06	0.024	GO:0016045	detection of bacterium
6	15	1.61	9.69E-08	0	GO:0008329	pattern recognition receptor activity
16	51	1.45	1.28E-16	0	GO:0008374	O-acyltransferase activity
28	132	1.23	5.74E-23	0	GO:0044420	extracellular matrix part
8	67	0.93	1.65E-05	0.036	GO:0030258	lipid modification
13	138	0.81	6.34E-07	0.001	GO:0045087	innate immune response

[a] Number of query proteins in the respective GO category.

[b] Total number of human proteins with the respective GO annotation.

Given the apparent enrichment in virulence mechanisms in the top 1% group, the next goal was to analyze this subset of human-bacteria protein similarities for known

and potentially novel virulence mechanisms. To organize the detected protein similarity relationships, the detected similarities were clustered and visualized in the form of a presence/absence-map (Figure 6.1). The presence/absence map graphically illustrates the detected matches between human proteins (x-axis) and protein in pathogen genomes (y-axis). Through analysis of the functions of the human and bacterial protein within these clusters, nineteen relationships were identified that represent known or potentially novel virulence mechanisms (Table 6.2). These include a considerable number of bacterial pathogen proteins/protein-families whose biological functions are currently unknown.

Figure 6.1 *(following page)*: **Presence/absence map of human versus bacterial-pathogen protein similarities.** The presence/absence of detected similarities between human proteins (x-axis) and proteins from bacterial pathogens (y-axis) is indicated in the 2D grid. The human proteins include the top 1% according to enrichment factor (enrichment of detected similarities in pathogens compared to non-pathogens). For each of these human proteins, the fraction of pathogen versus non-pathogen genomes containing a similar protein is shown below the grid. Human proteins were organized by hierarchical clustering of the presence/absence patterns in each column, and further annotated manually. Bacterial pathogens are ordered roughly according to a broad phyletic classification. Pathogens marked by asterisks were not included in the initial calculation due to redundancy (see Methods), but are included for completeness. See Appendix 7 for human protein NCBI IDs.

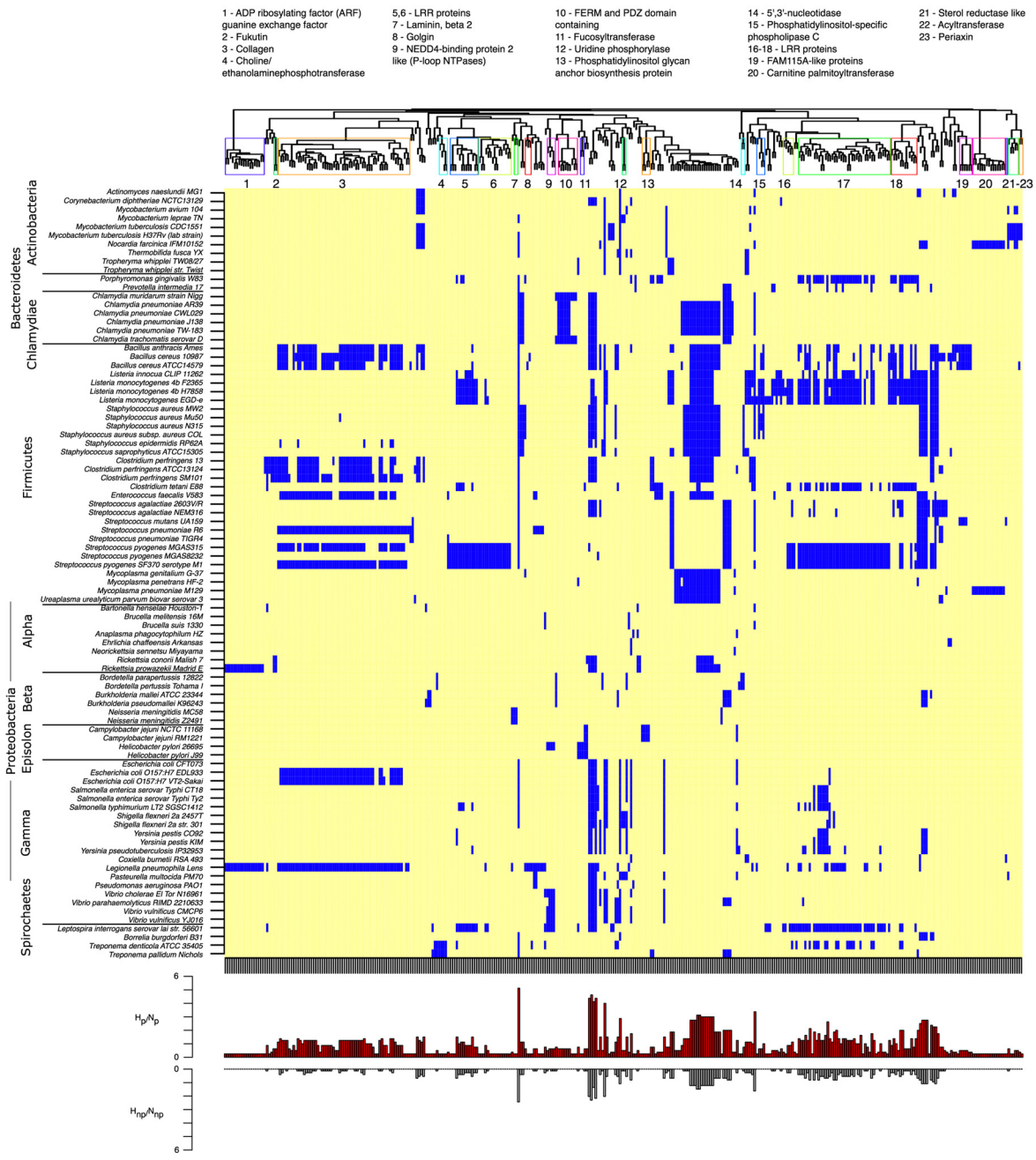

1 - ADP ribosylating factor (ARF) guanine exchange factor
2 - Fukutin
3 - Collagen
4 - Choline/ ethanolaminephosphotransferase
5,6 - LRR proteins
7 - Laminin, beta 2
8 - Golgin
9 - NEDD4-binding protein 2 like (P-loop NTPases)
10 - FERM and PDZ domain containing
11 - Fucosyltransferase
12 - Uridine phosphorylase
13 - Phosphatidylinositol glycan anchor biosynthesis protein
14 - 5',3'-nucleotidase
15 - Phosphatidylinositol-specific phospholipase C
16-18 - LRR proteins
19 - FAM115A-like proteins
20 - Carnitine palmitoyltransferase
21 - Sterol reductase like
22 - Acyltransferase
23 - Periaxin
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21-23
Bacteroidetes
Actinobacteria
Actinomyces naeslundii MG1
Corynebacterium diphtheriae NCTC13129
Mycobacterium avium 104
Mycobacterium leprae TN
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis H37Rv (lab strain)
Nocardia farcinica IFM10152
Thermobifida fusca YX
Tropheryma whipplei TW08/27
Tropheryma whipplei str. Twist
Porphyromonas gingivalis W83
Prevotella intermedia 17
Chlamydiae
Chlamydia muridarum strain Nigg
Chlamydia pneumoniae AR39
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae J138
Chlamydia pneumoniae TW-183
Chlamydia trachomatis serovar D
Firmicutes
Bacillus anthracis Ames
Bacillus cereus 10987
Bacillus cereus ATCC14579
Listeria innocua CLIP 11262
Listeria monocytogenes 4b F2365
Listeria monocytogenes 4b H7858
Listeria monocytogenes EGD-e
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Staphylococcus aureus subsp. aureus COL
Staphylococcus epidermidis RP62A
Staphylococcus saprophyticus ATCC15305
Clostridium perfringens 13
Clostridium perfringens ATCC13124
Clostridium perfringens SM101
Clostridium tetani E88
Enterococcus faecalis V583
Streptococcus agalactiae 2603V/R
Streptococcus agalactiae NEM316
Streptococcus mutans UA159
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus pyogenes MGAS315
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes SF370 serotype M1
Mycoplasma genitalium G-37
Mycoplasma penetrans HF-2
Mycoplasma pneumoniae M129
Ureaplasma urealyticum parvum biovar serovar 3
Proteobacteria
Alpha
Bartonella henselae Houston-1
Brucella melitensis 16M
Brucella suis 1330
Anaplasma phagocytophilum HZ
Ehrlichia chaffeensis Arkansas
Neorickettsia sennetsu Miyayama
Rickettsia conorii Malish 7
Rickettsia prowazekii Madrid E
Beta
Bordetella parapertussis 12822
Bordetella pertussis Tohama I
Burkholderia mallei ATCC 23344
Burkholderia pseudomallei K96243
Neisseria meningitidis MC58
Neisseria meningitidis Z2491
Epsilon
Campylobacter jejuni NCTC 11168
Campylobacter jejuni RM1221
Helicobacter pylori 26695
Helicobacter pylori J99
Gamma
Escherichia coli CFT073
Escherichia coli O157:H7 EDL933
Escherichia coli O157:H7 VT2-Sakai
Salmonella enterica serovar Typhi CT18
Salmonella enterica serovar Typhi Ty2
Salmonella typhimurium LT2 SGSC1412
Shigella flexneri 2a 2457T
Shigella flexneri 2a str. 301
Yersinia pestis CO92
Yersinia pestis KIM
Yersinia pseudotuberculosis IP32953
Coxiella burnetii RSA 493
Legionella pneumophila Lens
Pasteurella multocida PM70
Pseudomonas aeruginosa PAO1
Vibrio cholerae El Tor N16961
Vibrio parahaemolyticus RIMD 2210633
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Spirochaetes
Leptospira interrogans serovar lai str. 56601
Borrelia burgdorferi B31
Treponema denticola ATCC 35405
Treponema pallidum Nichols
6
0
H_p/N_p
0
H_{np}/N_{np}
6

Table 6.2: Nineteen putative virulence mechanisms identified via comparative analysis of host-protein similarities.

Cluster	Human Protein(s)	Pathogen Protein	Species	VF	Mechanism
1	ADP ribosylating factor (ARF) guanine exchange factor	lpl1919 (RalF)	*Legionella pneumophila*, *Rickettsia prowazekii*	Y	RalF recruits ARF GTPases for *Legionella* phagosome formation (Nagai et al., 2002)
2	Fukutin	CPF_0603 (LicD family)	*Clostridium perfringens*	?	Interaction with host alpha-dystroglycan; destabilization of host ECM-cytoskeletal network
3	Collagen	spr1403	Many	Y	Involvement in host-cell adherence/ invasion (Lukomski et al., 2000; Paterson et al., 2008)
4	Choline / ethanolamine phosphotransferase	TDE_0021	*Treponema* spp.	?	Production of phosphatidylcholine; host-phospholipid mimicry (Sohlenkamp et al., 2003; Kent et al., 2004)
5, 6, 16-18	Leucine-rich repeat proteins	spyM18_1868	Many	Y	Internalin-related; adhesion and invasion of host epithelia via via E-cadherin interaction (Gaillard et al., 1991)
7	Laminin, β 2	NMA0386	*Neisseria meningitidis*	?	Adhesion/invasion of host ECM (Eberhard et al., 1998; Orihuela et al., 2009)
8	Golgin	lpl2411 (lepB)	*Legionella pneumophila*	Y	Disruption of vesicular protein-trafficking (Ingmundson et al., 2007)
9	NEDD4-binding protein 2 like (P-loop NTPases)	VC_1610	*Vibrio* spp. *Helicobacter pylori*	?	Modulation of host external ATP pool and thus macrophage cell death (Zaborina et al., 1999; Punj et al., 2000; Trautmann, 2009)
10	FERM and PDZ domain containing	TC_0725 (Tail-specific proteases)	*Chlamydia* spp.	Y	Degradation of host immune response transcription factors (Lad et al., 2007)
11	Fucosyltransferase	HP0379	*Helicobacter pylori*	Y	Production of lewis x trisaccharide; mimicry of host cell-surface sugars; immune evasion (Sun et al., 2007)
12	Uridine phosphorylase	ML2177	*Mycobacterium leprae*	Y	Use of host uridine for pyrimidine scavenging (Wheeler, 1989)
13	Phosphatidylinositol glycan anchor biosynthesis protein	CJE_0829	*Campylobacter jejuni*	?	Production of GPIs; mimicy of host GPIs or modulation of macrophage function (van der Meer-Janssen et al., 2009)
14	5',3'-nucleotidase	BP0294	*Bordatella pertussis*	?	Hydrolysis of extracellular nucleotides (ENs); disruption of EN-mediated immune signaling (la Sala et al., 2003; de Almeida

Table 6.2: Nineteen putative virulence mechanisms identified via comparative analysis of host-protein similarities.

Cluster	Human Protein(s)	Pathogen Protein	Species	VF	Mechanism
					Marques-da-Silva et al., 2008)
15	Phosphatidylinositol-specific phospholipase C	lmo0201	*Listeria monocytogenes*, *Staphylococcus aureus*	Y	Aids in phagosome lysis (Freitag et al., 2009)
19	FAM115A-like proteins	BCE_5203 (enhancin metalloprotease family)	*Bacillus anthracis*, *Bacillus cereus*	?	Host-protein (e.g., mucin proteolysis) (Ivanova et al., 2003)
20	Carnitine palmitoyl-transferase	nfa31870	*Nocardia farcinica*, *Mycoplasma pneumoniae*	?	Biosynthesis of human-like sphingolipid precursors; sphingolipid metabolism is altered following *M. pneumoniae* infection (Yu et al., 2009)
21	Sterol reductase like	CBU_1158	*Coxiella burnetii*	?	Cholesterol metabolism involved in production of cholesterol-rich parasitophorous vacuole; perturbation of host cell cholesterol metabolism (Howe, 2006)
22	Acyltransferase	Rv1920	*Mycobacterium* spp.	?	Lipid biosynthesis from host-derived precursors; use as energy store and/or virulence-related, immunomodulatory lipids (Deb et al., 2009; Ehrt and Schnappinger, 2009)
23	Periaxin, apoB (leucine-proline rich repeats)	MT_1796 (PPE family domain)	*Mycobacterium* spp.	?	Targeting of DRP2-dystroglycan complex; interaction with host lipids? (Sherman et al., 2001; Marques et al., 2001); may also facilitate survival in macrophages

Several clusters (3, 5-7, 16-18) reflect potential virulence mechanisms involving ECM adhesion and invasion. A wide range of pathogens including species of *Clostridium*, *Bacillus*, *Listeria*, *E. coli* and *Legionella* contain elevated similarities to human collagen (cluster 3) and leucine-rich repeat (LRR) domains (clusters 5, 6, 16-18). The corresponding pathogen LRRs are homologs of the virulence factor, internalin, which mediates adhesion and invasion of the host ECM (Gaillard et al., 1991). Furthermore, the collagen-like proteins are similar to a known virulence factor from *Streptococcus pneumoniae* that mediates host-cell adherence and soft tissue pathology (Lukomski et al., 2000). Although the detected similarities are due to similar repetitive sequences, and not necessarily homology, these similarities are considerably weaker or absent in non-pathogens and are thus not an artifact of general, repetitive sequences in bacteria. In addition, while these proteins are predominantly in pathogens from the Firmicutes phylum (Figure 6.1), the pattern also appears to have been transferred to the highly pathogenic O157:H7 strain of *E. coli* as well as *Legionella pneumophila* but not other Gammaproteobacteria. In fact, while present in *E. coli* O157:H7, the collagen-like proteins are absent in the uropathogenic *E. coli* CFT073 strain, and the non-pathogenic K1 strain. Future work characterizing the role of these collagen-like proteins will be useful.

Similarities to the C-terminal domain of laminin β 2 (in cluster 7), another ECM component, were detected exclusively in *Neisseria meningitides*, the cause of meningococcal meningitis. This is a potentially important link given that laminin β 2 is highly neural specific and that the human laminin-receptor has been shown to initiate bacterial contact with the blood barrier in meningitis infection via an unknown mechanism (Orihuela et al., 2009). This *N. meningitides* protein (NMA0386) may play a role in the laminin-receptor interaction.

The dystroglycan complex, a transmembrane linkage between the ECM and cytoskeleton, is also an apparent target of several pathogen proteins detected in the analysis. Similarities to the human muscular dystrophy associated protein, fukutin (cluster 2), were detected exclusively in the pathogens, *Clostridium perfringens* and *Rickettsia* spp. Human fukutin is involved in interactions with alpha-dystroglycan, a key structural component found in muscle and other tissues, and a known receptor for major human viruses (Cao et al., 1998) and *Mycobacterium leprae* (Rambukkana et al., 1998). Alpha-dystroglycan may therefore be an unknown target of these bacterial pathogens mediated by bacterial fukutin-like proteins.

It is known that *M. leprae* (cause of leprosy) specifically invades human Schwann cells through an interaction with the dystroglycan complex. In addition, it is possible that *M. tuberculosis* may also invade human epithelial cells through an epithelial dystroglycan complex. Due to a common repetitive sequence and leucine-proline rich composition, BLAST detected similarity between a *M. tuberculosis* PPE protein (Rv1918c) and the human protein, periaxin (cluster 23), and to a lesser extent, apolipoprotein B. Interestingly, human periaxin is a Schwann-cell specific protein that is critical to formation of the dystroglycan-complex. While the detected similarity is not due to homology but rather a repetitive proline-rich composition (see Figure 6.2), it is tempting to speculate that PPE proteins may also interact with the dystroglycan-complex, or possibly act as membrane-interacting lipoproteins.

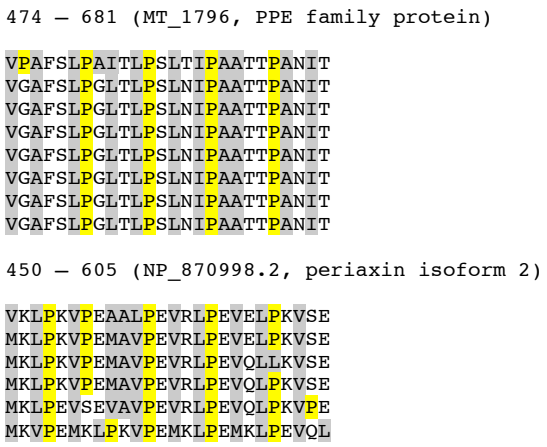


Figure 6.2: **Similar repeat patterns between human periaxin and a mycobacterial PPE protein.**

Another set of functions common to the proteins identified is that of lipid and sugar metabolism (Table 6.2). These cases likely reflect the unique abilities of bacterial pathogens to produce mimics of human lipids and sugars in order to modulate host processes and evade host immune recognition. They include human-like proteins involved in the production of phosphatidylcholine (*Treponema* spp.), cholesterol (*Coxiella burnetii*), and sugars (i.e., synthesis of the Lewis x trisaccharide via fucosyltransferase in *Helicobacter pylori*) (Table 6.2).

Homologs of human Carnitine O-palmitoyltransferase (CPT) were detected in *Mycobac-*

terium pneumonia and *Nocardia farcinica* (cluster 20), which are both causes of bacterial pneumonia. Lung infection by *M. pneumonia* has been linked to the pathogen's ability to mimic or modulate host sphingolipid biosynthesis (Yu et al., 2009). Human CPT converts Palmitoylcarnitine into Palmitoyl-CoA, which is a key precursor used in the production of sphingosine. The bacterial versions of CPT may therefore play a role in sphingolipid related virulence mechanisms of these pathogens.

A potential lipid-related virulence mechanism for *Campylobacter jejuni*, a major cause of human gastroenteritis, was also inferred. Previous studies suggest that *C. jejuni* invasion of host epithelial cells may involve lipid rafts or caveolae, but the details are unclear. The analysis revealed a unique *C. jejuni* protein (CJE_0829) with similarities to human phosphatidylinisitol glycan anchor (GPI-anchor) biosynthesis proteins (cluster 13). It is possible that this *C. jejuni* protein is involved in the production of human-like GPI-anchored proteins that allow the pathogen to enter human lipid-rafts, where GPI-anchored proteins are preferentially located.

Another known virulence mechanism of human pathogens involves modulation of the host external nucleotide pool (de Almeida Marques-da-Silva et al., 2008, la Sala et al., 2003). Cells damaged by invading pathogens release nucleotides into the external environment, which act as "danger signals" and stimulate pathogen-killing immune responses. The analysis identified a number of pathogen proteins with putative roles in modulation of these processes (Table 6.2), including a group of *Bordetella* spp. proteins with a unique similarity to human 5'nucleotidases (cluster 14), and human-like P-loop NTPases (cluster 9) from *Vibrio* spp. and *Helicobacter pylori*. These proteins may play host immunomodulatory roles via hydrolysis of extracellular nucleotides similar to that proposed for the protozoan pathogen, *Leishmania* (de Almeida Marques-da-Silva et al., 2008).

Several detected relationships involving known virulence factors are also worth mentioning. For instance, the well established mimicry mechanism of the *Legionella* (and *Rickettsia*) RalF protein is apparent in cluster 1. An additional *Legionella* effector (lepB), which is known to disrupt early secretory transport (Ingmundson et al., 2007), was also detected in cluster 8. Consistent with this function, the analysis detected similarity between lepB and human golgin 97, an ADP-ribosylation factor-like protein 1 (Arl1) GTPase that regulates secretory traffic (Lu et al., 2004).

The *Chlamydia* species possess a unique similarity pattern to a set of human PDZ

domain containing proteins (Figure 6.1, Table 6.2). The corresponding proteins are tail-specific proteases that have recently been found to degrade human transcription factors and thus interfere with the NF-kappa-B immune-response pathway (Lad et al., 2007).

Lastly, it is important to note that, while the analysis was restricted to the top 1% group, many mimicry/host-modulation mechanisms likely fall below the cutoff. For instance, the well known mimicry relationship for the *Yersinia* virulence factor, YopH, was detected in the top 2% of enrichment scores. The relationship was detected between YopH and human tyrosine phosphatases, which reflects the role of YopH in modulating host-proteins; more specifically, dephosphorylation of p130cas and the focal adhesion kinase (FAK), which prevents macrophage defense (Stebbins and Galan, 2001).

Analysis of bacterial protein similarities to human-ECM versus plant-ECM proteins

To test whether host-protein similarities distinguishes between pathogen-host-species relationships, the following analysis was performed. ECM-protein mimicry was chosen as a potential predictor of bacterial pathogenesis, and similarities were quantified between known human or plant pathogens/non-pathogens versus human or plant ECM proteins. As expected, the number of detected mimics of human ECM proteins were elevated in human pathogens vs non-pathogens ($p = 0.056$), and the number of detected mimics of plant ECM proteins were elevated in plant pathogens compared to non-pathogens ($p = 0.039$). However, plant pathogens did not have a greater number of human-ECM mimics than plant non-pathogens, and human pathogens did not have a greater number of plant-ECM mimics than human non-pathogens. This result shows that the trend of higher ECM mimicry in pathogens is specific to the host-pathogen relationship in question.

Furthermore, out of ten known plant pathogens, four were ranked in the top eight out of 157 genomes according to number of detected plant ECM mimics. These were species of *Xanthomonas*, and *Agrobacterium tumefaciens*, the cause of crown-gall disease. Interestingly, also among the top eight are bacteria that are plant pathogens or symbionts but were not included in the original annotation. These include *Sinorhizobium meliloti*, *Xanthomonas oryzae*, and *Xanthomonas campestris* (8004), the cause of black rot disease in plants

6.4 Discussion and Conclusion

In summary, the results clearly demonstrate that sequence mimicry of human proteins predicts virulence mechanisms of human pathogens. The preliminary analysis based on comparison to plant protein similarities, suggests that host-protein mimicry does broadly distinguish host-pathogen relationships.

Regarding the evolutionary mechanisms that underly the detected similarities between human and bacterial pathogen proteins, there are two main mechanisms responsible for these detected similarities: lateral transfer (e.g., carnitine palmitoyltransferase), and similarities due to repetitive sequences (e.g., collagen-like proteins) which may arise from convergent evolution. In the latter case, repetition appears to have been a successful strategy of bacterial pathogens to generate novel proteins composed of repeated motifs that functionally mimic or interfere with host proteins. Tandem repeats of bacterial sequence elements, rather than lateral transfer of entire host proteins, may be sufficient in many cases to generate novel mimics that contain key motifs that are functional in the host.

The predictions provide starting points for future experimental work characterizing the role of the detected proteins in bacterial pathogenesis. This study has analyzed only the tip of this spectrum, and future work expanding this analysis, and also evaluating the host-species specificity of this approach will be useful. Not all bacterial pathogens may contain structural or functional analogs/homologs of host proteins, and for some cases the functional similarities may be too subtle to detect using sequence methods alone. Thus, future use of more sophisticated statistical techniques and structural bioinformatics may provide added sensitivity. Prediction of novel pathogens, virulence mechanisms and host-pathogen relationships is becoming increasingly possible starting from genomic data alone. Rather than attempting to unify and predict virulence factors based on common sequence patterns, it can be highly useful to screen for the underlying determinants of pathogenicity.

Chapter 7

Conclusion

> "If we have learned anything at all in a century and a half of evolutionary biology, it is that facile generalizations are dangerous. The evolutionary process finds a way to create exceptions to every model we propose. Thus, it seems unwise to expect that functional diversification after gene duplication follows the same path every time" (Hughes, 2005).

Sequence similarity based approaches for assessing protein function operate on the assumption that sequence similarity implies similarity of function (Eisen, 1998). This assumption often makes sense because it seems reasonable that sequences that have diverged least from their common ancestor have also likely diverged least in terms of function. However, as demonstrated throughout this thesis, while this may be roughly correct as a general phenomenon, it can work poorly when assessing details of functional diversity. Proteins may have undergone key surface mutations, regulatory (e.g., gene expression) changes, novel repeats, recombination and other events that diversify function or scramble evolutionary relationships. In these situations, analyses based on overall similarity mask details of functional diversity and it becomes more useful to find and analyze additional traits that provide deeper clues into the basis of function or evolutionary origins.

As described in Chapter 2, assessment of overall similarity can mask important function-altering surface mutations. Chapters 3, 4, and 5 demonstrated how expression-related changes, repetition at the sub-domain level, mosaicism and recombination, can be important determinants of functional change—but are not captured by overall similarity. Finally,

Chapter 6 scaled up this idea to entire genomes, where overall similarity between genomes does not capture pathogenicity, and overall similarity cannot unify known virulence factors. In order to more accurately uncover function and/or evolutionary relationships in these cases, tailored approaches were developed, highly specific to the cases being examined as opposed to a general methodology. In AFPs and glycan-binding proteins, specific surface patterns were used to computationally recognize the determinants of ice-binding activity in unrelated structures. In *Arabidopsis* β-1,3-glucanases, known functional information integrated with gene expression divergence patterns was necessary to uncover patterns of functional diversification. In β-trefoils, consideration of internal repeats and carbohydrate-binding surface motifs was necessary to properly infer function and evolutionary history. In flagellins, phylogenetic mapping of the HVR (a key functional region) was necessary to reconstruct patterns of functional diversification that have occurred due to intragenic recombination. And in virulence factors, evaluation of genomic distributions of mimicry patterns in pathogenic versus non-pathogenic bacterial genomes was used to identify novel virulence factors of novel types and mechanisms. Consistent with the quotation that begins this concluding chapter, functional diversification does not follow the same path in each case. Throughout this thesis, this important idea was addressed by incorporating the biological specifics of the cases in question into computational analysis. Standard sequence-based analyses of conservation or adaptation (e.g., K_a/K_s ratios, site-specific rate shifts) could have inferred selection pressures in the sequences examined, but could not have uncovered the higher level functional patterns represented by surface motifs, gene expression traits, and the other traits examined.

Surface adaptations

In Chapter 2, ice-binding and glycan-binding proteins were investigated as examples of proteins that cannot be unified based on sequence. By computationally detecting a structural pattern that represents a more direct determinant of function, it was therefore possible to recognize structures, regardless of fold and evolutionary origin, that have converged upon the same function. Similar functional requirements (a necessity to bind ice, a glycan chain or similar ligand) have caused recurring molecular patterns in different structures that could be detected through sequence-independent, structure-based computational approaches. In both cases, "expert knowledge" was required to define these recurring patterns, which outlines a potential limitation of this approach as well as other general structure-based

methods of function prediction. No existing general method will measure such traits as 'ordered protein surfaces that mimics spacings of one or more ice planes' or 'protruding, coplanar aromatic sidechains in close proximity'. Existing methods focus rather on general traits (e.g., residue type, distance, hydrophobicity), which captures some but not enough functional information to recognize more complex structure-to-function relationships. Future methods that incorporate sophisticated 3D pattern-recognition approaches or detailed structure-to-function relationships described by experimental studies, may be necessary for effective large-scale structure-based functional annotation.

There are several implications and future directions regarding the work on AFPs. Computational detection of recurring surface features of AFPs (Doxey et al., 2006) supports previous models of AFP-ice binding, most notably those involving hydrophobic interactions and AFP-ice surface complementarity (Harding et al., 1999; Jia and Davies, 2002; Jorov and Zhorov, 2004). The analysis also quantitatively demonstrates convergent evolution of function in different folds, and reveals examples of molecular adaptation and functional diversification within protein families. LTP1 is one such example, and in related work, similar results have also been obtained when analyzing a β-1,3-glucanase from winter rye (Yaish et al., 2006).

Though not described in this thesis, computational modeling of the glucanase-AFP from winter rye (Yaish et al., 2006) revealed similar repetitive surface patterns that have developed in a specific member of the β-1,3-glucanase gene family. Winter rye glucanase-like AFP has developed an ordered surface with remarkable complementarity to the secondary prism plane of ice (Yaish et al., 2006). Like LTP, surface mutations in this protein therefore represent molecular adaptations to defense against ice-crystals. Ultimately, structural modeling of LTP-AFP and glucanase-AFP have supported the earlier hypothesis by Hon et al. (1995) that "subtle structural differences" may have evolved in the winter rye PR-AFPs to confer ice-binding activity.

Both LTPs (PR-14; Carvalho and Gomes, 2007; Sarowar et al., 2009) and β-1,-3-glucanases (PR-2; Leubner-Metzger and Meins, 1999) are known PR proteins. Other PR proteins with antifreeze activity have been isolated from winter rye and other plants, including thaumatin-like proteins (PR-5), chitinases, and polygalacturonase inhibitors (Hon et al., 1995; Griffith and Yaish, 2004). In winter rye, PR-AFPs exhibit hydrolytic activities, and cold-induced expression of these proteins confers increased resistance against freezing-induced injury or by pathogens present at freezing temperatures (i.e., snow mold)

(Hiilovaara-Teijo et al., 1999; Griffith and Yaish, 2004). Experimental work on winter rye glucanase-like AFP demonstrated that it is a cold-active protein, maintaining its hydrolytic function at sub-zero temperatures (Yaish et al., 2006). It is therefore possible that the true biological functions of LTP1, glucanase-AFP and other PR-AFPs, are as cold-active PR-proteins. Their ice-binding activity may be more of a secondary function than a primary one. Future work that tests winter rye AFP/PR-protein activity following cold/freezing stress and comparisons with PR-proteins without ice-binding activity will yield important insights into these questions.

The work on structure-based prediction of glycan-binding aromatic motifs (section 2.2) also revealed a putative molecular adaptation within a pathogenesis-related protein family (osmotin/thaumatin-like protein, PR-5d). Future experimental work is needed to further substantiate the hypothesis that the putative cellulose-binding motif (a PR-5d subfamily defining motif) is an evolutionary adaptation towards binding insoluble-carbohydrates (e.g., cellulose) in the cell walls of specific pathogens (e.g., *Phytophthora infestans*). *In vivo* or *in vitro* antifungal assays testing PR-5d's activity with and without the putative binding site is one possible direction. Comparative antifungal assays with other PR-5 proteins will also provide useful information.

The presence of recurring surface patterns in different AFP and GBP folds suggests that similar patterns may be engineered in other folds that are not related to known examples. Since the original publication (Doxey et al., 2006), the algorithm has been used to guide computational AFP design studies (Yu, 2010), and further work is in progress. The best use of structure-based algorithms like those developed in this thesis is probably in the early stages of screening to identify design scaffolds from the PDB. Candidate design templates would be those structures containing surface regions that, with a few minor modifications, may achieve the desired binding function.

Expression divergence

Surface patterns alone are insufficient to account for the full breadth of protein functional diversity. Molecular function, which was studied in Chapter 2, can remain relatively constant throughout the evolution of a protein family, but functions may diverge due to changes in gene expression. Many recent studies have demonstrated that expression divergence plays a significant role in functional diversification and adaptation of duplicate

genes (Makova and Li, 2003; Blanc and Wolfe 2004; He and Zhang, 2005; Duarte et al., 2006; Gu and Su, 2007; Ha et al., 2009; Fraser et al., 2010). Because gene expression traits are direct determinants of function (like the surface traits examined in Chapter 2), there is considerable potential for explaining functional diversity using microarray data (Woody et al., 2008).

In Chapter 3, patterns of gene expression divergence were investigated within the *Arabidopsis* β-1,3-glucanase gene family using microarray data. By combining microarray data with phylogeny, it was demonstrated that expression patterns converge and diverge independent of the evolutionary history of the coding sequence (Doxey et al., 2007). For example, the β-1,3-Gs involved in developmental processes versus pathogen-defense may look highly similar at the sequence level, but are part of two entirely different functional pathways. While this analysis focused largely on diverging patterns, again the idea of convergence was useful for functional annotation. By examining co-expression between β-1,3-glucanases and other genes outside of the family, it was possible to annotate new functions. Thus, while divergence occurred within the family, expression patterns of individual genes also converged upon those found in other families. This is possibly due to the nature of cis-regulatory motif evolution, which has a built-in capability for reversion and switching between common expression patterns encoded by the genome.

Classical work done over thirty years ago showed that isoenzymes, which may be an appropriate label for paralogous *Arabidopsis* β-1,3-glucanase genes, have diversified function through specialized expression in different tissues (Markert and Møller, 1959). Subsequent studies have explored regulatory divergence using large-scale microarray and genomic data (see Li et al., 2005 and Woody et al., 2008 for reviews). However, large-scale analyses have produced different results for different species. For example, gene expression divergence patterns in *Arabidopsis* are more unclear than in yeast (Li et al., 2005). As suggested by the quotation at the beginning of this chapter, while broad patterns may exist, in reality different genes/families/genomes adapt in different ways. The relative contribution of neofunctionalization, subfunctionalization, nonfunctionalization and other evolutionary forces is likely dependent on the case being analyzed. In the β-1,3-glucanase gene family, both neofunctionalization and subfunctionalization may have played roles. An example of neofunctionalization is the gain of extracellular expression and a pathogen-defense role in clade β1. It is also possible that the broad, ancestral expression pattern (M) may have been varied/split through subfunctionalization so that a full complement of β-1,3-G genes

is necessary to maintain full expression diversity.

For the purposes of functional annotation, perhaps what is more important than determination of the evolutionary forces contributing to expression divergence, is mapping relationships between changes in expression traits and specific biological functions. As demonstrated by the analysis in Chapter 3, known experimental information was critical for this purpose because it was able to determine whether the observed variations in gene expression are likely to be biologically relevant. However, as inferring functions by integrating microarray data and previous literature would be a laborious exercise to complete on a large-scale for each gene family, computational methods that distinguish adaptive evolution from neutral evolution of gene expression would be highly useful. A recent study introduced a computational test for lineage-specific selection on gene expression (Fraser et al., 2010). As stated by Fraser et al. (2010), there is a need for a "systematic, genome-scale approach to inferring regulatory adaptations". For future studies on functional divergence within gene families, phylogenomic approaches will be useful; that is, mapping of gene expression or other types of regulatory data (e.g., proteomics data) onto phylogenetic trees, and analysis of divergence patterns to infer adaptive evolution and functions of unannotated genes (Eisen, 1998; Sjölander, 2004).

Mosaicism, recombination, and lateral transfer

Chapters 2 and 3 dealt with cases in which functional alterations have occurred but the evolutionary relationships between sequences have remained largely intact. Thus, the problem required quantifying key surface- or expression-related patterns of functional diversification. Chapters 4 and 5 on the other hand dealt with cases in which overall similarity is a poor indicator of function because it is also a poor indicator of evolutionary history. The examples presented in these two chapters have been internally repeated or recombined and thus may be considered mosaic sequences. In Chapter 4, it was demonstrated that significant sequence and structural similarity between two β-trefoil proteins can exist while the sequences may be composed of distinct structural repeat units, in which case they have descended from a unique precursor repeat. In these cases, structure is conserved but the internal repeat (or symmetry) pattern can be different. Similar results have been obtained in analyses of β-propeller evolution (Chaudhuri et al., 2008), whereby individual blades have been amplified to generate novel β-propeller structures. Another recent study has ex-

amined this process on a broader scale, which they term internal amplification (Abraham et al., 2009).

The selective pressure for some proteins to maintain a high degree of internal symmetry is likely due to symmetry-related functions (e.g., multivalent binding). Abraham et al. (2009) have also suggested that internal amplifications may be maintained if symmetry is associated with selective features. Coevolution with the substrate may cause symmetry patterns to differ over time between species. This provides a model for the observation that PPODs have evolved a related but unique internal symmetry pattern from that found in the *Bacillus* family (section 4.2). In both of these cases, due to overall sequence similarity, these domains are annotated as "fascin-like", but their functions are likely different from fascin and can be revealed through inspection of their internal repeats rather than overall domain similarities.

An interesting commonality of the proteins identified by screening β-trefoils for recent examples of triplication is a role in carbohydrate-binding. While ricin-type and AbfB β-trefoils are associated with carbohydrate-binding, the identification of a carbohydrate-binding site (experimentally verified by Böttger et al., unpublished) is a more surprising result. A second surprising finding related to PPODs is that they are closely related to domains from bacterial carbohydrate-related enzymes, and thus appear to have been horizontally transferred to *Hydra*. Recently, the genome of *Hydra magnipapillata* has been sequenced (Chapman et al., 2010) and PPOD sequences are among those identified by screening for horizontally transferred genes. The authors also reported that numerous candidate horizontally transferred genes are sugar-modifying enzymes, three of which are normally involved in bacterial lipopolysaccharide synthesis (Chapman et al., 2010). Thus, PPODs appear to be part of a broader group of horizontally transferred bacterial genes that have imparted carbohydrate related functions to *Hydra*.

Flagellins (analyzed in section 5.2), like β-trefoils, required analysis of internal regions in order to properly characterize evolutionary relationships and infer functional similarities. In this case, similarities between the functionally important internal region (the HVR) were not necessarily consistent with overall similarity or with phylogenetic relationships of adjacent regions of the sequence. Two flagellins can be distantly related according to their N-terminal or C-terminal domains, but have acquired (likely through intragenic recombination) the same type of HVR (e.g., the collagenase-like HVR).

A number of previous studies have shown that flagellins exhibit mosaicism and undergo recombination (Frankel et al., 1989; Harrington et al., 1997; Smith et al., 1990; Smith and Selander, 1990; Wang et al., 2003). However, the phylogenetically distant homoplasies indicative of long-range intragenic recombination identified in section 5.2, have not been demonstrated previously. Typically, recombination-detection methods involves multiple alignments (most often of DNA) and/or phylogenetic trees (see Posada, 2002). Therefore, these analyses are restricted to sequences with a limited amount of sequence divergence (e.g., DNA sequences from closely related species). The flagellin HVR, however, evolves so rapidly that most of the flagellin HVR nucleotide or amino acid sequences could not be properly aligned. The BLAST-based clustering and phylogenetic reconstruction approach that was applied to flagellins is one way to partially overcome this problem, and was able to uncover phylogenetically distant patterns of homoplasy. It would be interesting to apply a similar technique to other rapidly-evolving regions such as HVRs in viral sequences to see whether similar evolutionary patterns have occurred.

CNTs (analyzed in section 5.1) represent another rapidly-evolving sequence family, whose evolutionary origins and relationships to other existing proteins are unclear. To infer relationships and putative ancestral functions, the neurotoxin gene cluster as a whole was therefore analyzed for recurring similarities. The approach involved analyzing sequence, structural, motif, and functional links between multiple (independent) proteins encoded within the neurotoxin gene cluster, with the assumption that the gene cluster as a whole is likely to function in the same pathway. Indeed, numerous adjacent genes (e.g., the hemagglutinin and non-hemaagglutinin genes) function as components of the neurotoxin progenitor complex (Ardnt, 2004).

Future work investigating further functional (or possible evolutionary) links between CNTs, collagenases and flagellins will be useful. Whether the identified similarities are the result of a series of local recombination events or more distant evolutionary relationships is too difficult to analyze using current data. Furthermore, the identified links between numerous proteins encoded in the neurotoxin gene cluster and collagenases, collagen-binding domains, and collagen itself is interesting from a functional standpoint. For instance, it would be interesting to test whether the collagen HCRn domain functionally mimics the collagen NC4 domain in order to interact with host proteins.

Mimicry and virulence factors

Lastly, in Chapter 6, a common theme among the proteins studied in this thesis (virulence) was explored in greater detail. Bacterial pathogen genomes and virulence factors cannot be predicted necessarily based on overall genome/sequence similarity. An underlying feature indicative of virulence, host-protein mimicry, was then explored as a computational predictor of existing and potentially novel virulence factors. Similar to the examples presented in Chapter 2, virulence represents a functional characteristic that has evolved independently in proteins by convergent or parallel evolution. Thus, the feature assessed needs to be independent of a single protein or taxonomic family. There are likely diverse evolutionary scenarios underlying the independent origins of this feature in numerous protein lineages. These include lateral transfer, convergent or parallel evolution of similar sequences in pathogens/hosts from similar ancestral scaffolds, repeats of sequence fragments containing motifs similar to motifs found in human proteins, among others.

This analysis was largely based on detectable sequence similarities. Thus, an obvious direction for future research is to apply some of the approaches developed earlier in the thesis (from Chapter 2 in particular) to detect structural mimicry that may not be detectable the sequence level. While there exist isolated examples of virulence factor structural mimicry with host proteins (Stebbins, 2001; Sikora et al., 2005), a large-scale structure-based analysis has not been performed.

The evolutionary arms race, coevolution, and the "Red Queen"

Many of the cases explored in this thesis are examples of host-pathogen coevolutionary adaptation at the molecular level. They may also be considered examples of what has been labeled by van Valen (1973) the "Red Queen Hypothesis". The name is taken from Lewis Carrol's "Through the Looking Glass", in which the Red Queen says to Alice "in this place it takes all the running you can do, to keep in the same place". In other words, an "evolutionary arms race" (see Dawkins and Krebs, 1979) takes place between hosts and pathogens which requires the continual development of evolutionary novelties in the genes and proteins involved just in order to maintain relative fitness.

Examples of molecular adaptations from this thesis that may have resulted from host-pathogen coevolutionary arms races are summarized below:

- Winter rye PR-AFPs: Development of ordered surfaces/ice-binding activity in these PR-proteins may facilitate PR activity in low temperatures where the plant is susceptible to psychrophilic or psychrotropic fungal pathogens such as snow mold.

- PR-5d: The coplanar aromatic surface motif that has originated in the PR-5d subfamily may be a defense adaptation against insoluble-carbohydrate (e.g., cellulose) containing pathogens such as *Phytophthora infestans*.

- β-1,3-glucanases: Changes in gene expression and key protein functional regions (e.g., GPI-anchor loss) of ancestral developmental genes may be responsible for the subfamily of extracellularly-secreted PR β-1,3-glucanases capable of hydrolyzing β-1,3-glucan in fungal pathogens. A counter-adaptation of pathogens is the production of β-1,3-glucanase inhibitor proteins (Stahl and Bishop, 2000).

- Internal repetition in ricin-type β-trefoils: Repeat amplification in ricin-type β-trefoils may enhance symmetric interactions with host cell carbohydrate receptors (Hazes, 1996; Pohleven et al., 2009), or rapidly change the sequence to avoid immune detection while maintaining 3D structure.

- Flagellin HVR recombination: Rapid mutation and recombination of HVRs may generate antigenic diversity in bacterial pathogens (Reid et al., 1999).

- CNTs: Extensive divergence and structural rearrangements in CNTs from ancestral proteins may be the result of extensive coevolution with host proteins. The detected sequence and structural similarities between the HCRn domain and collagen NC4 domain may be due to this.

- Pathogens/VF-mimics: Coevolution between host (human) and pathogenic bacteria has generated protein mimicry patterns that can be detected at the sequence and structural level.

Final remarks

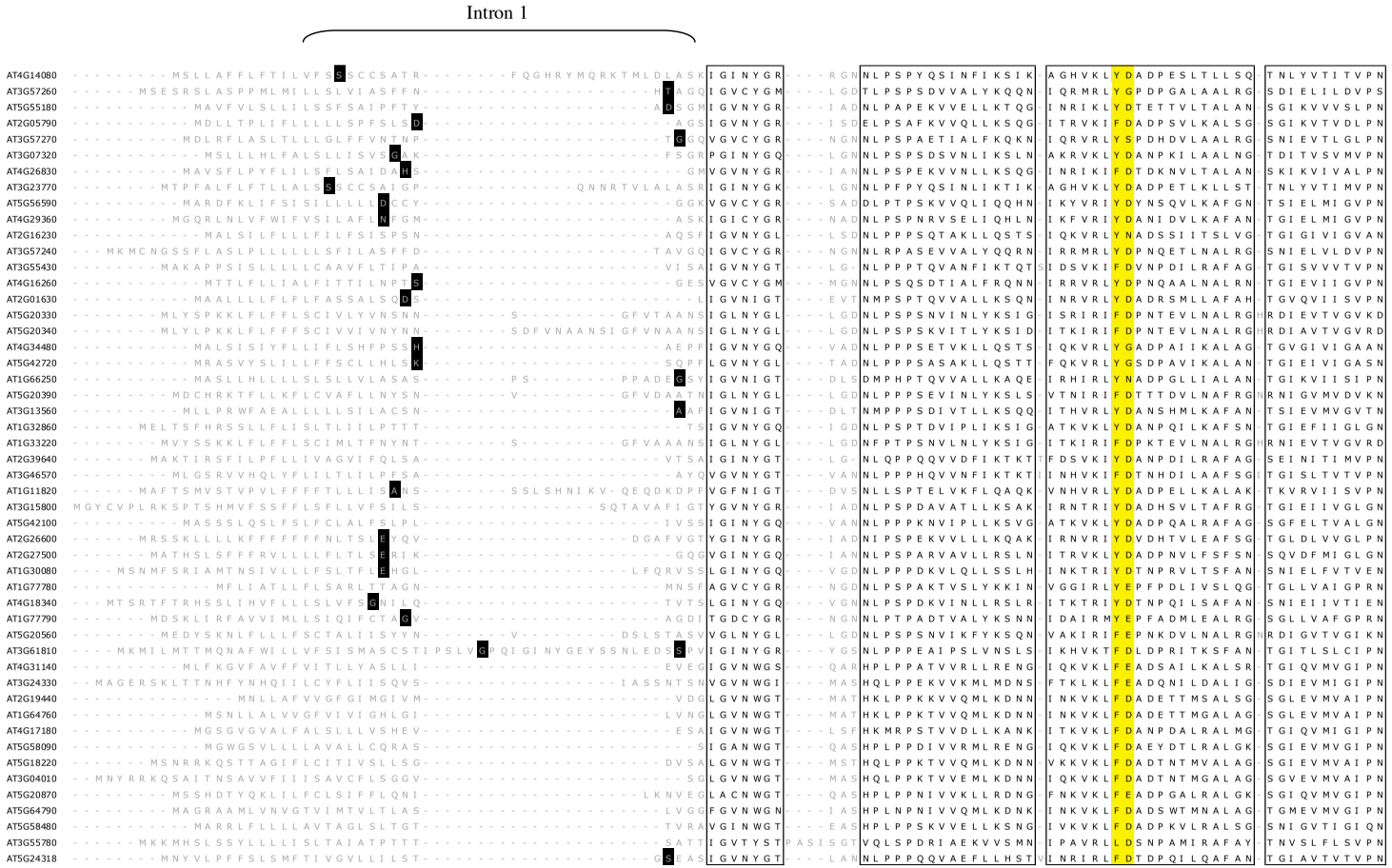
Despite the development of many bioinformatic tools over the last decade, the most significant approach in bioinformatics developed thus far remains the identification of homologous relationships through sequence alignments (i.e., the BLAST search; Altschul et al., 1990). For many cases, this approach answers two fundamental questions for any given sequence, that being its evolutionary relationships and function. One way to advance bioinformatics is to determine when and why this fundamental approach fails, and how to improve methods for inferring functions and evolutionary relationships. As suggested by the examples illustrated in this thesis, a general solution will likely require both lumping (grouping related sequences) and splitting (analyzing patterns of functional change). Beyond classification of sequences and evolutionary relationships, it will be necessary to examine diversification of protein function within multiple dimensions (e.g., binding, folding, catalysis, regulation), which will require considerable integration of experimental data and computational methods and a suite of computational tools.

Common sequence-based techniques for analyzing molecular adaptation are able to identify genes and subregions of genes that have undergone various types of selection (Yang and Bielawski, 2000). Bioinformatics approaches that rely on conservation also inherently aim to identify regions that undergo purifying selection due to functional constraints (Lichtarge et al., 1996). These techniques however do not indicate function itself. Future work in protein bioinformatics and computational biology will need to better integrate data that is more directly tied to function, and examine patterns of trait change through phylogenetic approaches (Eisen, 1998; Sjölander, 2004). Careful consideration will be needed to account for sequences that have been scrambled by recombination and other mechanisms, allowing for the possibility of mosaicism. Pathogen-defense proteins and virulence factors offer a window into these types of evolutionary change because of the frequent, ongoing nature of host-pathogen coevolution. While these cases represent ongoing examples of molecular evolution, they also offer a window into our distant, molecular evolutionary past.

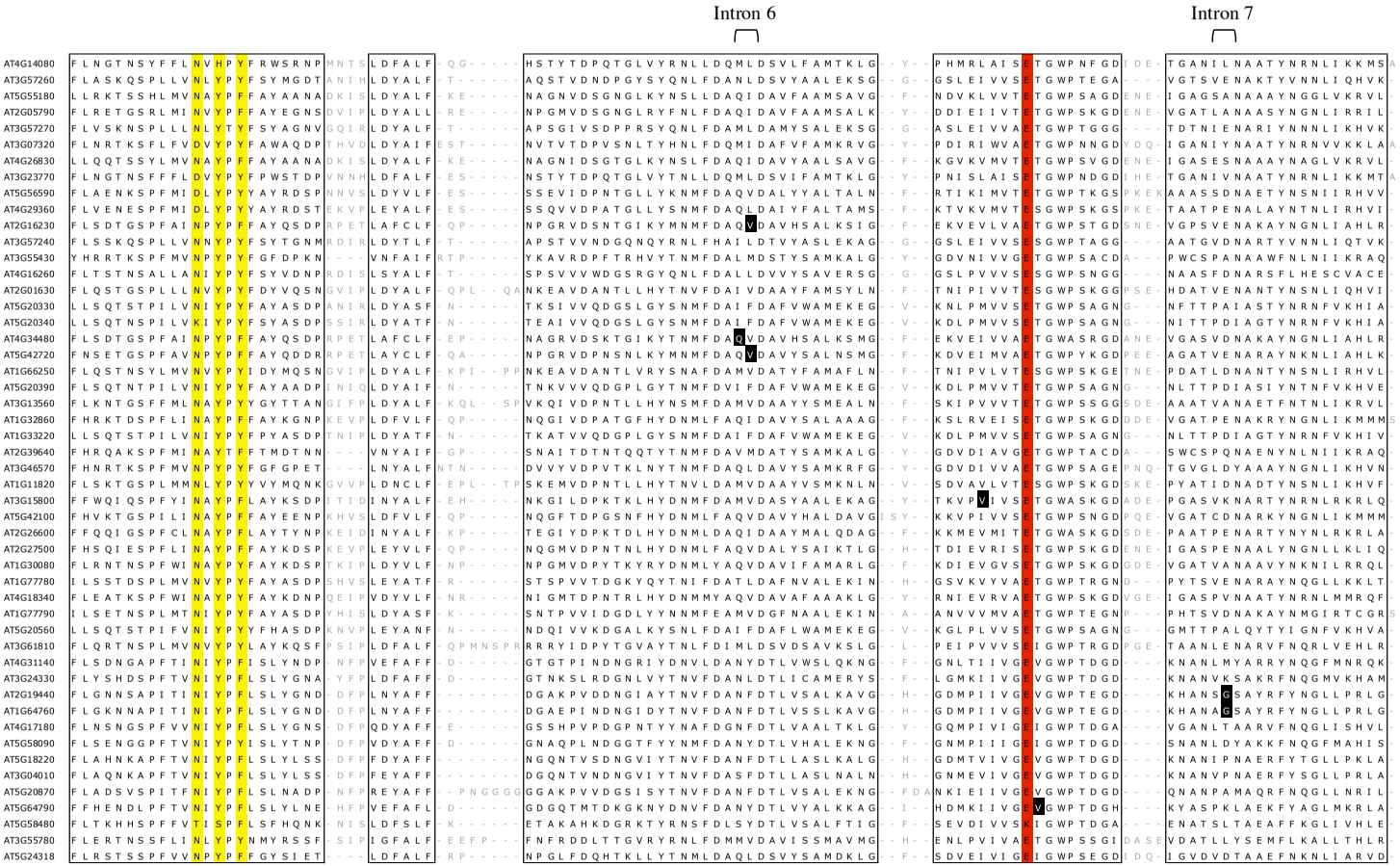
Appendices

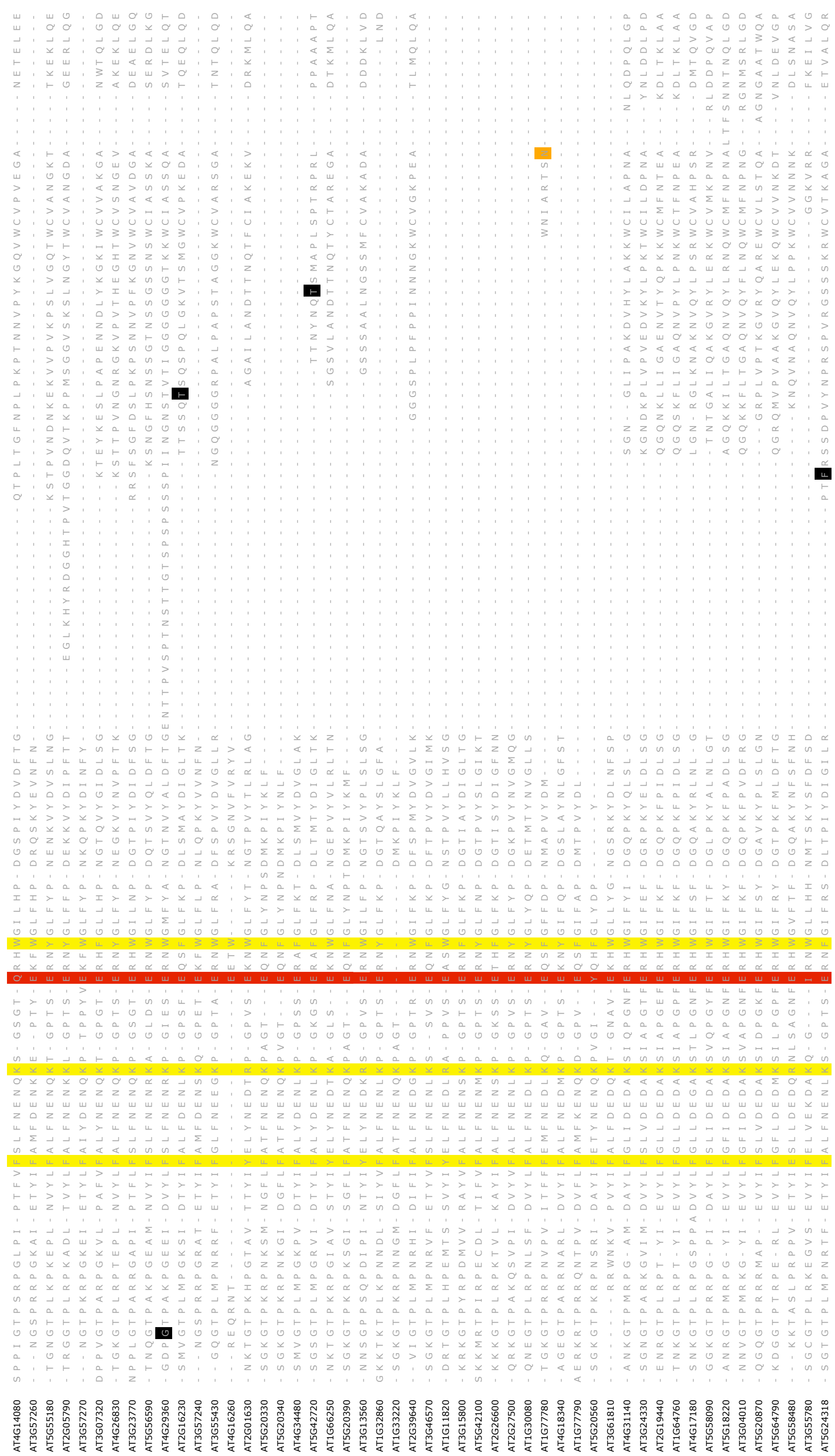
Appendix 1

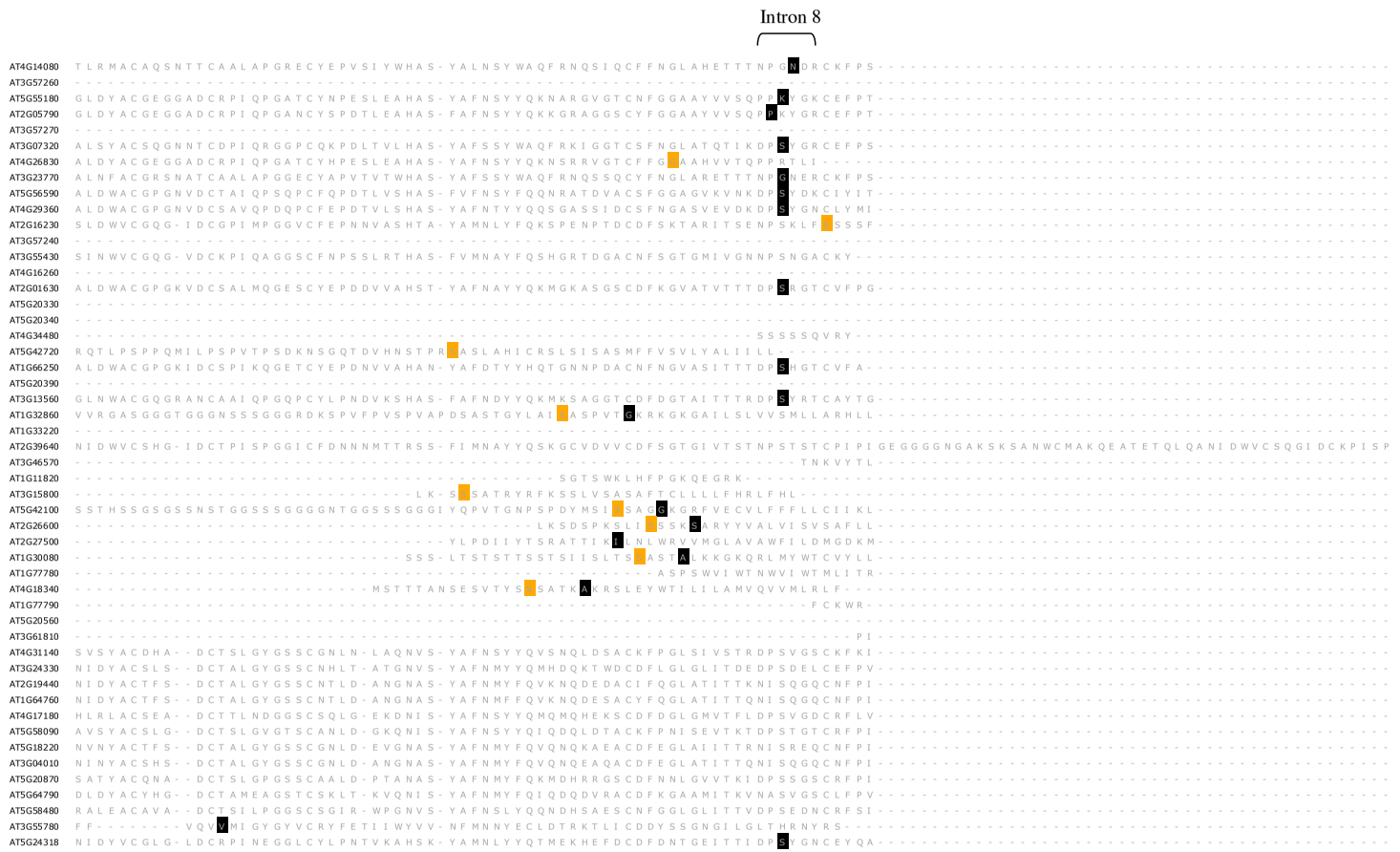
Multiple alignment of *Arabidopsis* β-1,3-glucanase sequences and key characteristics. Intron positions (black boxes), GPI attachment sites (orange), catalytic glutamatic acid residues (red) and other previously determined functional residues (yellow) are indicated.


Intron 1

Intron 2
Intron 3
Intron 4
Intron 5


Intron 6
Intron 7
AT4G14080
AT3G57260
AT5G55180
AT2G05790
AT3G57270
AT3G07320
AT4G26830
AT3G23770
AT5G56590
AT4G29360
AT2G16230
AT3G57240
AT3G55430
AT4G16260
AT2G01630
AT5G20330
AT5G20340
AT4G34480
AT5G42720
AT1G66250
AT5G20390
AT3G13560
AT1G32860
AT1G33220
AT2G39640
AT3G46570
AT1G11820
AT3G15800
AT5G42100
AT2G26600
AT2G27500
AT1G30080
AT1G77780
AT4G18340
AT1G77790
AT5G20560
AT3G61810
AT4G31140
AT3G24330
AT2G19440
AT1G64760
AT4G17180
AT5G58090
AT5G18220
AT3G04010
AT5G20870
AT5G64790
AT5G58480
AT3G55780
AT5G24318

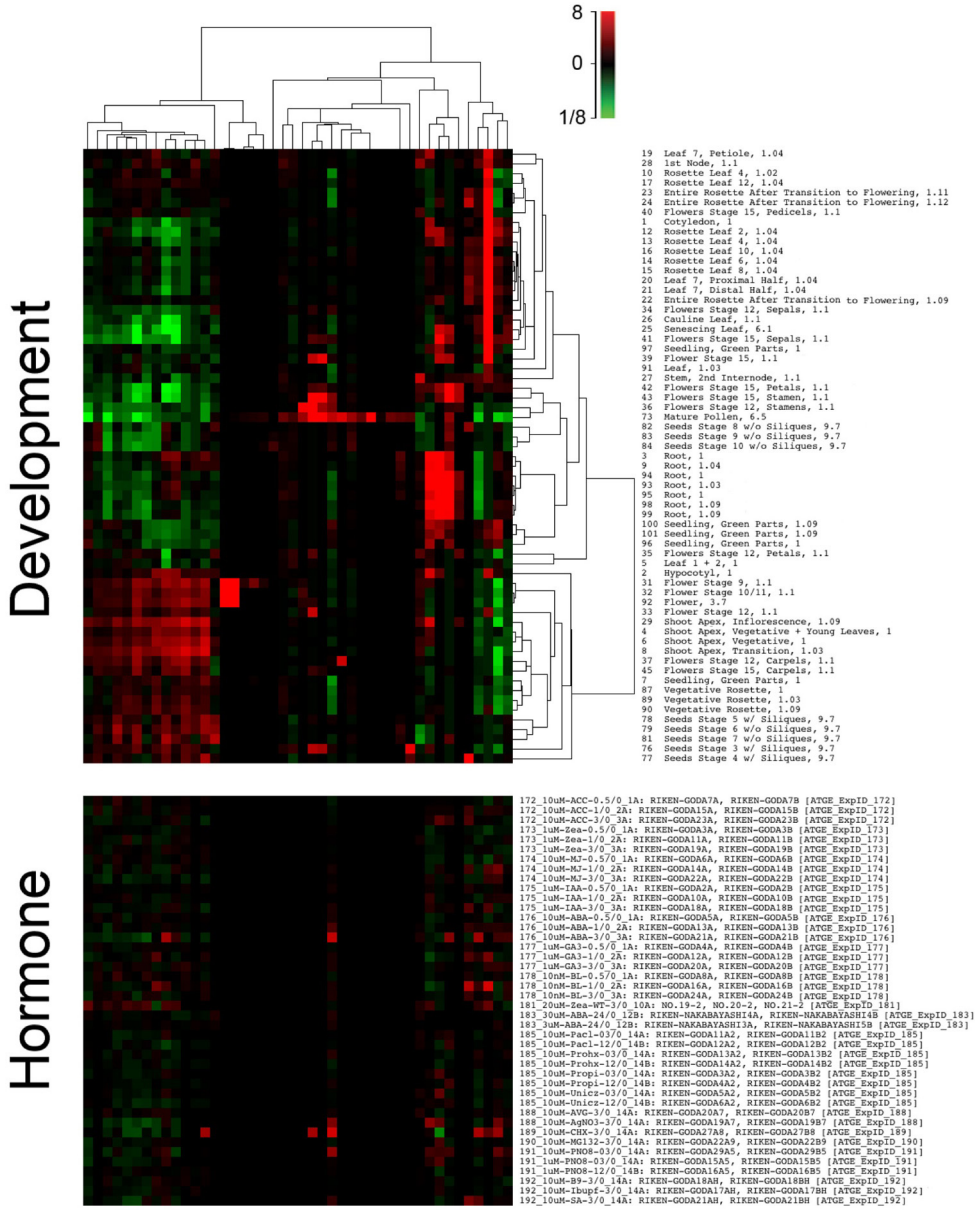


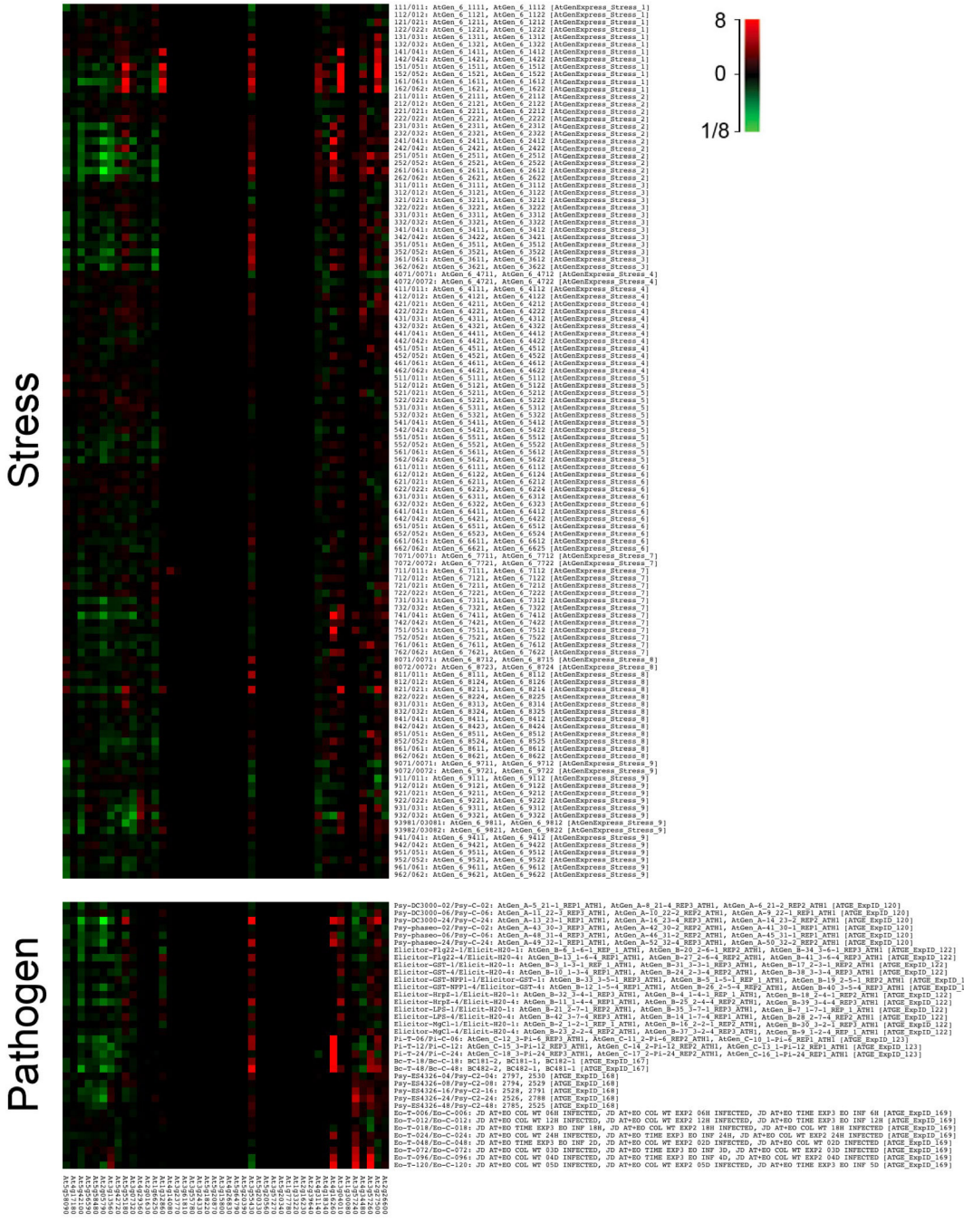

Intron 8
AT4G14080
AT3G57260
AT5G55180
AT2G05790
AT3G57270
AT3G07320
AT4G26830
AT3G23770
AT5G56590
AT4G29360
AT2G16230
AT3G57240
AT3G55430
AT4G16260
AT2G01630
AT5G20330
AT5G20340
AT4G34480
AT5G42720
AT1G66250
AT5G20390
AT3G13560
AT1G32860
AT1G33220
AT2G39640
AT3G46570
AT1G11820
AT3G15800
AT5G42100
AT2G26600
AT2G27500
AT1G30080
AT1G77780
AT4G18340
AT1G77790
AT5G20560
AT3G61810
AT4G31140
AT3G24330
AT2G19440
AT1G64760
AT4G17180
AT5G58090
AT5G18220
AT3G04010
AT5G20870
AT5G64790
AT5G58480
AT3G55780
AT5G24318

Intron 9
AT4G14080
AT3G57260
AT5G55180
AT2G05790
AT3G57270
AT3G07320
AT4G26830
AT3G23770
AT5G56590
AT4G29360
AT2G16230
AT3G57240
AT3G55430
AT4G16260
AT2G60630
AT5G20330
AT5G20340
AT4G34480
AT5G44720
AT1G66250
AT5G20390
AT3G13560
AT1G32860
AT1G33220
AT2G39640
AT3G46570
AT1G11820
AT3G15800
AT5G42100
AT2G26600
AT2G27500
AT1G30080
AT1G77780
AT4G18340
AT1G77790
AT5G20560
AT3G61810
AT4G31140
AT3G24330
AT2G19440
AT1G64760
AT4G17180
AT5G58090
AT5G18220
AT3G04010
AT5G20870
AT5G64790
AT5G58480
AT3G55780
AT5G24318

Appendix 2

Clustered *Arabidopsis* microarray data obtained from the Botany Array Resource (Toufighi et al., 2005) and the Genevestigator database (Zimmermann et al., 2004) divided into developmental-, hormone-, stress-, and pathogen-related experiments. Clustering was performed using developmental expression data. AtGenExpress sample IDs are indicated for developmental data, and AGI numbers are shown beneath the pathogen data (following page).




Stress
Pathogen
8
0
1/8

Appendix 3

Pairwise correlations involving *Arabidopsis* β-1,3-glucanases using gene expression data.

Correlation	Glucanase gene	Correlated gene	Description
Pairwise correlations using tissue expression data			
0.829	At3g13560	At1g02730	cellulose synthase family protein
0.853	At1g11820	At1g02730	cellulose synthase family protein
0.856	At5g58480	At1g02730	cellulose synthase family protein
0.858	At2g01630	At1g02730	cellulose synthase family protein
0.868	At5g55180	At1g02730	cellulose synthase family protein
0.882	At2g05790	At1g02730	cellulose synthase family protein
0.932	At4g29360	At1g02730	cellulose synthase family protein
0.939	At3g07320	At1g02730	cellulose synthase family protein
0.842	At5g55180	At1g20930	cell division control protein, putative
0.85	At3g07320	At1g20930	cell division control protein, putative
0.881	At4g29360	At1g20930	cell division control protein, putative
0.878	At5g20330	At1g23420	inner no outer protein
0.802	At5g58480	At1g76540	cell division control protein, putative
0.819	At1g66250	At1g76540	cell division control protein, putative
0.835	At1g11820	At1g76540	cell division control protein, putative
0.841	At2g05790	At1g76540	cell division control protein, putative
0.871	At2g01630	At1g76540	cell division control protein, putative
0.886	At5g55180	At1g76540	cell division control protein, putative
0.916	At3g07320	At1g76540	cell division control protein, putative
0.942	At4g29360	At1g76540	cell division control protein, putative
0.814	At3g13560	At1g78770	cell division cycle family protein
0.822	At1g66250	At1g78770	cell division cycle family protein
0.846	At1g11820	At1g78770	cell division cycle family protein
0.865	At3g07320	At1g78770	cell division cycle family protein
0.866	At2g01630	At1g78770	cell division cycle family protein
0.912	At4g29360	At1g78770	cell division cycle family protein
0.941	At3g57260	At2g14610	pathogenesis-related protein 1
0.801	At1g11820	At2g38620	cell division control protein, putative
0.804	At2g01630	At2g38620	cell division control protein, putative
0.868	At5g55180	At2g38620	cell division control protein, putative
0.879	At3g07320	At2g38620	cell division control protein, putative
0.91	At4g29360	At2g38620	cell division control protein, putative
0.844	At1g11820	At3g02210	COBRA cell expansion protein COBL3
0.856	At4g17180	At3g02210	COBRA cell expansion protein COBL3
0.859	At5g56590	At3g02210	COBRA cell expansion protein COBL3
0.866	At4g29360	At3g02210	COBRA cell expansion protein COBL3
0.866	At2g01630	At3g02210	COBRA cell expansion protein COBL3
0.906	At1g66250	At3g02210	COBRA cell expansion protein COBL3
0.966	At4g14080	At3g11980	male sterility protein 2
0.978	At3g23770	At3g11980	male sterility protein 2
0.909	At4g16260	At3g12500	basic endochitinase (PR-3)

Pairwise correlations involving *Arabidopsis* β-1,3-glucanases using gene expression data.

Correlation	Glucanase gene	Correlated gene	Description
Pairwise correlations using tissue expression data			
0.809	At3g07320	At3g25100	cell division control protein-related
0.81	At3g13560	At3g25100	cell division control protein-related
0.845	At4g29360	At3g25100	cell division control protein-related
0.849	At1g11820	At3g25100	cell division control protein-related
0.859	At2g01630	At3g25100	cell division control protein-related
0.802	At5g55180	At3g48150	cell division cycle family protein / CDC family protein
0.82	At1g11820	At3g48150	cell division cycle family protein / CDC family protein
0.832	At2g01630	At3g48150	cell division cycle family protein / CDC family protein
0.836	At4g29360	At3g48150	cell division cycle family protein / CDC family protein
0.847	At1g66250	At3g48150	cell division cycle family protein / CDC family protein
0.805	At5g42720	At4g34160	cyclin delta-3 (CYCD3)
0.813	At3g13560	At4g34160	cyclin delta-3 (CYCD3)
0.829	At5g58480	At4g34160	cyclin delta-3 (CYCD3)
0.838	At2g05790	At4g34160	cyclin delta-3 (CYCD3)
0.862	At1g11820	At4g34160	cyclin delta-3 (CYCD3)
0.87	At2g01630	At4g34160	cyclin delta-3 (CYCD3)
0.87	At3g07320	At4g34160	cyclin delta-3 (CYCD3)
0.888	At4g29360	At4g34160	cyclin delta-3 (CYCD3)
0.843	At5g56590	At5g18580	tonneau 2 (TON2)
0.844	At1g66250	At5g18580	tonneau 2 (TON2)
0.865	At4g29360	At5g18580	tonneau 2 (TON2)
0.872	At2g01630	At5g18580	tonneau 2 (TON2)
0.902	At5g20390	At5g39400	pollen-specific phosphatase (PTEN1)
0.927	At5g64790	At5g39400	pollen-specific phosphatase (PTEN1)
Pairwise correlations using stress expression data			
Correlation	Glucanase gene	Correlated gene	Description
0.802	At3g04010	At4g36010	pathogenesis-related thaumatin family protein

Appendix 4

NCBI annotation of a representative PPOD family protein.

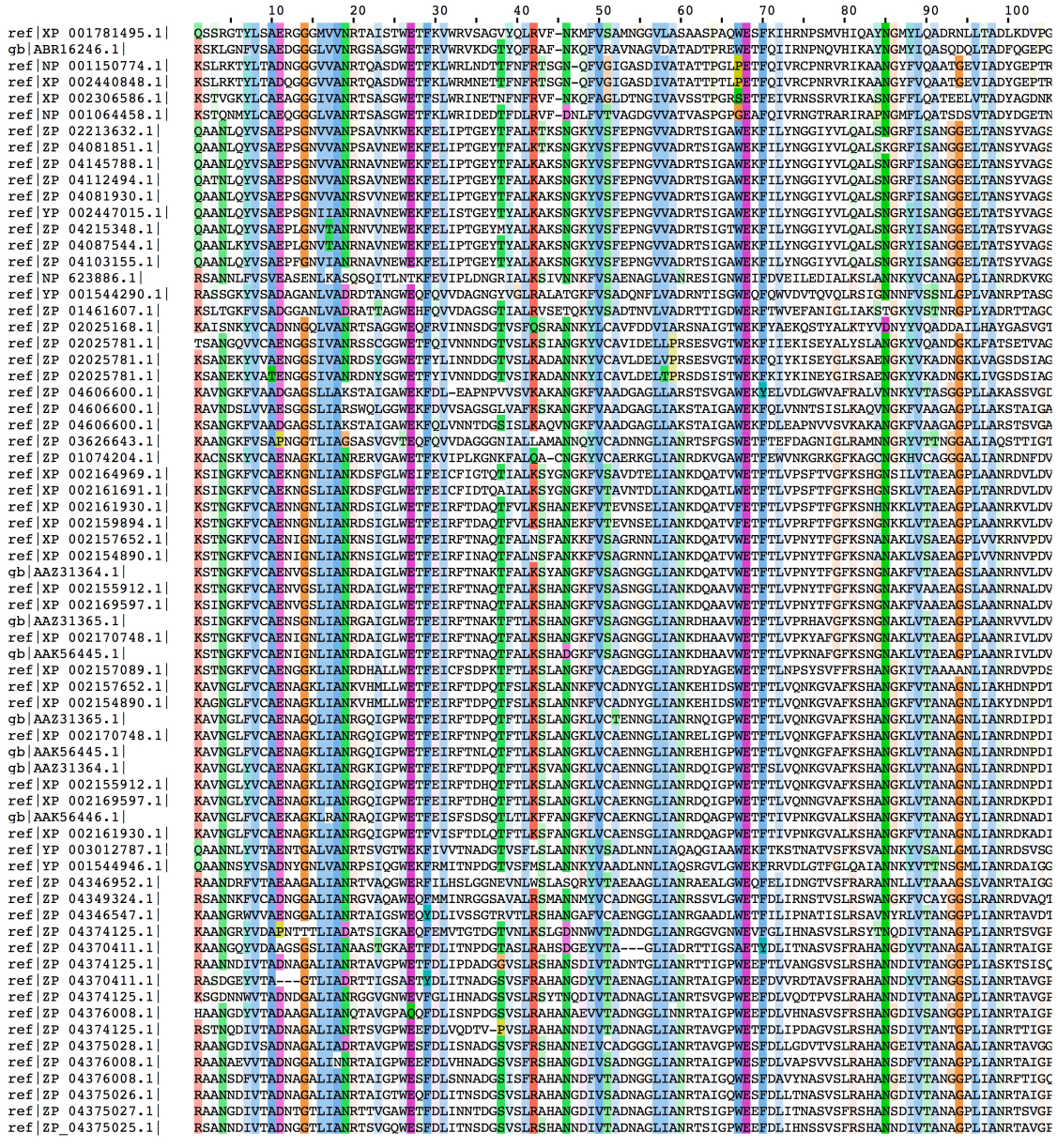
PPOD1 peroxidase [*Hydra vulgaris*]

GenBank: AAZ31364.1

```
FEATURES             Location/Qualifiers
     source          1..290
                     /organism="Hydra vulgaris"
                     /strain="AEP"
                     /db_xref="taxon:6087"
     Protein         1..290
                     /product="PPOD1 peroxidase"
                     /name="putative"
     Region          38..154
                     /region_name="Fascin"
                     /note="Fascin-like domain; members include
                     actin-bundling/crosslinking proteins facsin,
                     histoactophilin and singed;  identified in sea urchin,
                     Drosophila, Xenopus, rodents, and humans; The fascin-like
                     domain adopts a beta-trefoil topology and contains an...;
                     cl00187"
                     /db_xref="CDD:119602"
     Site            58
                     /site_type="other"
                     /note="PKC phosphorylation site"
                     /db_xref="CDD:29332"
     Region          117..244
                     /region_name="Fascin"
                     /note="Fascin-like domain; members include
                     actin-bundling/crosslinking proteins facsin,
                     histoactophilin and singed;  identified in sea urchin,
                     Drosophila, Xenopus, rodents, and humans; The fascin-like
                     domain adopts a beta-trefoil topology and contains an...;
                     cd00257"
                     /db_xref="CDD:29332"
     Site            145
                     /site_type="other"
                     /note="PKC phosphorylation site"
                     /db_xref="CDD:29332"
     CDS             1..290
                     /coded_by="DQ073555.1:38..910"
```

Appendix 5

Alignment of PPODs and related sequences used in phylogenetic analysis.



Appendix 6

Flagellin HVR cluster memberships (NCBI accession numbers) and tree mapping information.

Cluster 1: 63 477 578 721

gi|19910973|dbj|BAB87738.1| Clostridium_haemolyticum

gi|119898824|ref|YP_934037.1| Azoarcus_sp._BH72

gi|88858114|ref|ZP_01132756.1| Pseudoalteromonas_tunicata_D2

gi|95929889|ref|ZP_01312630.1| Desulfuromonas_acetoxidans_DSM_684

Cluster 2: 90 91 92 93 94 95 96 97 98

gi|150020412|ref|YP_001305766.1| Thermosipho_melanesiensis_BI429

gi|217077418|ref|YP_002335136.1| Thermosipho_africanus_TCF52B

gi|154250154|ref|YP_001410979.1| Fervidobacterium_nodosum_Rt17-B1

gi|170287974|ref|YP_001738212.1| Thermotoga_sp._RQ2

gi|157363124|ref|YP_001469891.1| Thermotoga_lettingae_TMO

gi|256543606|ref|ZP_05471002.1| Thermotoga_naphthophila_RKU-10

gi|160903242|ref|YP_001568823.1| Petrotoga_mobilis_SJ95

gi|154250151|ref|YP_001410976.1| Fervidobacterium_nodosum_Rt17-B1

gi|217076162|ref|YP_002333878.1| Thermosipho_africanus_TCF52B

Cluster 3: 105 180 1021 1022 1023 1024 1025 1026 1028 1029 1030 1031

gi|229541548|ref|ZP_04430608.1| Bacillus_coagulans_36D1

gi|167628838|ref|YP_001679337.1| Heliobacterium_modesticaldum_Ice1

gi|50082705|gb|AAT70050.1| Pectinatus_cerevisiiphilus

gi|50082711|gb|AAT70053.1| Pectinatus_frisingensis

gi|50082713|gb|AAT70054.1| Pectinatus_frisingensis

gi|50082715|gb|AAT70055.1| Pectinatus_frisingensis

gi|238927674|ref|ZP_04659434.1| Selenomonas_flueggei_ATCC_43531

gi|261877419|ref|ZP_06004024.1| Selenomonas_noxia_ATCC_43541

gi|260888511|ref|ZP_05899774.1| Selenomonas_sputigena_ATCC_35185

gi|260888404|ref|ZP_05899667.1| Selenomonas_sputigena_ATCC_35185

gi|238928168|ref|ZP_04659928.1| Selenomonas_flueggei_ATCC_43531

gi|261877280|ref|ZP_06003885.1| Selenomonas_noxia_ATCC_43541

Cluster 4: 130 133 134 135 244 245

gi|114565767|ref|YP_752921.1|

gi|229213021|ref|ZP_04339371.1| Dethiosulfovibrio_peptidovorans_DSM_11002

gi|229213309|ref|ZP_04339658.1| Dethiosulfovibrio_peptidovorans_DSM_11002

gi|229848920|ref|ZP_04468989.1| Thermanaerovibrio_acidaminovorans_DSM_6589

gi|227422491|ref|ZP_03905601.1| Denitrovibrio_acetiphilus_DSM_12809

gi|227422541|ref|ZP_03905651.1| Denitrovibrio_acetiphilus_DSM_12809

Cluster 5: 155 156 158 159 164

gi|2583084|gb|AAB82610.1| Butyrivibrio_fibrisolvens

gi|2583088|gb|AAB82613.1| Butyrivibrio_fibrisolvens

gi|113911603|gb|ABI48283.1| Lachnospiraceae_bacterium_A4

gi|113911605|gb|ABI48284.1| Lachnospiraceae_bacterium_A4

gi|225376513|ref|ZP_03753734.1| Roseburia_inulinivorans_DSM_16841

Cluster 6: 170 171 175 177

gi|160939645|ref|ZP_02086993.1| Clostridium_bolteae_ATCC_BAA-613

gi|266621937|ref|ZP_06114872.1| Clostridium_hathewayi_DSM_13479

gi|225570000|ref|ZP_03779025.1| Clostridium_hylemonae_DSM_15053

gi|225570005|ref|ZP_03779030.1| Clostridium_hylemonae_DSM_15053

Cluster 7: 182 183 184 185 186 187 188

gi|168335502|ref|ZP_02693583.1|

gi|168335504|ref|ZP_02693585.1|

gi|168335505|ref|ZP_02693586.1|

gi|168335503|ref|ZP_02693584.1|

gi|168335506|ref|ZP_02693587.1|

gi|168335507|ref|ZP_02693588.1|

gi|168335508|ref|ZP_02693589.1|

Cluster 8: 246 247 248 249 250 251 252 253

gi|160881968|ref|YP_001560936.1| Clostridium_phytofermentans_ISDg

gi|218134725|ref|ZP_03463529.1| Bacteroides_pectinophilus_ATCC_43243

gi|238917954|ref|YP_002931471.1| Eubacterium_eligens_ATCC_27750

gi|229825727|ref|ZP_04451796.1| Abiotrophia_defectiva_ATCC_49176

gi|225375130|ref|ZP_03752351.1| Roseburia_inulinivorans_DSM_16841

gi|257413133|ref|ZP_04742102.2| Roseburia_intestinalis_L1-82

gi|238922463|ref|YP_002935976.1| Eubacterium_rectale_ATCC_33656

gi|260439399|ref|ZP_05793215.1| Butyrivibrio_crossotus_DSM_2876

Cluster 9: 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 278 279 280 281 282 283 284 288 289 290 295 334 335 338

gi|120316|sp|P06175.2|FLIC_SALRU

gi|38049772|gb|AAR10505.1| Salmonella_enterica

gi|38049888|gb|AAR10563.1| Salmonella_enterica

gi|38050099|gb|AAR10668.1| Salmonella_enterica

gi|38050119|gb|AAR10678.1| Salmonella_enterica

gi|38050127|gb|AAR10682.1| Salmonella_enterica

gi|38050221|gb|AAR10729.1| Salmonella_enterica

gi|46359071|gb|AAR10751.1| Salmonella_enterica

gi|62180532|ref|YP_216949.1|

gi|116489768|gb|ABJ98791.1| Salmonella_typhimurium

gi|116489798|gb|ABJ98806.1| Salmonella_typhimurium

gi|161613476|ref|YP_001587441.1|

gi|168233721|ref|ZP_02658779.1|

gi|168259757|ref|ZP_02681730.1|

gi|168262130|ref|ZP_02684103.1|

gi|189490722|gb|ACE00522.1| Salmonella_typhimurium

gi|194444631|ref|YP_002041222.1|

gi|204930971|ref|ZP_03221797.1|

gi|224584549|ref|YP_002638347.1|

gi|834003|gb|AAB33952.1|

gi|38050223|gb|AAR10730.1| Salmonella_enterica

gi|116489816|gb|ABJ98815.1| Salmonella_typhimurium

gi|116489820|gb|ABJ98817.1| Salmonella_typhimurium

gi|38049682|gb|AAR10470.1| Salmonella_bongori

gi|38049684|gb|AAR10471.1| Salmonella_bongori

gi|38049688|gb|AAR10473.1| Salmonella_bongori

gi|50830960|gb|AAT81645.1| Salmonella_bongori_serovar_66

gi|51342391|gb|AAU01391.1| Salmonella_bongori

gi|1333832|emb|CAA27130.1|

gi|38050053|gb|AAR10645.1| Salmonella_enterica

gi|16129870|ref|NP_416433.1|

gi|89143092|emb|CAI29258.1| Escherichia_coli

gi|89143094|emb|CAI29259.1| Escherichia_coli

gi|30059884|gb|AAP13314.1| Escherichia_coli

gi|23477614|gb|AAN34780.1| Escherichia_coli

gi|33590252|gb|AAQ22687.1| Escherichia_coli

gi|46093563|dbj|BAD14977.1| Escherichia_coli

Cluster 10: 291 292 293 294 306 307 311 312 313 314 316 317 318 319 320 321 322 323 330 350 351

gi|30059870|gb|AAP13307.1| Escherichia_coli

gi|30059878|gb|AAP13311.1| Escherichia_coli

gi|30059906|gb|AAP13325.1| Escherichia_coli

gi|30059920|gb|AAP13332.1| Escherichia_coli

gi|30059868|gb|AAP13306.1| Escherichia_coli

gi|30059912|gb|AAP13328.1| Escherichia_coli

gi|110642029|ref|YP_669759.1| Escherichia_coli_536

gi|170682489|ref|YP_001743319.1| Escherichia_coli_SMS-3-5

gi|191168280|ref|ZP_03030073.1| Escherichia_coli_B7A

gi|215487128|ref|YP_002329559.1|

gi|30059876|gb|AAP13310.1| Escherichia_coli

gi|188494683|ref|ZP_03001953.1| Escherichia_coli_53638

gi|193071144|ref|ZP_03052068.1| Escherichia_coli_E110019

gi|33590257|gb|AAQ22689.1| Escherichia_coli

gi|218705418|ref|YP_002412937.1| Escherichia_coli_UMN026

gi|88683097|emb|CAJ77154.1| Escherichia_coli

gi|125659369|dbj|BAF46886.1| Escherichia_coli

gi|125659376|dbj|BAF46892.1| Escherichia_coli

gi|238764886|ref|ZP_04625826.1| Yersinia_kristensenii_ATCC_33638

gi|154710451|gb|ABS84292.1| Escherichia_coli

gi|237731969|ref|ZP_04562450.1| Citrobacter_sp._30_2

Cluster 11: 296 298 299 300 301 302 304 308 326 327 333 393

gi|33590233|gb|AAQ22678.1| Escherichia_coli

gi|30059896|gb|AAP13320.1| Escherichia_coli

gi|30059902|gb|AAP13323.1| Escherichia_coli

gi|33590231|gb|AAQ22677.1| Escherichia_coli

gi|37651054|emb|CAD97428.1| Escherichia_coli

gi|125659364|dbj|BAF46882.1| Escherichia_coli

gi|30059910|gb|AAP13327.1| Escherichia_coli

gi|30059928|gb|AAP13336.1| Escherichia_coli

gi|239918890|gb|ACS34794.1| Pantoea_agglomerans

gi|258635983|ref|ZP_05728744.1| Pantoea_sp._At-9b

gi|261821178|ref|YP_003259284.1| Pectobacterium_wasabiae_WPP163

gi|116696296|ref|YP_841872.1| Ralstonia_eutropha_H16

Cluster 12: 297 303 305 309 310 315 397

gi|6009841|dbj|BAA85085.1| Escherichia_coli

gi|6009851|dbj|BAA85090.1| Escherichia_coli

gi|6009849|dbj|BAA85089.1| Escherichia_coli

gi|33590241|gb|AAQ22682.1| Escherichia_coli

gi|91211142|ref|YP_541128.1| Escherichia_coli_UTI89

gi|218699513|ref|YP_002407142.1| Escherichia_coli_IAI39

gi|226944844|ref|YP_002799917.1| Azotobacter_vinelandii_DJ

Cluster 13: 332 339 340 341 342 343 344 345 346 347 348 349 352

gi|238794023|ref|ZP_04637641.1| Yersinia_intermedia_ATCC_29909

gi|16901494|gb|AAL30165.1|AF345848_1 Escherichia_coli

gi|33590237|gb|AAQ22680.1| Escherichia_coli

gi|112820171|gb|ABI23966.1| Escherichia_coli

gi|260868515|ref|YP_003234917.1|

gi|210062496|ref|YP_002300513.1|

gi|555642|gb|AAA53493.1|

gi|38050007|gb|AAR10622.1| Salmonella_enterica

gi|218548540|ref|YP_002382331.1| Escherichia_fergusonii_ATCC_35469

gi|46909844|gb|AAT06390.1| Citrobacter_freundii

gi|46909846|gb|AAT06391.1| Citrobacter_freundii

gi|227333567|ref|ZP_03837223.1| Citrobacter_youngae_ATCC_29220

gi|238912469|ref|ZP_04656306.1|

Cluster 14: 357 358 359 360

gi|1169696|sp|P42272.3|FLIC1_PROMI

gi|197285479|ref|YP_002151351.1| Proteus_mirabilis_HI4320

gi|227355969|ref|ZP_03840361.1| Proteus_mirabilis_ATCC_29906

gi|261344107|ref|ZP_05971751.1| Providencia_rustigianii_DSM_4541

Cluster 15: 408 409 410 411

gi|21242719|ref|NP_642301.1|

gi|58426798|gb|AAW75835.1|

gi|66768558|ref|YP_243320.1|

gi|86211138|gb|ABC87266.1| Xanthomonas_campestris_pv._campestris

Cluster 16: 432 433 434 435 845

gi|83647504|ref|YP_435939.1| Hahella_chejuensis_KCTC_2396

gi|83647506|ref|YP_435941.1| Hahella_chejuensis_KCTC_2396

gi|88799043|ref|ZP_01114624.1| Reinekea_sp._MED297

gi|94500260|ref|ZP_01306793.1| Oceanobacter_sp._RED65

gi|241775553|ref|ZP_04772836.1| Allochromatium_vinosum_DSM_180

Cluster 17: 448 449 450 451 452 453 454 455 456 457 458

gi|119898983|ref|YP_934196.1| Azoarcus_sp._BH72

gi|119898994|ref|YP_934207.1| Azoarcus_sp._BH72

gi|120613037|ref|YP_972715.1|

gi|241763661|ref|ZP_04761711.1| Acidovorax_delafieldii_2AN

gi|121611311|ref|YP_999118.1| Verminephrobacter_eiseniae_EF01-2

gi|121611312|ref|YP_999119.1| Verminephrobacter_eiseniae_EF01-2

gi|222112256|ref|YP_002554520.1| Diaphorobacter_sp._TPSY

gi|222112255|ref|YP_002554519.1| Diaphorobacter_sp._TPSY

gi|241763660|ref|ZP_04761710.1| Acidovorax_delafieldii_2AN

gi|218187850|gb|EEC70277.1| Oryza_sativa_Indica_Group

gi|120555503|ref|YP_959854.1| Marinobacter_aquaeolei_VT8

Cluster 18: 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 539 540 541 542 543 544

gi|17366557|sp|Q56570.1|FLAE_VIBAN

gi|260767717|ref|ZP_05876652.1| Vibrio_furnissii_CIP_102972

gi|261252311|ref|ZP_05944884.1| Vibrio_orientalis_CIP_102891

gi|260772313|ref|ZP_05881229.1| Vibrio_metschnikovii_CIP_69.14

gi|262404490|ref|ZP_06081045.1| Vibrio_sp._RC586

gi|37680677|ref|NP_935286.1| Vibrio_vulnificus_YJ016

gi|218710261|ref|YP_002417882.1| Vibrio_splendidus_LGP32

gi|254506945|ref|ZP_05119084.1| Vibrio_parahaemolyticus_16

gi|91223533|ref|ZP_01258798.1| Vibrio_alginolyticus_12G01

gi|260899922|ref|ZP_05908317.1| Vibrio_parahaemolyticus_AQ4037

gi|149189130|ref|ZP_01867418.1| Vibrio_shilonii_AK1

gi|17366561|sp|Q56571.3|FLAD_VIBAN

gi|17366567|sp|Q56574.3|FLAC_VIBAN

gi|260767760|ref|ZP_05876695.1| Vibrio_furnissii_CIP_102972

gi|148979885|ref|ZP_01815763.1| Vibrionales_bacterium_SWAT-3

gi|260767716|ref|ZP_05876651.1| Vibrio_furnissii_CIP_102972

gi|153836942|ref|ZP_01989609.1| Vibrio_parahaemolyticus_AQ3810

gi|260776494|ref|ZP_05885389.1| Vibrio_coralliilyticus_ATCC_BAA-450

gi|261252312|ref|ZP_05944885.1| Vibrio_orientalis_CIP_102891

gi|262404489|ref|ZP_06081044.1| Vibrio_sp._RC586

gi|260772283|ref|ZP_05881199.1| Vibrio_metschnikovii_CIP_69.14

gi|261253605|ref|ZP_05946178.1| Vibrio_orientalis_CIP_102891

gi|260772314|ref|ZP_05881230.1| Vibrio_metschnikovii_CIP_69.14

gi|260774947|ref|ZP_05883848.1| Vibrio_coralliilyticus_ATCC_BAA-450

gi|149188075|ref|ZP_01866370.1| Vibrio_shilonii_AK1

gi|149188076|ref|ZP_01866371.1| Vibrio_shilonii_AK1

gi|163801836|ref|ZP_02195733.1| Vibrio_sp._AND4

gi|218710260|ref|YP_002417881.1| Vibrio_splendidus_LGP32

gi|262404531|ref|ZP_06081086.1| Vibrio_sp._RC586

gi|17366564|sp|Q56572.3|FLAB_VIBAN

gi|27365271|ref|NP_760799.1| Vibrio_vulnificus_CMCP6

gi|218708814|ref|YP_002416435.1| Vibrio_splendidus_LGP32

gi|254506908|ref|ZP_05119047.1| Vibrio_parahaemolyticus_16

gi|261252313|ref|ZP_05944886.1| Vibrio_orientalis_CIP_102891

gi|262404488|ref|ZP_06081043.1| Vibrio_sp._RC586

gi|260767715|ref|ZP_05876650.1| Vibrio_furnissii_CIP_102972

gi|260772315|ref|ZP_05881231.1| Vibrio_metschnikovii_CIP_69.14

gi|209695749|ref|YP_002263679.1| Aliivibrio_salmonicida_LFI1238

gi|209695748|ref|YP_002263678.1| Aliivibrio_salmonicida_LFI1238

gi|209695752|ref|YP_002263682.1| Aliivibrio_salmonicida_LFI1238

gi|262276404|ref|ZP_06054213.1| Grimontia_hollisae_CIP_101886

gi|262276405|ref|ZP_06054214.1| Grimontia_hollisae_CIP_101886

gi|209695754|ref|YP_002263684.1| Aliivibrio_salmonicida_LFI1238

gi|17366580|sp|Q60246.1|FLAA_VIBAN

gi|258626881|ref|ZP_05721686.1| Vibrio_mimicus_VM603

gi|260772282|ref|ZP_05881198.1| Vibrio_metschnikovii_CIP_69.14

gi|260767761|ref|ZP_05876696.1| Vibrio_furnissii_CIP_102972

gi|254508448|ref|ZP_05120568.1| Vibrio_parahaemolyticus_16

gi|261253606|ref|ZP_05946179.1| Vibrio_orientalis_CIP_102891

gi|37679155|ref|NP_933764.1| Vibrio_vulnificus_YJ016

gi|148976402|ref|ZP_01813108.1| Vibrionales_bacterium_SWAT-3

gi|218708812|ref|YP_002416433.1| Vibrio_splendidus_LGP32

gi|254230132|ref|ZP_04923527.1| Vibrio_sp._Ex25

gi|149188074|ref|ZP_01866369.1| Vibrio_shilonii_AK1

gi|90411013|ref|ZP_01219027.1| Photobacterium_profundum_3TCK

gi|172087731|ref|YP_205245.3| Vibrio_fischeri_ES114

gi|209695935|ref|YP_002263865.1| Aliivibrio_salmonicida_LFI1238

gi|37679159|ref|NP_933768.1| Vibrio_vulnificus_YJ016

gi|163802593|ref|ZP_02196485.1| Vibrio_sp._AND4

gi|254230130|ref|ZP_04923525.1| Vibrio_sp._Ex25

gi|260899316|ref|ZP_05907711.1| Vibrio_parahaemolyticus_AQ4037

Cluster 19: 545 561 562 563 564 565 566

gi|126165554|gb|ABN80234.1| Pseudomonas_aeruginosa

gi|405547|gb|AAA62844.1| Pseudomonas_putida

gi|3098301|gb|AAC63948.1| Pseudomonas_putida

gi|126215004|emb|CAM35537.1| Pseudomonas_fluorescens

gi|229591862|ref|YP_002873981.1| Pseudomonas_fluorescens_SBW25

gi|167034934|ref|YP_001670165.1| Pseudomonas_putida_GB-1

gi|146281784|ref|YP_001171937.1| Pseudomonas_stutzeri_A1501

Cluster 20: 572 720 722 723

gi|239996310|ref|ZP_04716834.1| Alteromonas_macleodii_ATCC_27126

gi|77918423|ref|YP_356238.1| Pelobacter_carbinolicus_DSM_2380

gi|227326892|ref|ZP_03830916.1|

gi|260597239|ref|YP_003209810.1| Cronobacter_turicensis

Cluster 21: 682 683 684 686 687

gi|105894597|gb|ABF77762.1| Burkholderia_cenocepacia_AU_1054

gi|221202118|ref|ZP_03575153.1| Burkholderia_multivorans_CGD2M

gi|254246747|ref|ZP_04940068.1| Burkholderia_cenocepacia_PC184

gi|238025311|ref|YP_002909543.1| Burkholderia_glumae_BGR1

gi|187925802|ref|YP_001897444.1| Burkholderia_phytofirmans_PsJN

Cluster 22: 718 719 843 844

gi|68164529|gb|AAY87258.1| uncultured_bacterium_BAC17H8

gi|68164530|gb|AAY87259.1| uncultured_bacterium_BAC17H8

gi|192361599|ref|YP_001982391.1| Cellvibrio_japonicus_Ueda107

gi|192362369|ref|YP_001982389.1| Cellvibrio_japonicus_Ueda107

Cluster 23: 725 726 729 750 751 752 753

gi|49477319|ref|YP_035879.1|

gi|87242478|gb|ABD33778.1| Bacillus_thuringiensis

gi|189339478|gb|ABD33730.2| Bacillus_thuringiensis_serovar_kyushuensis

gi|71067029|gb|AAZ22698.1| Bacillus_cereus

gi|218230951|ref|YP_002366445.1| Bacillus_cereus_B4264

gi|87242350|gb|ABD33714.1| Bacillus_thuringiensis_serovar_guiyangiensis

gi|87242378|gb|ABD33728.1| Bacillus_thuringiensis_serovar_tolworthi

Cluster 24: 732 733 734 735 736

gi|87242358|gb|ABD33718.1| Bacillus_thuringiensis_serovar_finitimus

gi|87242434|gb|ABD33756.1| Bacillus_thuringiensis_serovar_poloniensis

gi|87242456|gb|ABD33767.1| Bacillus_thuringiensis_serovar_vazensis

gi|87242454|gb|ABD33766.1| Bacillus_thuringiensis_serovar_graciosensis

gi|87242450|gb|ABD33764.1| Bacillus_thuringiensis_serovar_bolivia

Cluster 25: 742 743 746 783 784

gi|87242484|gb|ABD33781.1| Bacillus_mycoides

gi|229029440|ref|ZP_04185523.1| Bacillus_cereus_AH1271

gi|87242342|gb|ABD33710.1| Bacillus_thuringiensis_serovar_malayensis

gi|87242352|gb|ABD33715.1| Bacillus_thuringiensis_serovar_chanpaisis

gi|87242490|gb|ABD33784.1| Bacillus_weihenstephanensis

Cluster 26: 754 755 760 762 765 775

gi|52143698|ref|YP_083130.1| Bacillus_cereus_E33L

gi|87242414|gb|ABD33746.1| Bacillus_thuringiensis_serovar_cameroun

gi|87242422|gb|ABD33750.1| Bacillus_thuringiensis_serovar_sooncheon

gi|87242464|gb|ABD33771.1| Bacillus_thuringiensis_serovar_azorensis

gi|196045108|ref|ZP_03112341.1| Bacillus_cereus_03BB108

gi|229090723|ref|ZP_04221956.1| Bacillus_cereus_Rock3-42

Cluster 27: 756 757 758 759 761 764 766 767 768 769 770 773

gi|87242436|gb|ABD33757.1| Bacillus_thuringiensis_serovar_palmanyolensis

gi|87242346|gb|ABD33712.1| Bacillus_thuringiensis_serovar_brasilensis

gi|189339486|gb|ABD33736.2| Bacillus_thuringiensis_serovar_pondicheriensis

gi|87242404|gb|ABD33741.1| Bacillus_thuringiensis_serovar_silo

gi|87242442|gb|ABD33760.1| Bacillus_thuringiensis_serovar_iberica

gi|189164128|gb|ABU96045.2| Bacillus_thuringiensis_serovar_novosibirsk

gi|87242418|gb|ABD33748.1| Bacillus_thuringiensis_serovar_seoulensis

gi|87242430|gb|ABD33754.1| Bacillus_thuringiensis_serovar_xiaguangiensis

gi|189339490|gb|ABD33740.2| Bacillus_thuringiensis_serovar_neoleonensis

gi|87242504|gb|ABD33791.1| Bacillus_weihenstephanensis

gi|87242444|gb|ABD33761.1| Bacillus_thuringiensis_serovar_pingluonsis

gi|229132588|ref|ZP_04261436.1| Bacillus_cereus_BDRD-ST196

Cluster 28: 838 839 840 841 842

gi|89094888|ref|ZP_01167820.1| Oceanospirillum_sp._MED92

gi|90021845|ref|YP_527672.1| Saccharophagus_degradans_2-40

gi|254785445|ref|YP_003072874.1| Teredinibacter_turnerae_T7901

gi|90021847|ref|YP_527674.1| Saccharophagus_degradans_2-40

gi|254785380|ref|YP_003072809.1| Teredinibacter_turnerae_T7901

Cluster 29: 930 931 932 933 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985

gi|94264908|ref|ZP_01288681.1| delta_proteobacterium_MLMS-1

gi|94270092|ref|ZP_01291650.1| delta_proteobacterium_MLMS-1

gi|94265621|ref|ZP_01289364.1| delta_proteobacterium_MLMS-1

gi|258594707|ref|ZP_05709517.1| Desulfurivibrio_alkaliphilus_AHT2

gi|224370746|ref|YP_002604910.1| Desulfobacterium_autotrophicum_HRM2

gi|144202|gb|AAA23027.1| Campylobacter_coli

gi|9296983|sp|Q46114.3|FLB2_CAMJE

gi|57238729|ref|YP_179511.1| Campylobacter_jejuni_RM1221

gi|167005900|ref|ZP_02271658.1|

gi|3290044|gb|AAC25647.1| Campylobacter_jejuni

gi|57168249|ref|ZP_00367388.1| Campylobacter_coli_RM2228

gi|57505230|ref|ZP_00371159.1| Campylobacter_upsaliensis_RM3195

gi|153951910|ref|YP_001397579.1|

gi|86151038|ref|ZP_01069254.1|

gi|1888390|emb|CAA72431.1| Campylobacter_sp.

gi|48716112|dbj|BAD22844.1| Campylobacter_lari

gi|48716120|dbj|BAD22848.1| Campylobacter_lari

gi|68533178|dbj|BAE06136.1| Campylobacter_lari

gi|68533179|dbj|BAE06137.1| Campylobacter_lari

gi|1888389|emb|CAA72430.1| Campylobacter_sp.

gi|68533172|dbj|BAE06131.1| Campylobacter_lari

gi|68533173|dbj|BAE06132.1| Campylobacter_lari

gi|16754889|dbj|BAB71799.1| Campylobacter_lari

gi|18857919|dbj|BAB85480.1| Campylobacter_lari

gi|28372324|dbj|BAC56987.1| Campylobacter_lari

gi|28372332|dbj|BAC56991.1| Campylobacter_lari

gi|28372340|dbj|BAC56995.1| Campylobacter_lari

gi|28372322|dbj|BAC56986.1| Campylobacter_lari

gi|729512|sp|Q07910.3|FLAB_HELMU

gi|262344292|gb|ACY56029.1| Helicobacter_hepaticus

gi|1706820|sp|P50612.1|FLAA_HELMU

gi|7387670|sp|Q9XB38.3|FLAA_HELFE

gi|261839570|gb|ACX99335.1| Helicobacter_pylori_52

gi|237751661|ref|ZP_04582141.1| Helicobacter_bilis_ATCC_43879

gi|154175408|ref|YP_001408971.1| Campylobacter_curvus_525.92

gi|157164988|ref|YP_001467626.1| Campylobacter_concisus_13826

gi|223040859|ref|ZP_03611124.1| Campylobacter_rectus_RM3267

gi|255322858|ref|ZP_05363999.1| Campylobacter_showae_RM3277

gi|32266863|ref|NP_860895.1| Helicobacter_hepaticus_ATCC_51449

gi|224437397|ref|ZP_03658368.1| Helicobacter_cinaedi_CCUG_18818

gi|237751363|ref|ZP_04581843.1| Helicobacter_bilis_ATCC_43879

gi|149194794|ref|ZP_01871888.1| Caminibacter_mediatlanticus_TB-2

gi|224373626|ref|YP_002607998.1| Nautilia_profundicola_AmH

gi|224373627|ref|YP_002607999.1| Nautilia_profundicola_AmH

gi|5441621|emb|CAB46859.1| Helicobacter_felis

gi|261838983|gb|ACX98748.1| Helicobacter_pylori_52

gi|34558473|ref|NP_908288.1| Wolinella_succinogenes_DSM_1740

gi|237753123|ref|ZP_04583603.1| Helicobacter_winghamensis_ATCC_BAA-430

gi|261886307|ref|ZP_06010346.1|

gi|261886308|ref|ZP_06010347.1|

gi|34557839|ref|NP_907654.1| Wolinella_succinogenes_DSM_1740

gi|224419151|ref|ZP_03657157.1| Helicobacter_canadensis_MIT_98-5491

gi|237752810|ref|ZP_04583290.1| Helicobacter_winghamensis_ATCC_BAA-430

gi|242309434|ref|ZP_04808589.1| Helicobacter_pullorum_MIT_98-5489

gi|229531686|ref|ZP_04421071.1| Sulfurospirillum_deleyianum_DSM_6946

Cluster 30: 997 998 999 1000 1001

gi|1169694|sp|P46210.1|FLAA_AQUPY

gi|15606990|ref|NP_214372.1| Aquifex_aeolicus_VF5

gi|163783304|ref|ZP_02178297.1| Hydrogenivirga_sp._128-5-R1-1

gi|206891026|ref|YP_002249309.1| Thermodesulfovibrio_yellowstonii_DSM_11347

gi|152990389|ref|YP_001356111.1| Nitratiruptor_sp._SB155-2

Cluster 31: 1013 1014 1015 1016 1017 1018 1019

gi|32472702|ref|NP_865696.1| Rhodopirellula_baltica_SH_1

gi|32473113|ref|NP_866107.1| Rhodopirellula_baltica_SH_1

gi|32474579|ref|NP_867573.1| Rhodopirellula_baltica_SH_1

gi|87306354|ref|ZP_01088501.1| Blastopirellula_marina_DSM_3645

gi|87312157|ref|ZP_01094261.1| Blastopirellula_marina_DSM_3645

gi|149176412|ref|ZP_01855026.1| Planctomyces_maris_DSM_8797

gi|229537726|ref|ZP_04426862.1| Planctomyces_limnophilus_DSM_3776

Cluster 32: 1107 1108 1111 1112 1113

gi|152222|gb|AAA26278.1|

gi|15964423|ref|NP_384776.1| Sinorhizobium_meliloti_1021

gi|15964424|ref|NP_384777.1| Sinorhizobium_meliloti_1021

gi|150395494|ref|YP_001325961.1| Sinorhizobium_medicae_WSM419

gi|150395492|ref|YP_001325959.1| Sinorhizobium_medicae_WSM419

Appendix 7

Human proteins and bacterial protein BLAST hits used in Figure 6.1.

Human Protein	Pathogen Protein	E-value	Cluster #
gi|45827716|ref|NP_996792.1|	lpl1919	8.00E-14	1
gi|117606360|ref|NP_056125.3|	lpl1919	2.00E-13	1
gi|56790299|ref|NP_002770.3|	lpl1919	9.00E-15	1
gi|56788370|ref|NP_036587.2|	lpl1919	9.00E-13	1
gi|4758964|ref|NP_004753.1|	lpl1919	8.00E-30	1
gi|110349763|ref|NP_059430.2|	lpl1919	1.00E-29	1
gi|53832026|ref|NP_055890.1|	lpl1919	4.00E-19	1
gi|50582989|ref|NP_055684.3|	lpl1919	5.00E-18	1
gi|7019505|ref|NP_037517.1|	lpl1919	3.00E-27	1
gi|71143127|ref|NP_056047.1|	lpl1919	6.00E-19	1
gi|4758416|ref|NP_004184.1|	lpl1919	3.00E-26	1
gi|47157332|ref|NP_004219.2|	lpl1919	1.00E-29	1
gi|8670546|ref|NP_059431.1|	lpl1919	1.00E-29	1
gi|150417986|ref|NP_006411.2|	lpl1919	2.00E-29	1
gi|4758968|ref|NP_004218.1|	lpl1919	2.00E-27	1
gi|51479145|ref|NP_006412.2|	lpl1919	2.00E-32	1
gi|14150035|ref|NP_115665.1|	lpl1919	1.00E-13	1
gi|169217073|ref|XP_001714463.1|	lpl1919	2.00E-15	1
gi|18104959|ref|NP_005774.2|	CPF_2664	9.00E-07	NA
gi|148727333|ref|NP_689540.2|	CPE2384	6.00E-10	NA
gi|169175458|ref|XP_378090.5|	CPF_1202	4.00E-06	NA
gi|10092615|ref|NP_061108.2|	CPR_0594	3.00E-06	NA
gi|119395714|ref|NP_006722.2|	CPF_0603	4.00E-14	2
gi|119395712|ref|NP_001073270.1|	CPF_0603	4.00E-14	2
gi|11386161|ref|NP_001843.1|	BA_3841	6.00E-10	3
gi|115392133|ref|NP_690850.2|	BCE_3739	3.00E-48	3
gi|156616290|ref|NP_001096078.1|	spr1403	2.00E-33	3
gi|32140760|ref|NP_116277.2|	spr1403	5.00E-16	3
gi|89363017|ref|NP_000384.2|	BCE_3739	1.00E-14	3
gi|20357512|ref|NP_619729.1|	Z2340	1.00E-07	3
gi|4505259|ref|NP_002436.1|	Z2340	2.00E-07	3
gi|20357515|ref|NP_619730.1|	Z2340	2.00E-08	3
gi|4757760|ref|NP_004788.1|	Z2147	2.00E-07	3
gi|87196339|ref|NP_001839.2|	spr1403	6.00E-13	3
gi|47778921|ref|NP_001849.2|	lpl2569	6.00E-13	3
gi|154759255|ref|NP_001032852.2|	lpl2569	2.00E-09	3
gi|61699226|ref|NP_003010.4|	spr1403	5.00E-10	3
gi|98985808|ref|NP_542197.2|	spr1403	4.00E-15	3
gi|98985806|ref|NP_001845.3|	spr1403	4.00E-15	3
gi|98985810|ref|NP_542196.2|	spr1403	4.00E-15	3
gi|65301115|ref|NP_848635.2|	spr1403	1.00E-14	3
gi|56847616|ref|NP_001007538.1|	spr1403	7.00E-15	3
gi|122937273|ref|NP_001073909.1|	spr1403	9.00E-10	3
gi|98986321|ref|NP_861454.2|	EF_2090	1.00E-06	3
gi|55743104|ref|NP_476507.2|	Z1483	3.00E-10	3
gi|55743106|ref|NP_476508.2|	Z1483	3.00E-10	3
gi|55743100|ref|NP_476505.2|	Z1483	3.00E-10	3
gi|55743102|ref|NP_476506.2|	Z1483	3.00E-10	3
gi|55743098|ref|NP_004360.2|	Z1483	3.00E-10	3
gi|115527062|ref|NP_001840.3|	SpyM3_0738	2.00E-09	3
gi|115527070|ref|NP_478055.2|	SpyM3_0738	2.00E-09	3
gi|115527066|ref|NP_478054.2|	SpyM3_0738	4.00E-09	3
gi|4502961|ref|NP_000085.1|	spr1403	2.00E-30	3
gi|111118970|ref|NP_542411.2|	spr1403	2.00E-18	3
gi|111118968|ref|NP_542410.2|	spr1403	2.00E-18	3
gi|111118972|ref|NP_542412.2|	spr1403	2.00E-18	3
gi|38570073|ref|NP_115907.2|	spr1403	5.00E-15	3
gi|38570075|ref|NP_942014.1|	spr1403	8.00E-18	3
gi|119508426|ref|NP_001844.3|	spr1403	9.00E-17	3
gi|111118974|ref|NP_149162.2|	spr1403	8.00E-14	3
gi|111118976|ref|NP_001835.3|	spr1403	6.00E-14	3
gi|110735435|ref|NP_056534.2|	spr1403	2.00E-16	3

Human proteins and bacterial protein BLAST hits used in Figure 6.1.

Human Protein	Pathogen Protein	E-value	Cluster #
gi|73486664|ref|NP_511040.2|	spr1403	9.00E-25	3
gi|73486666|ref|NP_001842.3|	spr1403	3.00E-28	3
gi|169166080|ref|XP_001717794.1|	spr1403	4.00E-11	3
gi|169165280|ref|XP_001718018.1|	spr1403	4.00E-11	3
gi|169165757|ref|XP_001718121.1|	spr1403	4.00E-11	3
gi|116256356|ref|NP_000083.3|	lpl2569	1.00E-10	3
gi|40805823|ref|NP_690848.1|	spr1403	1.00E-07	3
gi|110349772|ref|NP_000079.2|	lpl2569	2.00E-06	3
gi|89142733|ref|NP_112730.2|	spr1403	1.00E-16	3
gi|89142730|ref|NP_000082.2|	spr1403	5.00E-17	3
gi|89142735|ref|NP_112733.2|	spr1403	5.00E-16	3
gi|89142737|ref|NP_112734.2|	spr1403	1.00E-15	3
gi|55743096|ref|NP_066933.1|	spr1403	2.00E-06	3
gi|116256354|ref|NP_001837.2|	spr1403	6.00E-32	3
gi|18780273|ref|NP_110447.2|	spr1403	5.00E-28	3
gi|29725624|ref|NP_775736.2|	spr1403	6.00E-23	3
gi|148536825|ref|NP_001836.2|	spr1403	6.00E-20	3
gi|148536823|ref|NP_001838.2|	spr1403	1.00E-20	3
gi|16357503|ref|NP_378667.1|	spr1403	2.00E-20	3
gi|5803080|ref|NP_006761.1|	spr1403	7.00E-09	3
gi|22027583|ref|NP_542993.2|	spr1403	1.00E-12	3
gi|22027595|ref|NP_542998.2|	spr1403	4.00E-12	3
gi|169213233|ref|XP_001716101.1|	spr0581	9.00E-06	NA
gi|148839305|ref|NP_056110.2|	SMU_1513	2.00E-05	NA
gi|10864075|ref|NP_066929.1|	CPF_2923	2.00E-06	NA
gi|156616294|ref|NP_443122.3|	BC_3637	4.00E-11	NA
gi|16418405|ref|NP_443123.1|	pnf1380	5.00E-14	NA
gi|56550073|ref|NP_065126.2|	pnf1380	3.00E-11	NA
gi|4827036|ref|NP_005082.1|	MAV_0206	7.00E-15	NA
gi|74136743|ref|NP_003419.2|	BPSS0088	1.00E-12	NA
gi|56090146|ref|NP_001005920.2|	BMA_A0793	6.00E-06	NA
gi|169168918|ref|XP_001720060.1|	BMA_1027	3.00E-06	NA
gi|8923460|ref|NP_060316.1|	TP_0356	3.00E-05	NA
gi|20127551|ref|NP_057197.2|	TP_0835	2.00E-07	NA
gi|40068047|ref|NP_951060.1|	TDE_1243	1.00E-06	NA
gi|50083289|ref|NP_277040.1|	TDE_0021	9.00E-06	4
gi|50726996|ref|NP_064629.2|	TP_0671	5.00E-09	4
gi|56118223|ref|NP_001007795.1|	TDE_0021	6.00E-11	4
gi|5174415|ref|NP_006081.1|	TDE_0021	6.00E-11	4
gi|4502313|ref|NP_001685.1|	spyM18_0145	1.00E-07	NA
gi|169211810|ref|XP_001126648.2|	spyM18_1868	3.00E-05	5
gi|169210944|ref|XP_001126626.2|	spyM18_1868	3.00E-05	5
gi|169212041|ref|XP_001720073.1|	spyM18_1868	3.00E-05	5
gi|30425563|ref|NP_848665.1|	spyM18_1868	3.00E-14	5
gi|7019381|ref|NP_037363.1|	spyM18_1868	1.00E-10	5
gi|38202222|ref|NP_938205.1|	LA2448	6.00E-15	5
gi|7019383|ref|NP_037413.1|	LA2448	6.00E-15	5
gi|153792227|ref|NP_060804.3|	spyM18_1868	1.00E-09	5
gi|153791507|ref|NP_001093130.1|	spyM18_1868	1.00E-09	5
gi|153792651|ref|NP_001093128.1|	spyM18_1868	1.00E-09	5
gi|4758460|ref|NP_004479.1|	spyM18_1868	4.00E-18	5
gi|19743846|ref|NP_598010.1|	SpyM3_1561	9.00E-11	5
gi|4503271|ref|NP_001911.1|	SpyM3_1561	9.00E-11	5
gi|75677612|ref|NP_955372.2|	spyM18_1868	2.00E-06	6
gi|116325993|ref|NP_001006608.2|	spyM18_1868	1.00E-06	6
gi|110665701|ref|NP_055649.3|	spyM18_1868	1.00E-06	6
gi|27363458|ref|NP_076941.2|	spyM18_1868	2.00E-08	6
gi|46094076|ref|NP_056331.2|	spyM18_1868	2.00E-10	6
gi|88702793|ref|NP_612449.2|	spyM18_1868	5.00E-06	6
gi|169161015|ref|XP_001713994.1|	spyM18_1868	5.00E-06	6
gi|113411943|ref|XP_948754.2|	spyM18_1868	2.00E-08	6
gi|113403736|ref|XP_059074.5|	spyM18_1868	3.00E-08	6
gi|17511205|ref|NP_116166.6|	spyM18_1868	8.00E-08	6
gi|21281681|ref|NP_644807.1|	spyM18_1868	2.00E-05	6
gi|4826816|ref|NP_005088.1|	spyM18_1868	4.00E-05	6
gi|16418445|ref|NP_443185.1|	spyM18_1868	1.00E-08	6

Human proteins and bacterial protein BLAST hits used in Figure 6.1.

Human Protein	Pathogen Protein	E-value	Cluster #
gi|53829385|ref|NP_443120.2|	spyM18_1868	4.00E-07	6
gi|67003570|ref|NP_001019849.1|	spyM18_1868	1.00E-06	6
gi|115648142|ref|NP_055771.4|	lpl2411	2.00E-08	NA
gi|52630440|ref|NP_036313.3|	NMA0273	3.00E-05	7
gi|119703755|ref|NP_002283.3|	NMA0386	1.00E-06	7
gi|145553959|ref|NP_060597.4|	BCE_4210	6.00E-19	NA
gi|46409264|ref|NP_997192.1|	SACOL2161	6.00E-37	NA
gi|156627575|ref|NP_003106.3|	SSP0716	2.00E-54	NA
gi|87298937|ref|NP_008949.4|	lpl2411	6.00E-11	8
gi|55770834|ref|NP_057427.3|	lpl2411	2.00E-06	8
gi|4504063|ref|NP_002068.1|	lpl2411	2.00E-06	8
gi|58530840|ref|NP_004406.2|	lpl2411	2.00E-14	NA
gi|29171753|ref|NP_803172.1|	lpl2411	3.00E-08	NA
gi|4505983|ref|NP_003617.1|	lpl2411	2.00E-08	NA
gi|109659849|ref|NP_055705.2|	lpl2411	1.00E-06	NA
gi|109659845|ref|NP_878913.2|	lpl2411	5.00E-07	NA
gi|109659847|ref|NP_001035924.1|	lpl2411	6.00E-07	NA
gi|23503275|ref|NP_699178.1|	LB269	3.00E-12	NA
gi|31742492|ref|NP_060647.2|	VC_1610	2.00E-08	9
gi|7656971|ref|NP_055702.1|	VV1614	5.00E-11	9
gi|119226224|ref|NP_438169.2|	VC_1610	1.00E-15	9
gi|119226227|ref|NP_001073159.1|	VP1457	4.00E-15	9
gi|45356743|ref|NP_065115.3|	BPP4384	2.00E-08	NA
gi|110346440|ref|NP_001035977.1|	TC_0725	3.00E-06	10
gi|110346449|ref|NP_001035981.1|	TC_0725	5.00E-06	10
gi|110346445|ref|NP_001035990.1|	TC_0725	5.00E-06	10
gi|63252916|ref|NP_001017929.1|	TC_0725	3.00E-06	10
gi|110346447|ref|NP_001035980.1|	TC_0725	4.00E-06	10
gi|110346443|ref|NP_001035989.1|	TC_0725	4.00E-06	10
gi|169194806|ref|XP_001714660.1|	TC_0725	2.00E-05	10
gi|110346438|ref|NP_689641.4|	TC_0725	2.00E-05	10
gi|110346436|ref|NP_001018081.2|	TC_0725	2.00E-05	10
gi|77812678|ref|NP_008834.3|	LA4039	4.00E-08	NA
gi|40805106|ref|NP_116053.3|	jhp0596	8.00E-05	11
gi|145580617|ref|NP_775811.2|	HP0379	4.00E-06	11
gi|37594467|ref|NP_803877.2|	HP0507	1.00E-30	NA
gi|118200356|ref|NP_055209.2|	jhp0854	3.00E-09	NA
gi|14150047|ref|NP_115671.1|	lpl2474	6.00E-20	NA
gi|46361976|ref|NP_996996.1|	PA1569	3.00E-26	NA
gi|38093649|ref|NP_938018.1|	BCE_0729	2.00E-13	NA
gi|49619231|ref|NP_061837.3|	PA1569	4.00E-18	NA
gi|4758648|ref|NP_004512.1|	lpl2411	1.00E-06	NA
gi|30425526|ref|NP_848652.1|	y0742	6.00E-09	NA
gi|28631173|ref|NP_789776.1|	SSP0643	2.00E-06	NA
gi|7019551|ref|NP_037451.1|	c_4883	2.00E-15	NA
gi|169217968|ref|XP_001715197.1|	c_4365	8.00E-07	NA
gi|93141204|ref|NP_060866.2|	Rv1498c	7.00E-06	NA
gi|52856442|ref|NP_859076.2|	Rv1498c	2.00E-06	NA
gi|52856440|ref|NP_001005372.1|	Rv1498c	2.00E-06	NA
gi|58743373|ref|NP_001011713.1|	VP1850	3.00E-08	NA
gi|116875826|ref|NP_057147.2|	PA1878	1.00E-15	NA
gi|4505235|ref|NP_002426.1|	VPA1425	3.00E-52	NA
gi|31742508|ref|NP_853628.1|	ML2177	4.00E-20	12
gi|4507839|ref|NP_003355.1|	ML2177	4.00E-20	12
gi|106879206|ref|NP_078920.2|	MAV_3361	4.00E-13	NA
gi|90819239|ref|NP_060385.2|	BR_1413	8.00E-06	NA
gi|4502169|ref|NP_003896.1|	CBU_0876	5.00E-09	NA
gi|132626688|ref|NP_055456.2|	APH_0455	2.00E-11	NA
gi|6912618|ref|NP_036545.1|	PIN_A0590	2.00E-07	NA
gi|169211607|ref|XP_001725259.1|	RC0168	8.00E-06	NA
gi|4503987|ref|NP_003869.1|	RP404	6.00E-06	NA
gi|6912500|ref|NP_036459.1|	CJE_0829	5.00E-06	13
gi|29029537|ref|NP_789744.1|	CJE_0829	5.00E-06	13
gi|38045917|ref|NP_690577.2|	CJE_0829	5.00E-07	13
gi|23397648|ref|NP_116023.2|	CJE_0829	9.00E-07	13
gi|5174717|ref|NP_006044.1|	CPR_1614	4.00E-07	NA

Human proteins and bacterial protein BLAST hits used in Figure 6.1.

Human Protein	Pathogen Protein	E-value	Cluster #
gi|19924145|ref|NP_006010.2|	CPE1643	6.00E-07	NA
gi|42741679|ref|NP_036595.2|	CTC_02331	2.00E-07	NA
gi|85386056|ref|NP_570855.2|	CTC_02331	8.00E-10	NA
gi|85386547|ref|NP_570856.2|	CTC_02331	8.00E-10	NA
gi|85386053|ref|NP_065683.2|	CTC_02331	8.00E-10	NA
gi|13399322|ref|NP_078938.1|	VPA1750	1.00E-25	NA
gi|4506201|ref|NP_002788.1|	Tfu_1790	6.00E-12	NA
gi|113420096|ref|XP_001129390.1|	lmo2618	2.00E-05	NA
gi|113420404|ref|XP_001127371.1|	gbs0072	3.00E-08	NA
gi|169173286|ref|XP_001720813.1|	gbs0072	3.00E-08	NA
gi|89038393|ref|XP_370865.3|	MPN208	2.00E-07	NA
gi|169208651|ref|XP_001722684.1|	MPN208	2.00E-07	NA
gi|169208936|ref|XP_001726335.1|	MPN208	2.00E-07	NA
gi|147907441|ref|NP_001083057.1|	BC_3825	2.00E-05	NA
gi|169161807|ref|XP_001724899.1|	MG_070	3.00E-05	NA
gi|169161803|ref|XP_001724900.1|	MG_070	3.00E-05	NA
gi|169161805|ref|XP_001724897.1|	MG_070	3.00E-05	NA
gi|113408840|ref|XP_001134241.1|	BC_3825	3.00E-07	NA
gi|88951511|ref|XP_949264.1|	BC_3825	3.00E-07	NA
gi|169160827|ref|XP_001719654.1|	BC_3825	3.00E-07	NA
gi|9845502|ref|NP_002286.2|	BC_3825	1.00E-07	NA
gi|59859885|ref|NP_001012321.1|	BC_3825	1.00E-07	NA
gi|169205444|ref|XP_001724023.1|	CpB0723	3.00E-08	NA
gi|169205442|ref|XP_001724027.1|	CpB0723	3.00E-08	NA
gi|169205440|ref|XP_001724025.1|	CpB0723	3.00E-08	NA
gi|169204115|ref|XP_001726573.1|	CpB0723	3.00E-08	NA
gi|41201737|ref|XP_370697.1|	CpB0723	3.00E-08	NA
gi|169204113|ref|XP_001726574.1|	CpB0723	3.00E-08	NA
gi|169204812|ref|XP_001723476.1|	CpB0723	1.00E-07	NA
gi|169204808|ref|XP_001723478.1|	CpB0723	1.00E-07	NA
gi|169204810|ref|XP_001723479.1|	CpB0723	1.00E-07	NA
gi|46255039|ref|NP_598193.2|	NMA1658	7.00E-05	NA
gi|47419914|ref|NP_004175.2|	gbs2127	3.00E-10	NA
gi|47419916|ref|NP_776049.1|	gbs2127	3.00E-10	NA
gi|47419918|ref|NP_998810.1|	gbs2127	2.00E-10	NA
gi|47419920|ref|NP_998811.1|	gbs2127	2.00E-10	NA
gi|41327715|ref|NP_291028.3|	CpB0730	4.00E-04	NA
gi|19924133|ref|NP_002866.2|	MG_339	4.00E-05	NA
gi|7427519|ref|NP_005906.2|	STM3899	2.00E-07	NA
gi|61966919|ref|NP_001013758.1|	YPTB2659	1.00E-06	NA
gi|21361743|ref|NP_060846.2|	BP1916	2.00E-06	14
gi|9910372|ref|NP_064586.1|	BP0294	4.00E-34	14
gi|40254908|ref|NP_060303.2|	LMOh7858_1569	1.00E-05	NA
gi|93352549|ref|NP_001035282.1|	LMOh7858_1569	1.00E-05	NA
gi|50659093|ref|NP_056089.1|	CPF_2337	1.00E-07	NA
gi|18087837|ref|NP_542395.1|	CTC_02197	2.00E-08	NA
gi|7661568|ref|NP_056414.1|	CPR_2324	4.00E-10	NA
gi|153791497|ref|NP_055490.3|	lmo1224	7.00E-06	15
gi|53828920|ref|NP_001005473.1|	lmo0201	4.00E-07	15
gi|8922995|ref|NP_060860.1|	lmo0201	4.00E-12	15
gi|169217751|ref|XP_001719290.1|	lmo0201	7.00E-11	15
gi|10863957|ref|NP_066978.1|	LA3320	3.00E-06	NA
gi|47607495|ref|NP_999840.1|	LA3320	1.00E-07	NA
gi|42542390|ref|NP_055072.3|	LA3320	1.00E-07	NA
gi|16306582|ref|NP_036289.2|	LMOf2365_0495	4.00E-08	NA
gi|27734755|ref|NP_116264.2|	EF_2248	2.00E-07	NA
gi|4504851|ref|NP_003731.1|	lin2165	1.00E-05	NA
gi|157785649|ref|NP_001099129.1|	LMOh7858_0301	6.00E-06	NA
gi|20336726|ref|NP_073603.2|	lin1865	2.00E-06	NA
gi|20143971|ref|NP_006059.2|	LA2452	3.00E-06	16
gi|23097240|ref|NP_690852.1|	LA2964	2.00E-05	16
gi|5031895|ref|NP_005573.1|	LMOf2365_1254	2.00E-05	16
gi|20302168|ref|NP_619542.1|	lmo0333	5.00E-11	16
gi|19924149|ref|NP_612564.1|	spyM18_1868	7.00E-09	16
gi|38490688|ref|NP_849144.2|	LA3323	1.00E-09	NA
gi|7657419|ref|NP_055174.1|	LA3319	7.00E-07	NA

Human proteins and bacterial protein BLAST hits used in Figure 6.1.

Human Protein	Pathogen Protein	E-value	Cluster #
gi|40254971|ref|NP_079101.2|	LA3322	1.00E-11	17
gi|4507061|ref|NP_003052.1|	PG_1864	4.00E-14	17
gi|4759146|ref|NP_004778.1|	LA3320	5.00E-16	17
gi|41350214|ref|NP_060150.3|	STM0800	1.00E-12	17
gi|7706093|ref|NP_057646.1|	LA2450	8.00E-12	17
gi|5031707|ref|NP_005503.1|	PG_1864	6.00E-20	17
gi|19718734|ref|NP_003255.2|	SpyM3_1561	5.00E-05	17
gi|4505047|ref|NP_002336.1|	spyM18_1868	2.00E-10	17
gi|16751843|ref|NP_003259.2|	LA3322	8.00E-12	17
gi|71040111|ref|NP_002014.2|	spyM18_1868	3.00E-15	17
gi|4506041|ref|NP_002716.1|	SpyM3_1561	2.00E-10	17
gi|41349454|ref|NP_958505.1|	SpyM3_1561	2.00E-10	17
gi|45505137|ref|NP_714914.2|	LA2448	5.00E-15	17
gi|4557543|ref|NP_001384.1|	LA3027	6.00E-12	17
gi|4826876|ref|NP_005005.1|	LA2452	6.00E-10	17
gi|4507531|ref|NP_003256.1|	spyM18_1868	1.00E-15	17
gi|86990456|ref|NP_849161.2|	spyM18_1868	2.00E-19	17
gi|7662102|ref|NP_056379.1|	SpyM3_1561	2.00E-24	17
gi|54607118|ref|NP_056356.2|	spyM18_1868	2.00E-16	17
gi|8394456|ref|NP_059138.1|	spyM18_1868	9.00E-11	17
gi|38348406|ref|NP_940967.1|	spyM18_1868	1.00E-16	17
gi|11321571|ref|NP_003053.1|	spyM18_1868	3.00E-15	17
gi|42544231|ref|NP_006329.2|	spyM18_1868	1.00E-15	17
gi|42544233|ref|NP_963924.1|	spyM18_1868	1.00E-15	17
gi|50263044|ref|NP_116197.4|	LA3320	1.00E-16	17
gi|7662320|ref|NP_055628.1|	spyM18_1868	9.00E-15	17
gi|122937309|ref|NP_001073926.1|	spyM18_1868	5.00E-13	17
gi|15029530|ref|NP_071426.1|	spyM18_1868	8.00E-15	17
gi|85986601|ref|NP_067647.2|	spyM18_1868	2.00E-23	17
gi|12597641|ref|NP_075052.1|	SpyM3_1561	4.00E-05	17
gi|156139147|ref|NP_079269.4|	spyM18_1868	4.00E-20	17
gi|109809759|ref|NP_821079.3|	spyM18_1868	1.00E-13	17
gi|153791466|ref|NP_065754.2|	spyM18_1868	2.00E-06	17
gi|51317373|ref|NP_065980.1|	spyM18_1868	4.00E-19	17
gi|76880480|ref|NP_055632.2|	spyM18_1868	3.00E-16	17
gi|21281673|ref|NP_644813.1|	spyM18_1868	2.00E-07	17
gi|22749183|ref|NP_689783.1|	SpyM3_1561	2.00E-06	17
gi|153251229|ref|NP_001258.2|	spyM18_1868	5.00E-13	17
gi|157426829|ref|NP_001094861.1|	spyM18_1868	6.00E-07	17
gi|66912176|ref|NP_001019782.1|	LA2452	5.00E-06	17
gi|40217820|ref|NP_055741.2|	spyM18_1868	2.00E-06	17
gi|18677729|ref|NP_570718.1|	spyM18_1868	4.00E-21	17
gi|13899243|ref|NP_113615.1|	BCE_1444	1.00E-07	18
gi|157674358|ref|NP_848547.4|	PG_1864	3.00E-13	18
gi|31377705|ref|NP_078824.2|	PG_1864	7.00E-14	18
gi|4826651|ref|NP_004919.1|	LMOh7858_0499	3.00E-06	18
gi|42542396|ref|NP_964013.1|	PG_1864	3.00E-19	18
gi|5901898|ref|NP_008923.1|	PG_1864	2.00E-17	18
gi|120953300|ref|NP_001073379.1|	PG_1864	4.00E-10	18
gi|120953243|ref|NP_115541.3|	PG_1864	5.00E-10	18
gi|61966709|ref|NP_001013648.1|	lmo0433	2.00E-07	18
gi|33636689|ref|NP_112584.2|	PG_1864	4.00E-14	18
gi|169218126|ref|XP_001719886.1|	PG_1864	6.00E-07	18
gi|55743114|ref|NP_060161.2|	PG_1864	1.00E-14	18
gi|4557367|ref|NP_000377.1|	EF_0302	1.00E-99	NA
gi|4505289|ref|NP_002452.1|	lin0011	5.00E-33	NA
gi|5031751|ref|NP_005509.1|	EF_1363	2.00E-32	NA
gi|148298764|ref|NP_002121.4|	EF_1363	4.00E-27	NA
gi|148298677|ref|NP_001091742.1|	EF_1363	4.00E-27	NA
gi|29788758|ref|NP_060904.2|	gbs2008	5.00E-05	NA
gi|18426967|ref|NP_550438.1|	BCE_0015	1.00E-10	NA
gi|4758768|ref|NP_004535.1|	SSP2158	2.00E-09	NA
gi|14150078|ref|NP_115689.1|	spyM18_0302	8.00E-23	NA
gi|8922701|ref|NP_060708.1|	BA_0323	3.00E-07	NA
gi|169213669|ref|XP_001726237.1|	UU375	1.00E-13	NA
gi|169214163|ref|XP_001722837.1|	UU375	2.00E-12	NA

Human proteins and bacterial protein BLAST hits used in Figure 6.1.

Human Protein	Pathogen Protein	E-value	Cluster #
gi|46243671|ref|NP_996880.1|	SAG_2063	7.00E-09	NA
gi|147903302|ref|NP_078873.2|	SAG_2063	1.00E-08	NA
gi|24308177|ref|NP_060439.1|	BCE_0607	2.00E-08	NA
gi|20270337|ref|NP_620147.1|	ECH_0498	6.00E-08	NA
gi|19923967|ref|NP_612446.1|	BA_2967	1.00E-10	NA
gi|51702248|ref|NP_001004051.1|	BA_2967	1.00E-10	NA
gi|56676399|ref|NP_054783.2|	BC_2891	4.00E-06	NA
gi|169166991|ref|XP_001715696.1|	BCE_3139	3.00E-09	19
gi|169166369|ref|XP_001714661.1|	BCE_3139	3.00E-09	19
gi|169166783|ref|XP_001715610.1|	BCE_3139	3.00E-09	19
gi|169166531|ref|XP_001714772.1|	BCE_3139	3.00E-09	19
gi|169217984|ref|XP_001716220.1|	BCE_5203	2.00E-09	19
gi|169171568|ref|XP_001714107.1|	BCE_5203	2.00E-09	19
gi|22748777|ref|NP_689572.1|	nfa31870	5.00E-39	20
gi|23238258|ref|NP_689453.1|	nfa31870	1.00E-28	20
gi|119433675|ref|NP_065574.2|	nfa31870	6.00E-42	20
gi|119433669|ref|NP_066265.2|	nfa31870	5.00E-42	20
gi|119433673|ref|NP_066264.2|	nfa31870	5.00E-42	20
gi|119433671|ref|NP_066266.2|	nfa31870	5.00E-42	20
gi|4503023|ref|NP_000089.1|	nfa31870	6.00E-44	20
gi|31542325|ref|NP_066974.2|	nfa31870	6.00E-44	20
gi|4758050|ref|NP_004368.1|	nfa31870	9.00E-45	20
gi|23238254|ref|NP_689451.1|	nfa31870	9.00E-45	20
gi|23238256|ref|NP_689452.1|	nfa31870	9.00E-45	20
gi|21618334|ref|NP_003994.2|	nfa31870	4.00E-52	20
gi|21618331|ref|NP_000746.2|	nfa31870	5.00E-52	20
gi|73623030|ref|NP_001867.2|	nfa31870	3.00E-42	20
gi|73623028|ref|NP_001027017.1|	nfa31870	1.00E-40	20
gi|117414150|ref|NP_003264.2|	CBU_1158	2.00E-48	21
gi|153266878|ref|NP_004473.2|	Rv0696	1.00E-07	22
gi|37537527|ref|NP_079374.2|	Rv1920	1.00E-06	22
gi|51317389|ref|NP_001002254.1|	Rv1920	3.00E-05	22
gi|14149793|ref|NP_115510.1|	MT_0370	2.00E-12	22
gi|22748883|ref|NP_689630.1|	nfa38270	3.00E-14	22
gi|116686120|ref|NP_870998.2|	MT_1796	1.00E-14	23
gi|105990532|ref|NP_000375.2|	Rv1918c	1.00E-09	23

Appendix 8

Enriched GO terms in top 1% subset (no restrictions)

N	X	LOD	P	Padj	attrib ID	attrib name
4	4	2.72	7.94E-08	0	GO:0004102	choline O-acetyltransferase activity
4	4	2.72	7.94E-08	0	GO:0008745	N-acetylmuramoyl-L-alanine amidase activity
4	4	2.72	7.94E-08	0	GO:0032648	regulation of interferon-beta production
4	4	2.72	7.94E-08	0	GO:0045356	positive regulation of interferon-alpha biosynthetic process
4	4	2.72	7.94E-08	0	GO:0045357	regulation of interferon-beta biosynthetic process
4	4	2.72	7.94E-08	0	GO:0045359	positive regulation of interferon-beta biosynthetic process
10	11	2.62	1.78E-17	0	GO:0016406	carnitine O-acyltransferase activity
3	3	2.61	4.75E-06	0.014	GO:0004421	hydroxymethylglutaryl-CoA synthase activity
3	3	2.61	4.75E-06	0.014	GO:0005594	collagen type IX
8	9	2.53	5.36E-14	0	GO:0005587	collagen type IV
7	8	2.47	2.88E-12	0	GO:0004095	carnitine O-palmitoyltransferase activity
7	8	2.47	2.88E-12	0	GO:0016416	O-palmitoyltransferase activity
5	6	2.34	7.84E-09	0	GO:0006853	carnitine shuttle
5	6	2.34	7.84E-09	0	GO:0030020	ECM structural constituent conferring tensile strength
7	9	2.25	1.28E-11	0	GO:0030934	anchoring collagen
4	5	2.25	3.92E-07	0.001	GO:0004830	tryptophan-tRNA ligase activity
4	5	2.25	3.92E-07	0.001	GO:0006436	tryptophanyl-tRNA aminoacylation
4	5	2.25	3.92E-07	0.001	GO:0007529	establishment of synaptic specificity at neuromuscular junction
4	5	2.25	3.92E-07	0.001	GO:0043179	rhythmic excitation
4	5	2.25	3.92E-07	0.001	GO:0045354	regulation of interferon-alpha biosynthetic process
8	11	2.16	9.54E-13	0	GO:0030935	sheet-forming collagen
7	10	2.10	4.20E-11	0	GO:0005583	fibrillar collagen
26	40	2.06	5.98E-37	0	GO:0005581	collagen
6	9	2.04	1.77E-09	0	GO:0016413	O-acetyltransferase activity
4	6	2.03	1.16E-06	0.002	GO:0005593	FACIT collagen
4	6	2.03	1.16E-06	0.002	GO:0016019	peptidoglycan receptor activity
4	6	2.03	1.16E-06	0.002	GO:0060024	rhythmic synaptic transmission
15	26	1.91	1.23E-20	0	GO:0043022	ribosome binding
11	19	1.91	1.80E-15	0	GO:0032012	regulation of ARF protein signal transduction
4	7	1.88	2.67E-06	0.003	GO:0000270	peptidoglycan metabolic process
4	7	1.88	2.67E-06	0.003	GO:0009253	peptidoglycan catabolic process
4	7	1.88	2.67E-06	0.003	GO:0032479	regulation of type I interferon production
4	7	1.88	2.67E-06	0.003	GO:0032647	regulation of interferon-alpha production
11	20	1.86	3.94E-15	0	GO:0005086	ARF guanyl-nucleotide exchange factor activity
5	9	1.86	1.58E-07	0.001	GO:0015879	carnitine transport
5	10	1.77	3.11E-07	0.001	GO:0032365	intracellular lipid transport
5	11	1.70	5.63E-07	0.001	GO:0042834	peptidoglycan binding
4	9	1.68	9.36E-06	0.024	GO:0016045	detection of bacterium
4	9	1.68	9.36E-06	0.024	GO:0017169	CDP-alcohol phosphatidyltransferase activity
7	17	1.63	6.14E-09	0	GO:0016409	palmitoyltransferase activity
6	15	1.61	9.69E-08	0	GO:0008329	pattern recognition receptor activity
7	19	1.55	1.55E-08	0	GO:0009595	detection of biotic stimulus
15	45	1.49	4.12E-16	0	GO:0043021	ribonucleoprotein binding
27	85	1.47	2.28E-27	0	GO:0005201	extracellular matrix structural constituent

Enriched GO terms in top 1% subset (no restrictions)

N	X	LOD	P	Padj	attrib ID	attrib name
16	51	1.45	1.28E-16	0	GO:0008374	O-acyltransferase activity
5	17	1.41	6.94E-06	0.02	GO:0015909	long-chain fatty acid transport
9	31	1.40	1.45E-09	0	GO:0030199	collagen fibril organization
7	26	1.36	1.82E-07	0.001	GO:0006635	fatty acid beta-oxidation
6	23	1.34	1.74E-06	0.002	GO:0015908	fatty acid transport
6	24	1.32	2.29E-06	0.002	GO:0016780	phosphotransferase activity, for other substituted phosphate groups
28	132	1.23	5.74E-23	0	GO:0044420	extracellular matrix part
7	34	1.21	1.33E-06	0.002	GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain
7	37	1.16	2.43E-06	0.002	GO:0009062	fatty acid catabolic process
7	39	1.14	3.53E-06	0.004	GO:0019395	fatty acid oxidation
7	39	1.14	3.53E-06	0.004	GO:0034440	lipid oxidation
7	47	1.04	1.29E-05	0.031	GO:0016054	organic acid catabolic process
7	47	1.04	1.29E-05	0.031	GO:0046395	carboxylic acid catabolic process
13	97	0.98	9.32E-09	0	GO:0030198	extracellular matrix organization
10	81	0.94	1.06E-06	0.001	GO:0016407	acetyltransferase activity
8	67	0.93	1.65E-05	0.036	GO:0030258	lipid modification
17	147	0.91	5.07E-10	0	GO:0043062	extracellular structure organization
26	263	0.84	4.94E-13	0	GO:0005578	proteinaceous extracellular matrix
23	235	0.83	1.39E-11	0	GO:0016746	transferase activity, transferring acyl groups
15	157	0.82	7.21E-08	0	GO:0022627	cytosolic small ribosomal subunit
13	138	0.81	6.34E-07	0.001	GO:0045087	innate immune response
20	220	0.80	1.13E-09	0	GO:0008415	acyltransferase activity
26	290	0.79	4.69E-12	0	GO:0031012	extracellular matrix
20	223	0.79	1.43E-09	0	GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups
11	135	0.74	1.86E-05	0.038	GO:0046578	regulation of Ras protein signal transduction
15	185	0.74	6.15E-07	0.001	GO:0015935	small ribosomal subunit
50	810	0.64	1.88E-15	0	GO:0007155	cell adhesion
50	810	0.64	1.88E-15	0	GO:0022610	biological adhesion
50	923	0.57	2.92E-13	0	GO:0005615	extracellular space
57	1067	0.57	9.11E-15	0	GO:0044421	extracellular region part
48	927	0.55	4.63E-12	0	GO:0005198	structural molecule activity
65	2049	0.33	4.42E-07	0.001	GO:0005576	extracellular region

Enriched GO terms in top 1% subset (restricted to human proteins with unique top matches in pathogen genomes).

N	X	LOD	P	Padj	attrib ID	attrib name
2	2	2.888710354	4.18E-05	0.0774	GO:0001729	ceramide kinase activity
3	4	2.560866776	1.06E-06	0.0015	GO:0008745	N-acetylmuramoyl-L-alanine amidase activity
3	6	2.192849176	5.27E-06	0.0069	GO:0016019	peptidoglycan receptor activity
3	7	2.083684298	9.18E-06	0.0128	GO:0000270	peptidoglycan metabolic process
3	7	2.083684298	9.18E-06	0.0128	GO:0009253	peptidoglycan catabolic process
3	9	1.923942636	2.18E-05	0.0258	GO:0016045	detection of bacterium
3	11	1.807396243	4.25E-05	0.0827	GO:0042834	peptidoglycan binding

Enriched GO terms in top 1% subset (restricted to human proteins with unique top matches in pathogen genomes).

N	X	LOD	P	Padj	attrib ID	attrib name
4	19	1.658678382	6.10E-06	0.0076	GO:0009595	detection of biotic stimulus
3	15	1.639823496	0.000114886	0.1389	GO:0008329	pattern recognition receptor activity
4	34	1.364403839	6.77E-05	0.1012	GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain
9	263	0.779254254	5.65E-05	0.0951	GO:0005578	proteinaceous extracellular matrix
9	290	0.734905688	0.000119052	0.1418	GO:0031012	extracellular matrix
29	1989	0.420178683	2.59E-05	0.0285	GO:0016740	transferase activity

Enriched GO terms in top 1% subset (restricted to proteins with top reciprocal matches in pathogen genomes).

N	X	LOD	P	Padj	attrib ID	attrib name
2	2	3.173746432	1.15E-05	0.027	GO:0001729	ceramide kinase activity
3	4	2.84884978	1.52E-07	0	GO:0008745	N-acetylmuramoyl-L-alanine amidase activity
3	6	2.480832278	7.55E-07	0.001	GO:0016019	peptidoglycan receptor activity
3	7	2.371667449	1.32E-06	0.002	GO:0000270	peptidoglycan metabolic process
3	7	2.371667449	1.32E-06	0.002	GO:0009253	peptidoglycan catabolic process
3	9	2.211925884	3.15E-06	0.006	GO:0016045	detection of bacterium
3	11	2.095379589	6.16E-06	0.011	GO:0042834	peptidoglycan binding
3	15	1.927807037	1.68E-05	0.031	GO:0008329	pattern recognition receptor activity
3	17	1.863308311	2.50E-05	0.042	GO:0050830	defense response to Gram-positive bacterium
3	19	1.807151627	3.54E-05	0.064	GO:0009595	detection of biotic stimulus
4	81	1.249438194	0.000169774	0.181	GO:0016407	acetyltransferase activity
6	235	0.947222539	0.000151275	0.176	GO:0016746	transferase activity, transferring acyl groups
20	1989	0.576072808	7.19E-06	0.011	GO:0016740	transferase activity
38	6080	0.438494207	1.72E-05	0.031	GO:0003824	catalytic activity

Permissions to reproduce articles

Article: Doxey AC, Yaish MW, Griffith M, McConkey BJ. (2006) Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nat. Biotechnol.* 24:852-5.

This is a License Agreement between Andrew C Doxey ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

License Number 2396700187820

License date Mar 26, 2010

Licensed content publisher Nature Publishing Group

Licensed content publication Nature Biotechnology

Licensed content title Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions

Licensed content author Andrew C Doxey, Mahmoud W Yaish, Marilyn Griffith, Brendan J McConkey

Year of publication 2006

Portion used Full paper

Requestor type Student

Type of Use Thesis / Dissertation

Company Andrew C Doxey

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use, subject to the conditions below:

NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does

not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version. Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run). NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

The credit line should read:

Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

For AOP papers, the credit line should read:

Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

Translations of 401 words up to a whole article require NPG approval. Please visit :

http://www.macmillanmedicalcommunications.com for more information. Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Article: Doxey AC, Yaish MW, Moffatt BA, Griffith M, McConkey BJ. (2007) Functional divergence in the *Arabidopsis* β-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states. *Mol. Biol. Evol.* 24:1045-55.

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oxfordjournals.org

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

Article: Doxey AC, Lynch MD, Müller KM, Meiering EM, McConkey BJ. (2008b) Insights into the evolutionary origins of clostridial neurotoxins from analysis of the *Clostridium botulinum* strain A neurotoxin gene cluster. *BMC Evol. Biol.* 8:316.

From: david.roman@biomedcentral.com

Subject: 00320081: thesis and copyright

Date: December 14, 2009 10:38:29 AM GMT-05:00

To: acdoxey@sciborg.uwaterloo.ca

Dear Andrew

Thank you for contacting BioMed Central.

BMC Evolutionary Biology is an open access journal therefore you are free to include the publication as requested. Please ensure you attribute BioMed Central as the original publisher.

If you have any questions please don't hesitate to contact me.

Best wishes

David Roman

david.roman@biomedcentral.com

www.biomedcentral.com

Bibliography

1. Aarts MG, et al. (1997) The *Arabidopsis* MALE STERILITY 2 protein shares similarity with reductases in elongation/condensation complexes. *Plant J.* 12:615-23.

2. Abascal F, Zardoya R, Posada D. (2005) ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* 21:2104-5.

3. Abhiman S, Sonnhammer EL. (2005) FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.* 33:D197-200.

4. Abraham AL, Pothier J, Rocha EP. (2009) Alternative to homo-oligomerisation: the creation of local symmetry in proteins by internal amplification. *J. Mol. Biol.* 394:522-34.

5. Albert LJ, Inman RD. (1999) Molecular mimicry and autoimmunity. *N. Engl. J. Med.* 341:2068-74.

6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-10.

7. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

8. Andersen-Nissen E, et al. (2005) Evasion of Toll-like receptor 5 by flagellated bacteria. *Proc. Natl. Acad. Sci. U. S. A.*102:9247-52.

9. Andrade MA, Perez-Iratxeta C, Ponting CP. (2001) Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* 134:117-31.

10. Anisimova M, Gascuel O. (2006) Approximate likelihood ratio test for branchs: A fast, accurate and powerful alternative. *Syst. Biol.* 55:539-52.

11. Antson AA, et al. (2001) Understanding the mechanism of ice binding by type III antifreeze proteins. *J. Mol. Biol.* 305:875-89.

12. Antuch W, et al. (1994) NMR solution structure of the recombinant tick anticoagulant protein (rTAP), a factor Xa inhibitor from the tick *Ornithodoros moubata*. *FEBS Lett.* 352:251-7.

13. Apweiler R, et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 29:37-40.

14. Arndt JW, et al. (2005) The structure of the neurotoxin-associated protein HA33/A from *Clostridium botulinum* suggests a reoccurring β-trefoil fold in the progenitor toxin complex. *J. Mol. Biol.* 346:1083-93.

15. Arora SK, Neely AN, Blair B, Lory S, Ramphal R. (2005) Role of motility and flagellin glycosylation in the pathogenesis of *Pseudomonas aeruginosa* burn wound infections. *Infect. Immun.* 73:4395-8.

16. Arora SK, Ritchings BW, Almira EC, Lory S, Ramphal R. (1996) Cloning and characterization of *Pseudomonas aeruginosa* fliF, necessary for flagellar assembly and bacterial adherence to mucin. *Infect. Immun.* 64:2130-6.

17. Ashburner M, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25:25-9.

18. Ausiello G, et al. (2008) FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics* 9 Suppl 2:S2.

19. Bachman ES, McClay DR. (1996) Molecular cloning of the first metazoan β-1,3 glucanase from eggs of the sea urchin *Strongylocentrotus purpuratus*. *Proc. Natl. Acad. Sci. U. S. A.* 93:6808-13.

20. Bagley SC, Altman RB. (1995) Characterizing the microenvironment surrounding protein sites. *Protein Sci.* 4:622-35.

21. Barral P, et al. (2005) An olive pollen protein with allergenic activity, Ole e 10, defines a novel family of carbohydrate-binding modules and is potentially implicated in pollen germination. *Biochem. J.* 390:77-84.

22. Beatson SA, Minamino T, Pallen MJ. (2006) Variation in bacterial flagellins: from sequence to structure. *Trends Microbiol.* 14:151-5.

23. Bejerano G, Haussler D, Blanchette M. (2004) Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics* 20 Suppl 1:i40-8.

24. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res.* 28:235-42.

25. Berriz GF, King OD, Bryant B, Sander C, Roth FP. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics* 19:2502-4.

26. Blanc G, Wolfe KH. (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16:1679-91.

27. Blasi J, et al. (1993) Botulinum neurotoxin A selectively cleaves the synaptic protein SNAP-25. *Nature* 365:160-3.

28. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* 382:769-81.

29. Borner GH, Lilley KS, Stevens TJ, Dupree P. (2003) Identification of glycosylphosphatidylinositol-anchored proteins in *Arabidopsis*. A proteomic and genomic analysis. *Plant Physiol.* 132:568-77.

30. Borner GH, Sherrier DJ, Stevens TJ, Arkin IT, Dupree P. (2002) Prediction of glycosylphosphatidylinositol-anchored proteins in *Arabidopsis*. A genomic analysis. *Plant Physiol.* 129:486-99.

31. Bosch TC. (2007) Symmetry breaking in stem cells of the basal metazoan *Hydra*. *Prog. Mol. Subcell. Biol.* 45:61-78.

32. Boyes DC, et al. (2001) Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell* 13:1499-1510.

33. Bravo LA, Griffith M. (2005) Characterization of antifreeze activity in Antarctic plants. *J. Exp. Bot.* 56:1189-96.

34. Bray MR, et al. (1996) Probing the role of tryptophan residues in a cellulose-binding domain by chemical modification. *Protein Sci.* 5:2311-8.

35. Brown D, Sjölander K. (2006) Functional classification using phylogenomic inference. *PLoS Comput. Biol.* 2:e77.

36. Buell CR, Somerville SC. (1997) Use of *Arabidopsis* recombinant inbred lines reveals a monogenic and a novel digenic resistance mechanism to *Xanthomonas campestris* pv campestris. *Plant J.* 12:21-9.

37. Burgoyne NJ, Jackson RM. (2006) Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics* 22:1335-42.

38. Campbell P, Braam J. (1999) Xyloglucan endotransglycosylases: diversity of genes, enzymes and potential wall-modifying functions. *Trends Plant Sci.* 4:361-6.

39. Cantarel BL, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37:D233-8.

40. Cao W, et al. (1998) Identification of α-dystroglycan as a receptor for lymphocytic choriomeningitis virus and Lassa fever virus. *Science* 282:2079-81.

41. Carvalho Ade O, Gomes VM. (2007) Role of plant lipid transfer proteins in plant cell physiology-a concise review. *Peptides* 28:1144-53.

42. Casari G, Sander C, Valencia A. (1995) A method to predict functional residues in proteins. *Nat Struct Biol.* 2:171-8.

43. Chai Q, et al. (2006) Structural basis of cell surface receptor recognition by botulinum neurotoxin B. *Nature* 444:1096-100.

44. Chao H, Sönnichsen FD, DeLuca CI, Sykes BD, Davies PL. (1994) Structure-function relationship in the globular type III antifreeze protein: identification of a cluster of surface residues required for binding to ice. *Protein Sci.* 3:1760-9.

45. Chapman JA, et al. (2010) The dynamic genome of *Hydra*. *Nature* [Epub ahead of print].

46. Chaudhuri I, Söding J, Lupas AN (2008) Evolution of the β-propeller fold. *Proteins* 71:795-803.

47. Chen L, et al. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* 33:D325-8.

48. Chenna R, et al. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31:3497-500.

49. Chothia C. (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357:543-4.

50. Collins MD, East AK. (1998) Phylogeny and taxonomy of the food-borne pathogen *Clostridium botulinum* and its neurotoxins. *J. Appl. Microbiol.* 84:5-17.

51. Crooks GE, Hon G, Chandonia JM, Brenner SE. (2004) WebLogo: A sequence logo generator. *Genome Res.* 14:1188-90.

52. Davey NE, Shields DC, Edwards RJ. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res.* 34:3546-54.

53. Dawkins R, Krebs JR. (1979) Arms races between and within species. *Proc R Soc Lond B Biol Sci.* 205:489-511.

54. de Almeida Marques-da-Silva E, et al. (2008) Extracellular nucleotide metabolism in Leishmania: influence of adenosine in the establishment of infection. *Microbes Infect.* 10:850-7.

55. de Maagd RA, Bravo A, Berry C, Crickmore N, Schnepf HE. (2003) Structure, diversity, and evolution of protein toxins from spore-forming entomopathogenic bacteria. *Annu. Rev. Genet.* 37:409-33.

56. De Souza SJ, Pereira HM, Jacchieri S, Brentani RR. (1996) Collagen/collagenase interaction: does the enzyme mimic the conformation of its own substrate? *FASEB J.* 10:927-30.

57. Deb C, et al. (2009) A novel in vitro multiple-stress dormancy model for *Mycobacterium tuberculosis* generates a lipid-loaded, drug-tolerant, dormant pathogen. *PLoS One.* 4:e6077.

58. DeLuca CI, Davies PL, Ye Q, Jia Z. (1998) The effects of steric mutations on the structure of type III antifreeze protein and its interaction with ice. *J. Mol. Biol.* 275:515-25.

59. DeVries AL. (1986) Antifreeze glycopeptides and peptides: interactions with ice and water. *Methods Enzymol.* 127:293-303.

60. Delp G, Palva ET. (1999) A novel flower-specific *Arabidopsis* gene related to both pathogen-induced and developmentally regulated plant β-1,3-glucanase genes. *Plant Mol. Biol.* 39:565-75.

61. Dineen SS, Bradshaw M, Johnson EA: Neurotoxin gene clusters in *Clostridium botulinum* type A strains. (2003) sequence comparison and evolutionary implications. *Curr. Microbiol.* 46:345-52.

62. Dong M, Yeh F, Tepp WH, Dean C, Johnson EA, Janz R, Chapman ER. (2006) SV2 is the protein receptor for botulinum neurotoxin A. *Science* 312:592-6.

63. Dong X, Mindrinos M, Davis KR, Ausubel FM. (1991) Induction of *Arabidopsis* defense genes by virulent and avirulent *Pseudomonas syringae* strains and by a cloned avirulence gene. *Plant Cell* 3:61-72.

64. Doolittle RF. (1994) Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* 19:15-8.

65. Doxey AC, Yaish MW, Griffith M, McConkey BJ. (2006) Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nat. Biotechnol.* 24:852-5.

66. Doxey AC, Yaish MW, Moffatt BA, Griffith M, McConkey BJ. (2007) Functional divergence in the *Arabidopsis* β-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states. *Mol. Biol. Evol.* 24:1045-55.

67. Doxey AC, Cheng Z, McConkey BJ. (2008a) Discrimination of insoluble-carbohydrate binding proteins and their binding sites using a 3D motif detection method. In *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, pp 207-13.

68. Doxey AC, Lynch MD, Müller KM, Meiering EM, McConkey BJ. (2008b) Insights into the evolutionary origins of clostridial neurotoxins from analysis of the *Clostridium botulinum* strain A neurotoxin gene cluster. *BMC Evol. Biol.* 8:316.

69. Duarte JM, et al. (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol. Biol. Evol.* 23:469-78.

70. Duman J, DeVries A. (1974) Freezing resistance in winter flounder *Pseudopleuronectes americanus*. *Nature* 247:237-8.

71. Duman JG, Serianni AS. (2002) The role of endogenous antifreeze protein enhancers in the hemolymph thermal hysteresis activity of the beetle *Dendroides canadensis*. *J Insect Physiol.* 48:103-111.

72. D'haeseleer P. (2005) How does gene expression clustering work? *Nat. Biotechnol.* 23:1499-1501.

73. East AK, Bhandari M, Stacey JM, Campbell KD, Collins MD. (1996) Organization and phylogenetic interrelationships of genes encoding components of the botulinum toxin complex in proteolytic *Clostridium botulinum* types A, B, and F: evidence of chimeric sequences in the gene encoding the nontoxic nonhemagglutinin component. *Int. J. Syst. Bacteriol.* 46:1105-12.

74. Eberhard T, Virkola R, Korhonen T, Kronvall G, Ullberg M. (1998) Binding to human extracellular matrix by *Neisseria meningitidis. Infect. Immun.* 66:1791-4.

75. Eddy SR. (1998) Profile hidden Markov models. *Bioinformatics* 14:755-63.

76. Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792-7.

77. Ehrt S, Schnappinger D. (2009) *Mycobacterium tuberculosis* virulence: lipids inside and out. *Nat. Med.* 13:284-5.

78. Eisen JA. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8:163-7.

79. Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95:14863-8.

80. Elder JF, Turner BJ. (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. *Q. Rev. Biol.* 70:297-320.

81. Elortza F, et al. (2003) Proteomic analysis of glycosylphosphatidylinositol-anchored membrane proteins. Mol. Cell. *Proteomics* 2: 1261-70.

82. Endersby, J. (2009) Lumpers and splitters: Darwin, Hooker, and the search for order. *Science* 326:1496-9.

83. Ernst D, Schraudner M, Langebartels C, Sandermann H Jr. (1992) Ozone-induced changes of mRNA levels of β-1,3-glucanase, chitinase and 'pathogenesis-related' protein 1b in tobacco plants. *Plant Mol. Biol.* 20:673-82.

84. Fankhauser N, Mäser P. (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* 21:1846-52.

85. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R. (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins* 58:134-43.

86. Ferrier DE, Minguillón C. (2003) Evolution of the Hox/ParaHox gene clusters. *Int J Dev Biol.* 47:605-11.

87. Finlay BB, Cossart P. (1997) Exploitation of mammalian host cell functions by bacterial pathogens. *Science* 276:718-25.

88. Finn RD, et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.* 34:D247-51.

89. Fischer D, et al. (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins* 3:209-17.

90. Frankel G, Newton SM, Schoolnik GK, Stocker BA. (1989) Intragenic recombination in a flagellin gene: characterization of the H1-j gene of *Salmonella typhi. EMBO J.* 8:3149-52.

91. Fraser HB, Moses AM, Schadt EE. (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc. Natl. Acad. Sci. U. S. A.* 107:2977-82.

92. Fraune S, Bosch TCG. (2007) Long-term maintenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. *Proc. Natl. Acad. Sci. U. S. A.* 104:13146-51.

93. Freitag NE, Port GC, Miner MD. (2009) *Listeria monocytogenes* - from saprophyte to intracellular pathogen. *Nat. Rev. Microbiol.* 7:623-8.

94. Fulcher R, McCully M, Setterfield G, Sutherland J. (1976) β-1,3-glucans may be associated with cell-plate formation during cytokinesis. *Can. J. Bot.* 54:539-42.

95. Gaillard JL, Berche P, Frehel C, Gouin E, Cossart P. (1991) Entry of *L. monocytogenes* into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. *Cell* 65:1127-41.

96. Gallagher KR, Sharp KA. (2003) Analysis of thermal hysteresis protein hydration using the random network model. *Biophys. Chem.* 105:195-209.

97. Galtier N, Gouy M, Gautier C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* 12:543-8.

98. Gardel, CL, Mekalanos JJ. (1996) Alterations in *Vibrio cholerae* motility phenotypes correlate with changes in virulence factor expression. *Infect. Immun.* 64:2246-55.

99. Gaucher EA, Gu X, Miyamoto MM, Benner SA. (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* 27:315-21.

100. Gauthier SY, Kay CM, Sykes BD, Walker VK, Davies PL. (1998) Disulfide bond mapping and structural characterization of spruce budworm antifreeze protein. *Eur. J. Biochem.* 258:445-53.

101. Gibrat JF, Madej T, Bryant SH. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* 6:377-85.

102. Glaser F, et al. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19:163-4.

103. Graether SP, Sykes BD. (2004) Cold survival in freeze-intolerant insects: the structure and function of β-helical antifreeze proteins. *Eur. J. Biochem.* 271:3285-96.

104. Graether SP, et al. (1999) Quantitative and qualitative analysis of type III antifreeze protein structure and function. *J. Biol. Chem.* 274:11842-7.

105. Grahn E, et al. (2007) Crystal structure of the *Marasmius oreades* mushroom lectin in complex with a xenotransplantation epitope. *J. Mol. Biol.* 369:710-21.

106. Griffith M, Yaish MW. (2004) Antifreeze proteins in overwintering plants: a tale of two activities. *Trends Plant Sci.* 9:399-405.

107. Gu X, Su Z. (2007) Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc. Natl. Acad. Sci. U. S. A.* 104:2779-84.

108. Gu X, Vander Velden K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* 18:500-1.

109. Gu X, Zhang Z, Huang W. (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Natl. Acad. Sci. U. S. A.* 102:707-12.

110. Gu X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16:1664-74.

111. Guerry P, et al. (2006) Changes in flagellin glycosylation affect Campylobacter autoagglutination and virulence. *Mol. Microbiol.* 60:299-311.

112. Guex N, Peitsch MC. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18:2714-23.

113. Guindon S, Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696-704.

114. Gulbins E, Dreschers S, Wilker B, Grassmé H. (2004) Ceramide, membrane rafts and infections. *J. Mol. Med.* 82:357-63.

115. Gupta R, Ting JT, Sokolov LN, Johnson SA, Luan S. (2002) A tumor suppressor homolog, AtPTEN1, is essential for pollen development in *Arabidopsis*. *Plant Cell* 14:2495-507.

116. Gutteridge A, Bartlett GJ, Thornton JM. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* 330:719-34.

117. Ha M, Kim ED, Chen ZJ. (2009) Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc. Natl. Acad. Sci. U. S. A*. 106:2295-300.

118. Habazettl J, et al. (1992) Structure of hisactophilin is similar to interleukin-1 β and fibroblast growth factor. *Nature* 359:855-8.

119. Hanson MA, Stevens RC. (2000) Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 A resolution. *Nat. Struct. Biol.* 7:687-92.

120. Harding MM, Ward LG, Haymet AD. (1999) Type I 'antifreeze' proteins. Structure-activity studies and mechanisms of ice growth inhibition. *Eur. J. Biochem.* 264:653-65.

121. Harrington CS, Thomson-Carter FM, Carter PE. (1997) Evidence for recombination in the flagellin locus of *Campylobacter jejuni*: implications for the flagellin gene typing scheme. *J. Clin. Microbiol.* 35:2386-92.

122. Harrington DJ. (1996) Bacterial collagenases and collagen-degrading enzymes and their potential role in human disease. *Infect. Immun.* 64:1885-9.

123. Hatheway CL. (1990) Toxigenic clostridia. *Clin. Microbiol. Rev.* 3:66-98.

124. Hayashi F, et al. (2001) The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* 410:1099-103.

125. Hazes B. (1996) The (QxW)3 domain: a flexible lectin scaffold. *Protein Sci.* 5:1490-501.

126. He X, Zhang J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157-64.

127. Heger A, Holm L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* 41:224-37.

128. Hennig L, Menges M, Murray JA, Gruissem W. (2003) *Arabidopsis* transcript profiling on Affymetrix GeneChip arrays. *Plant Mol. Biol.* 53:457-65.

129. Henrissat B, Davies GJ. (2000) Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiol.* 124:1515-9.

130. Hiilovaara-Teijo M, Hannukkala A, Griffith M, Yu XM, Pihakaski-Maunsbach K. (1999) Snow-mold-induced apoplastic proteins in winter rye leaves lack antifreeze activity. *Plant Physiol.* 121:665-74.

131. Hill KK, et al. (2007) Genetic diversity among Botulinum Neurotoxin-producing clostridial strains. *J. Bacteriol.* 189:818-32.

132. Hill KK, et al. (2009) Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and *Clostridium butyricum* type E strains. *BMC Biol.* 7:66.

133. Hincha DK, et al. (2001) Cabbage cryoprotectin is a member of the nonspecific plant lipid transfer protein gene family. *Plant Physiol.* 125:835-46.

134. Hines HB, Lebeda F, Hale M, Brueggemann EE. (2005) Characterization of botulinum progenitor toxins by mass spectrometry. *Appl. Environ. Microbiol.* 71:4478-86.

135. Hird DL, et al. (1993) The anther-specific protein encoded by the *Brassica napus* and *Arabidopsis thaliana* A6 gene displays similarity to β-1,3-glucanases. *Plant J.* 4:1023-33.

136. Hoffmeister S, Schaller HC. (1985) A new biochemical marker for foot-specific cell differentiation in *Hydra*. *Dev. Genes Evol.* 194:453-61.

137. Hoffmeister-Ullerich SA, Herrmann D, Kielholz J, Schweizer M, Schaller HC. (2002) Isolation of a putative peroxidase, a target for factors controlling foot-formation in the coelenterate *Hydra*. *Eur. J. Biochem.* 269:4597-606.

138. Hoj PB and Fincher GB. (1995) Molecular evolution of plant β-glucan endohydrolases. *Plant J.* 7:367-79.

139. Holm L, Sander C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* 20:478-80.

140. Hon WC, Griffith M, Chong P, Yang D. (1994) Extraction and isolation of antifreeze proteins from winter rye (*Secale cereale* L.) leaves. *Plant Physiol.* 104:971-80.

141. Hon WC, Griffith M, Mlynarz A, Kwok YC, Yang DS. (1995) Antifreeze proteins in winter rye are similar to pathogenesis-related proteins. *Plant Physiol.* 109:879-89.

142. Hooft RW, Vriend G, Sander C, Abola EE. (1996) Errors in protein structures. *Nature* 381:272.

143. Howe D, Heinzen RA. (2006) *Coxiella burnetii* inhabits a cholesterol-rich vacuole and influences cellular cholesterol metabolism. *Cell Microbiol.* 8:496-507.

144. Huecas S, Villalba M, Rodriguez R. (2001) Ole e 9, a major olive pollen allergen is a 1,3-β-glucanase. Isolation, characterization, amino acid sequence, and tissue specificity. *J. Biol. Chem.* 276:27959-66.

145. Hughes AL. (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci.* 256:119-24.

146. Hughes AL. (2005) Gene duplication and the origin of novel proteins. *Proc. Natl. Acad. Sci. U. S. A.* 102:8791-2.

147. Hughes TR, et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102:109-26.

148. Huiskonen JT, Manole V, Butcher SJ. (2007) Tale of two spikes in bacteriophage PRD1. *Proc. Natl. Acad. Sci. U. S. A.* 104:6666-71.

149. Hulo N, et al. (2006) The PROSITE database. *Nucleic Acids Res.* 34:D227-30.

150. Hurst LD. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486.

151. Huynen M, Snel B, Lathe W, Bork P. (2000) Exploitation of gene context. *Curr. Opin. Struct. Biol.* 10:366-70.

152. Ichinose Y, et al. (2008) Role of flagellin glycosylation in bacterial virulence. In *Pseudomonas syringae* pathovars and related pathogens — Identification, epidemiology and genomics. Springer, Netherlands.

153. Ingmundson A, Delprato A, Lambright DG, Roy CR. (2007) *Legionella pneumophila* proteins that regulate Rab1 membrane cycling. *Nature* 450:365-9.

154. Ivanova N, et al. (2003) Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature* 423:87-91.

155. Jach G, et al. (1995) Enhanced quantitative resistance against fungal disease by combinatorial expression of different barley antifungal proteins in transgenic tobacco. *Plant J.* 8:97-109.

156. Jacobson MJ, Lin G, Raphael B, Andreadis J, Johnson EA. (2008) Analysis of neurotoxin cluster genes in *Clostridium botulinum* strains producing botulinum neurotoxin serotype A subtypes. *Appl. Environ. Microbiol.* 74:2778-86.

157. Jia Z, Davies PL. (2002) Antifreeze proteins: an unusual receptor-ligand interaction. *Trends Biochem. Sci.* 27:101-6.

158. Jia Z, DeLuca CI, Chao H, Davies PL. (1996) Structural basis for the binding of a globular antifreeze protein to ice. *Nature* 384:285-8.

159. Jin R, Rummel A, Binz T, Brunger AT. (2006) Botulinum neurotoxin B recognizes its protein receptor with high affinity and specificity. *Nature* 444:1092-5.

160. Jin W, Horner HT, Palmer RG, Shoemaker RC. (1999) Analysis and mapping of gene families encoding β-1,3-glucanases of soybean. *Genetics* 153:445-52.

161. Jones S, Thornton JM. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* 272:133-43.

162. Jones S, Thornton JM. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* 8:3-7.

163. Jorov A, Zhorov BS, Yang DS. (2004) Theoretical study of interaction of winter flounder antifreeze protein with ice. *Protein Sci.* 13:1524-37.

164. Karp PD. (1998) What we do not know about sequence analysis and sequence databases. *Bioinformatics* 14:753-4.

165. Kelley LA, Sternberg MJ. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* 4:363-71.

166. Kelley LA, Sutcliffe MJ. (1997) OLDERADO: on-line database of ensemble representatives and domains. On Line Database of Ensemble Representatives And DOmains. *Protein Sci.* 6:2628-30.

167. Kent C, Gee P, Lee SY, Bian X, Fenno JC. (2004) A CDP-choline pathway for phosphatidylcholine biosynthesis in *Treponema denticola*. *Mol. Microbiol.* 51:471-81.

168. Kerff F, et al. (2008) Crystal structure and activity of Bacillus subtilis YoaJ (EXLX1), a bacterial expansin that promotes root colonization. *Proc. Natl. Acad. Sci. U. S. A.* 105:16876-81.

169. Knight CA. (2000) Adding to the antifreeze agenda. *Nature* 406:249-51.

170. Knudsen B, Miyamoto MM. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl. Acad. Sci. U. S. A.* 98:14512-7.

171. Koiwa H, et al. (1997) Purification and characterization of tobacco pathogenesis-related protein PR-5d, an antifungal thaumatin-like protein. *Plant Cell Physiol.* 38:783-91.

172. Koiwa H, et al. (1999) Crystal structure of tobacco PR-5d protein at 1.8 *Å* resolution reveals a conserved acidic cleft structure in antifungal thaumatin-like proteins. *J. Mol. Biol.* 286:1137-45.

173. Kovall RA, Hendrickson WA. (2004) Crystal structure of the nuclear effector of Notch signaling, CSL, bound to DNA. *EMBO J.* 23:3441-51.

174. Kureishy N, Sapountzi V, Prag S, Anilkumar N, Adams JC. (2002) Fascins, and their roles in cell structure and function. *Bioessays* 24:350-61.

175. la Sala A, et al. (2003) Alerting and tuning the immune response by extracellular nucleotides. *J. Leukoc. Biol.* 73:339-43.

176. Lacy DB, Tepp W, Cohen AC, DasGupta BR, Stevens RC. (1998) Crystal structure of botulinum neurotoxin type A and implications for toxicity. *Nat. Struct. Biol.* 5:898-902.

177. Lad SP, et al. (2007) Chlamydial CT441 is a PDZ domain-containing tail-specific protease that interferes with the NF-kappaB pathway of immune response. *J. Bacteriol.* 189:6619-25.

178. Langebartels C, Ernst D, Kangasjarvi J, Sandermann H Jr. (2000) Ozone effects on plant defense. *Methods Enzymol.* 319:520-35.

179. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. (1996) Protein clefts in molecular recognition and function. *Protein Sci.* 5:2438-52.

180. Lee D, Redfern O, Orengo C. (2007) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell. Biol.* 8:995-1005.

181. Lehtio J, et al. (2003) The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules. *Proc. Natl. Acad. Sci. U. S. A.* 100:484-9.

182. Leppänen VM, et al. (2007) Crystal structure of the N-terminal NC4 domain of collagen IX, a zinc binding member of the laminin-neurexin-sex hormone binding globulin (LNS) domain family. *J. Biol. Chem.* 282:23219-30.

183. Leubner-Metzger G, Meins F Jr. (1999) Functions and regulation of plant β-1,3-glucanases (PR-2). In: Datta SK and Muthukrishnan S, editors. Pathogenesis-related proteins in plants. Boca Raton, FL: CRC Press LLC. p. 49-76.

184. Leubner-Metzger G, Meins F Jr. (2000). Sense transformation reveals a novel role for class I β-1,3-glucanase in tobacco seed germination. Plant J. 23: 215-21.

185. Leubner-Metzger G. (2003) Functions and regulation of β-1,3-glucanases during seed germination, dormancy release and after-ripening. *Seed Sci. Res.* 13:17-34.

186. Leubner-Metzger G. (2005). β-1,3-glucanase gene expression in low-hydrated seeds as a mechanism for dormancy release during tobacco after-ripening. *Plant J.* 41: 133-45.

187. Li W, Godzik A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-9.

188. Li WH, Yang J, Gu X. (2005) Expression divergence between duplicate genes. *Trends Genet.* 21:602-7.

189. Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB. (2003) WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res.* 31:3324-7.

190. Lichtarge O, Bourne HR, Cohen FE. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257:342-58.

191. Lillehoj EP, Kim BT, Kim KC. (2002) Identification of *Pseudomonas aeruginosa* flagellin as an adhesin for Muc1 mucin. *Am. J. Physiol. Lung. Cell. Mol. Physiol.* 282:L751-6.

192. Linthorst HJ, et al. (1990) Analysis of gene families encoding acidic and basic β-1,3-glucanases of tobacco. *Proc. Natl. Acad. Sci. U. S. A.* 87:8756-60.

193. Liu C, Xu Z, Gupta D, Dziarski R. (2001) Peptidoglycan recognition proteins: a novel family of four human innate immunity pattern recognition molecules. *J. Biol. Chem.* 276:34686-94.

194. Liu R, Ochman H. (2007) Stepwise formation of the bacterial flagellar system. *Proc. Natl. Acad. Sci. U. S. A.* 104:7116-21.

195. Liu Y, et al. (2000) Crystal structure of the cysteine-rich domain of mannose receptor complexed with a sulfated carbohydrate ligand. *J. Exp. Med.* 191:1105-16.

196. Logsdon JM Jr, Doolittle WF. (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94:3485-7.

197. Longly B, Waterkeyn L. (1977) Les stades callosiques de la plaque cellulaire somatique. *Cellule* 72:195-224.

198. Lu L, Tai G, Hong W. (2004) Autoantigen Golgin-97, an effector of Arl1 GTPase, participates in traffic from the endosome to the trans-golgi network. *Mol. Biol. Cell.* 15:4426-43.

199. Lukomski S, et al. (2000) Identification and characterization of the scl gene encoding a group A Streptococcus extracellular protein virulence factor with similarity to human collagen. *Infect. Immun.* 68:6542-53.

200. Lupas AN, Ponting CP, Russell RB. (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* 134:191-203.

201. Lynch M, Force A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459-73.

202. Maddison WP, Maddison DR. (2005) Mesquite: a modular system for evolutionary analysis. Available at:

http://www.mesquiteproject.org

203. Madura JD, Baran K, Wierzbicki A. (2000) Molecular recognition and binding of thermal hysteresis proteins to ice. *J. Mol. Recognit.* 13:101-13.

204. Mahalingam R, et al. (2003) Characterizing the stress/defense transcriptome of *Arabidopsis*. *Genome Biol.* 4:R20.

205. Makova KD, Li WH. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* 13:1638-45.

206. Malik A, Ahmad S. (2007) Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. *BMC Struct. Biol.* 7:1.

207. Marchler-Bauer A, et al. (2009) CDD: specific functional annotation with the conserved domain database. *Nucleic Acids Res.* 37:D205-10.

208. Marcotte EM, et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285:751-3.

209. Markert CL, Møller F. (1959) Multiple forms of enzymes: tissue, ontogenetic, and species specific patterns. *Proc. Natl. Acad. Sci. U. S. A.* 45:753-63.

210. Marques MA, Antonio VL, Sarno EN, Brennan PJ, Pessolani MC. (2001) Binding of alpha2-laminins by pathogenic and non-pathogenic mycobacteria and adherence to Schwann cells. *J. Med. Microbiol.* 50:23-8.

211. Marshall KM, Bradshaw M, Pellett S, Johnson EA. (2007) Plasmid encoded neurotoxin genes in *Clostridium botulinum* serotype A subtypes. *Biochem. Biophys. Res. Commun.* 361:49-54.

212. Martin-Cuadrado AB, Duenas E, Sipiczki M, Vazquez de Aldana CR, del Rey F. (2003) The endo-β-1,3-glucanase eng1p is required for dissolution of the primary septum during cell separation in *Schizosaccharomyces pombe*. *J. Cell. Sci.* 116:1689-98.

213. Maruyama Y, Momma M, Mikami B, Hashimoto W, Murata K. (2008) Crystal structure of a novel bacterial cell-surface flagellin binding to a polysaccharide. *Biochemistry* 47:1393-402.

214. McConkey BJ, Sobolev V, Edelman M. (2002) Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics* 18:1365-73.

215. McGee, K, Horstedt P, Milton DL. (1996) Identification and characterization of additional flagellin genes from *Vibrio anguillarum*. *J. Bacteriol.* 178:5188-98.

216. McLachlan AD. (1982) Rapid comparison of protein structures. *Acta Cryst.* A38:871-973.

217. McLean BW, et al. (2000) Analysis of binding of the family 2a carbohydrate-binding module from *Cellulomonas fimi* xylanase 10a to cellulose: specificity and identification of functionally important amino acid residues. *Protein Eng.* 13:801-9.

218. Memelink J, Linthorst HJ, Schilperoort RA, Hoge JH. (1990) Tobacco genes encoding acidic and basic isoforms of pathogenesis-related proteins display different expression patterns. *Plant Mol. Biol.* 14:119-26.

219. Miyanaga A, et al. (2004) Crystal structure of a family 54 α-L-arabinofuranosidase reveals a novel carbohydrate-binding module that can bind arabinose. *J. Biol. Chem.* 279:44907-14.

220. Montecucco C, Molgo J. (2005) Botulinal neurotoxins: revival of an old killer. *Curr. Opin. Pharmacol.* 5:274-9.

221. Moran NA. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108:583-6.

222. Moriishi K, et al. (1996) Mosaic structures of neurotoxins produced from *Clostridium botulinum* types C and D organisms. *Biochim. Biophys. Acta.* 1307:123-6.

223. Mortimer CK, Gharbia SE, Logan JM, Peters TM, Arnold C. (2007) Flagellin gene sequence evolution in Salmonella. *Infect. Genet. Evol.* 7:411-5.

224. Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536-40.

225. Murzin AG, Lesk AM, Chothia C. (1992) β-Trefoil fold: Patterns of structure and sequence in the Kunitz inhibitors interleukins-1β and 1α and fibroblast growth factors. *J. Mol. Biol.* 223:531-43.

226. Nagai H, Kagan JC, Zhu X, Kahn RA, Roy CR. (2002) A bacterial guanine nucleotide exchange factor activates ARF on Legionella phagosomes. *Science* 295:679-82.

227. Neduva V, Russell RB. (2005) Linear motifs: Evolutionary interaction switches. *FEBS Lett.* 579:3324-5.

228. Neduva V, Russell RB. (2006) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res.* 34:W350-5.

229. Nicol F, Hofte H. (1998) Plant cell expansion: scaling the wall. *Curr. Opin. Plant Biol.* 1:12-7.

230. Notenboom V, Boraston AB, Williams SJ, Kilburn DG, Rose D. (2002) High-resolution crystal structures of the lectin-like xylan binding domain from *Streptomyces lividans* xylanase 10A with bound substrates reveal a novel mode of xylan binding. *Biochemistry* 41:4246-54.

231. Ohno S. (1970) Evolution by gene duplication. Springer. Berlin. Alemania. 160 p.

232. Orengo CA, Jones DT, Thornton JM. (1994) Protein superfamilies and domain superfolds. *Nature* 372:631-4.

233. Orihuela CJ, et al. (2009) Laminin receptor initiates bacterial contact with the blood brain barrier in experimental meningitis models. *J. Clin. Invest.* 119:1638-46.

234. Ornitz DM, Itoh N. (2001) Fibroblast growth factors. *Genome Biol.* 2:REVIEWS3005.

235. Osmond RIW, Hrmova M, Fontaine F, Imberty A, Fincher GB. (2001) Binding interactions between barley thaumatin-like proteins and (1,3)-β-D-glucans: Kinetics, specificity, structural analysis and biological implications. *FEBS J.* 268:4190-9.

236. Paterson GK, Nieminen L, Jefferies JM, Mitchell TJ. (2008) PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol. Lett.* 285:170-6.

237. Pearson WR, Lipman DJ. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85:2444-8.

238. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96:4285-8.

239. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. (2001) The Comprehensive microbial resource. *Nucleic Acids Res.* 29:123-5.

240. Pohleven J, et al. (2009) Purification, characterization and cloning of a ricin B-like lectin from mushroom *Clitocybe nebularis* with antiproliferative activity against human leukemic T cells. *Biochim. Biophys. Acta.* 1790:173-81.

241. Polacco BJ, Babbitt PC. (2006) Automated discovery of 3D motifs for protein function annotation. *Bioinformatics* 22:723-30.

242. Ponting CP, Russell RB. (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all β-trefoil proteins. *J. Mol. Biol.* 302:1041-7.

243. Ponting CP. (2000) Novel repeats in ryanodine and IP3 receptors and protein O-mannosyltransferases. *Trends Biochem. Sci.* 25:48-50.

244. Posada D. (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.* 19:708-17.

245. Priestle JP, Schär HP, Grütter MG. (1989) Crystallographic refinement of interleukin 1 β at 2.0 A resolution. *Proc. Natl. Acad. Sci. U. S. A.* 86:9667-71.

246. Proux-Gillardeaux V, Gavard J, Irinopoulou T, Mge RM, Galli T. (2005) Tetanus neurotoxin-mediated cleavage of cellubrevin impairs epithelial cell migration and integrin-dependent cell adhesion. *Proc. Natl. Acad. Sci. U. S. A.* 102:6362-7.

247. Punj V, et al. (2000) Phagocytic cell killing mediated by secreted cytotoxic factors of *Vibrio cholerae*. *Infect. Immun.* 68:4930-7.

248. Puntervoll P, et al. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* 31:3625-30.

249. Rabaan AA, Gryllos I, Tomás JM, Shaw JG. (2001) Motility and the polar flagellum are required for *Aeromonas caviae* adherence to HEp-2 cells. *Infect. Immun.* 69:4257-67.

250. Rambukkana A, et al. (1998) Role of α-dystroglycan as a Schwann cell receptor for *Mycobacterium leprae*. *Science* 282:2076-9.

251. Ramos HC, Rumbo M, Sirard JC. (2004) Bacterial flagellins: mediators of pathogenicity and host immune responses in mucosa. *Trends Microbiol.* 12:509-17.

252. Rawlings ND, Morton FR, Barrett AJ. (2006) MEROPS: the peptidase database. *Nucleic Acids Res.* 34:D270-2.

253. Reid SD, Selander RK, Whittam TS. (1999) Sequence diversity of flagellin (fliC) alleles in pathogenic *Escherichia coli*. *J. Bacteriol.* 181:153-60.

254. Rinne PL, Kaikuranta PM, van der Schoot C. (2001) The shoot apical meristem restores its symplasmic organization during chilling-induced release from dormancy. *Plant J.* 26:249-64.

255. Ronquist F, Huelsenbeck JP. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-4.

256. Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.* 12:85-94.

257. Roudier F, Schindelman G, DeSalle R, Benfey PN. (2002) The COBRA family of putative GPI-anchored proteins in *Arabidopsis*. A new fellowship in expansion. *Plant Physiol.* 130:538-48.

258. Rutenber E, Ready M, Robertus JD. (1987) Structure and evolution of ricin B chain. *Nature* 326:624-26.

259. Sadreyev RI, Tang M, Kim BH, Grishin NV. (2007) COMPASS server for remote homology inference. *Nucleic Acids Res.* 35:W653-8.

260. Saitou N, Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-2.

261. Salazar-Gonzalez RM, McSorley SJ. (2005) Salmonella flagellin, a microbial target of the innate and adaptive immune system. *Immunol. Lett.* 101:117-22.

262. Samatey FA, et al. (2001) Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature* 410:331-7.

263. Samuel AD, et al. (1999) Flagellar determinants of bacterial sensitivity to chi-phage. *Proc. Natl. Acad. Sci. U. S. A.* 96:9863-6.

264. Samuels AL, Giddings TH Jr, Staehelin LA. (1995) Cytokinesis in tobacco BY-2 and root tip cells: a new model of cell plate formation in higher plants. *J. Cell Biol.* 130:1345-57.

265. Sarowar S, et al. (2009) Overexpression of lipid transfer protein (LTP) genes enhances resistance to plant pathogens and LTP functions in long-distance systemic signaling in tobacco. *Plant Cell Rep.* 28:419-27.

266. Schmid M, et al. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* 37: 501-6.

267. Schultz J, Milpetz F, Bork P, Ponting CP. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* 95:5857-64.

268. Schwede T, Kopp J, Guex N, Peitsch MC. (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* 31:3381-5.

269. Sebaihia M, et al. (2001) Genome sequence of a proteolytic (Group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes. *Genome Res.* 17:1082-92.

270. Sebestyén A, Muskotál A, Végh BM, Vonderviszt F. (2008) The hypervariable D3 domain of Salmonella flagellin is an autonomous folding unit. *Protein Pept. Lett.* 15:54-7.

271. Sela-Buurlage MB, et al. (1993) Only specific tobacco (*Nicotiana tabacum*) chitinases and β-1,3-glucanases exhibit antifungal activity. *Plant Physiol.* 101:857-63.

272. Sherman DL, Fabrizi C, Gillespie CS, Brophy PJ. (2001) Specific disruption of a schwann cell dystrophin-related protein complex in a demyelinating neuropathy. *Neuron* 30:677-87.

273. Shionyu-Mitsuyama C, Shirai T, Ishida H, Yamane T. (2003) An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins. *Protein Eng.* 16:467-78.

274. Sieg F, Schroder W, Schmitt JM, Hincha DK. (1996) Purification and characterization of a cryoprotective protein (cryoprotectin) from the leaves of cold-acclimated cabbage. *Plant Physiol.* 111:215-21.

275. Sikora S, Strongin A, Godzik A. (2005) Convergent evolution as a mechanism for pathogenic adaptation. *Trends Microbiol.* 13:522-7.

276. Simpson LL. (1980) Kinetic studies on the interaction between botulinum toxin type A and the cholinergic neuromuscular junction. *J. Pharmacol. Exp. Ther.* 212:16-21.

277. Sjölander K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 20(2):170-9.

278. Smith JM. (1992) Analyzing the mosaic structure of genes. *J. Mol. Evol.* 34:126-9.

279. Smith KD, Ozinsky A. (2002) Toll-like receptor-5 and the innate immune response to bacterial flagellin. *Curr. Top. Microbiol. Immunol.* 270:93-108.

280. Smith NH, Beltran P, Selander RK. (1990) Recombination of Salmonella phase 1 flagellin genes generates new serovars. *J. Bacteriol.* 172:2209-16.

281. Smith NH, Selander RK. (1990) Sequence invariance of the antigen-coding central region of the phase 1 flagellar filament gene (fliC) among strains of *Salmonella typhimurium*. *J. Bacteriol.* 172:603-9.

282. Smith TJ, et al. (2007) Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3, /Ba4 and /B1 clusters are located within plasmids. *PLoS ONE* 2:e1271.

283. Snider C, Hsiang T, Zhao G, Griffith M. (2000) Role of ice nucleation and antifreeze activities in pathogenesis and growth of snow molds. *Phytopathology* 90:354-61.

284. Söding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951-60.

285. Söding J, Lupas AN (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* 25:837-46.

286. Söding J, Remmert M, Biegert A (2006) HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res.* 34:W137-42.

287. Sohlenkamp C, López-Lara IM, Geiger. (2003) Biosynthesis of phosphatidylcholine in bacteria. *Prog. Lipid Res.* 42:115-62.

288. Stark A, Russell RB. (2003) Annotation in three dimensions. PINTS: Patterns in non-homologous tertiary structures. *Nucleic Acids Res.* 31: 3341-4.

289. Stahl EA, Bishop JG. (2000) Plant-pathogen arms races at the molecular level. *Curr Opin Plant Biol.* 3:299-304.

290. Stebbins CE, Galán JE. (2001) Structural mimicry in bacterial virulence. *Nature* 412:701-5.

291. Steiniger-White M, Rayment I, Reznikoff WS. (2004) Structure/function insights into Tn5 transposition. *Curr. Opin. Struct. Biol.* 14:50-7.

292. Stern A, et al. (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res.* 35:W506-11.

293. Stevens R, Goble C, Baker P, Brass A. (2001) A classification of tasks in bioinformatics. *Bioinformatics* 17:180-8.

294. Stone BA, Clarke AE. (1992) Chemistry and biology of (1-3)-β-D-glucans. Victoria, Australia: La Trobe University Press.

295. Sun L, Gurnon JR, Adams BJ, Graves MV, Van Etten LJ. (2000) Characterization of a β-1,3-glucanase encoded by chlorella virus PBCV-1. *Virology* 276:27-36.

296. Sun HY, et al. (2007) Structure and mechanism of *Helicobacter pylori* fucosyltransferase. A basis for lipopolysaccharide variation and inhibitor design. *J. Biol. Chem.* 282:9973-82.

297. Swaminathan S, Eswaramoorthy S. (2000) Structural analysis of the catalytic and binding sites of *Clostridium botulinum* neurotoxin B. *Nat. Struct. Biol.* 7:693-9.

298. Takeda H, et al. (2004) Molecular cloning of two exo-β-glucanases and their in vivo substrates in the cell walls of lily pollen tubes. *Plant Cell Physiol.* 45:436-44.

299. Takishita K, Inagaki Y. (2009) Eukaryotic origin of glyceraldehyde-3-phosphate dehydrogenase genes in *Clostridium thermocellum* and *Clostridium cellulolyticum* genomes and putative fates of the exogenous gene in the subsequent genome evolution. *Gene* 441(1-2):22-7.

300. Taroni C, Jones S, Thornton JM. (2000) Analysis and prediction of carbohydrate binding sites. *Protein Eng.* 13:89-98.

301. Thiruv B, Quon G, Saldanha SA, Steipe B. (2005) Nh3d: a reference dataset of non-homologous protein structures. *BMC Struct. Biol.* 5:12.

302. Thomas PD, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13:2129-41.

303. Thomsen S, Bosch TCG. (2006) Foot differentation and genomic plasticity in *Hydra*: lessons from the PPOD gene family. *Dev. Genes Evol.* 216:57-68.

304. Thornton JM, Orengo CA, Todd AE, Pearl FM. (1999) Protein folds, functions and evolution. *J. Mol. Biol.* 293:333-42.

305. Todd AE, Orengo CA, Thornton JM. (1999) Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.* 3:548-56.

306. Tominaga A. (2004) Characterization of six flagellin genes in the H3, H53 and H54 standard strains of *Escherichia coli*. *Genes Genet. Syst.* 79:1-8.

307. Toufighi K, Brady SM, Austin R, Ly E, Provart NJ. (2005) The botany array resource: e-northerns, expression angling, and promoter analyses. *Plant J.* 43: 153-63.

308. Transue TR, Smith AK, Mo H, Goldstein IJ, Saper MA. (1997) Structure of benzyl T-antigen disaccharide bound to *Amaranthus caudatus* agglutinin. *Nat. Struct. Biol.* 4:779-83.

309. Trautmann A. (2009) Extracellular ATP in the Immune System: More Than Just a "Danger Signal". *Sci. Signal.* 2:pe6.

310. Uchida T, et al. (2004) Crystal structure of the hemolytic lectin CEL-III isolated from the marine invertebrate *Cucumaria echinata*: implications of domain structure for its membrane pore-formation mechanism. *J Biol Chem.* 279:37133-41.

311. Uchiyama T, et al. (2001) Roles of the exposed aromatic residues in crystalline chitin hydrolysis by chitinase A from *Serratia marascens* 2170. *J. Biol. Chem.* 276:41343-9.

312. Valencia A. (2005) Automatic annotation of protein function. *Curr. Opin. Struct. Biol.* 15:267-74.

313. van der Meer-Janssen YP, van Galen J, Batenburg JJ, Helms JB. (2009) Lipids in host-pathogen interactions: Pathogens exploit the complexity of the host cell lipidome. *Prog. Lipid Res.* 49:1-26.

314. van Noort V, Snel B, Huynen MA. (2003) Predicting gene function by conserved co-expression. *Trends Genet.* 19:238-42.

315. van Valen L. (1973) A new evolutionary law. *Evol. Theory.* 1:1-30.

316. Varghese JN, et al. (1994) Three-dimensional structures of two plant β-glucan endohydrolases with distinct substrate specificities. *Proc. Natl. Acad. Sci. U. S. A.* 91:2785-9.

317. Villanueva JM, et al. (1999) INNER NO OUTER regulates abaxial- adaxial patterning in *Arabidopsis* ovules. *Genes Dev.* 13:3160-9.

318. von Mering C, et al. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31:258-61.

319. Wang G, Dunbrack RL, Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19:1589-91.

320. Wang L, Rothemund D, Curd H, Reeves PR. (2003) Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. *J. Bacteriol.* 185:2936-43.

321. Ward E, et al. (1991) Differential regulation of β-1,3-glucanase messenger RNAs in response to pathogen infection. *Plant Physiol.* 96:390-7.

322. Watson JD, Laskowski RA, Thornton JM. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* 15:275-84.

323. Wheeler PR. (1989) Pyrimidine scavenging by *Mycobacterium leprae*. *FEMS Microbiol. Lett.* 48:179-84.

324. Whisstock JC, Lesk AM. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* 36:307-40.

325. Wiener M, Freymann D, Ghosh P, Stroud RM. (1997) Crystal structure of colicin Ia. *Nature* 385:461-4.

326. Wittlieb J, Khalturin K, Lohmann JU, Anton-Erxleben F, Bosch TC. (2006) Transgenic *Hydra* allow in vivo tracking of individual stem cells during morphogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 103:6208-11.

327. Wolf YI, Grishin NV, Koonin EV. (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* 299:897-905.

328. Woloshuk CP, Meulenhoff JS, Sela-Buurlage M, van den Elzen PJ, Cornelissen BJ. (1991) Pathogen-induced proteins with inhibitory activity toward *Phytophthora infestans*. *Plant Cell.* 3:619-28.

329. Wolynes PG. (1996) Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci. U. S. A.* 93:14249-55.

330. Woody OZ, Doxey AC, McConkey BJ. (2008) Assessing the evolution of gene expression using microarray data. *Evol. Bioinform. Online.* 4:139-52.

331. Wu LF, Hughes TR, et al. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31:255-65.

332. Yaish MWF, Doxey AC, McConkey BJ, Moffatt BA, Griffith M. (2006) Cold active winter rye glucanases with ice-binding capacity. *Plant Physiol.* 141:1459-72.

333. Yang DS, et al. (1998) Identification of the ice-binding surface on a type III antifreeze protein with a "flatness function" algorithm. *Biophys. J.* 74:2142-51.

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

335. Yao R, Burr DH, Doig P, Trust TJ, Guerry P. (1994) Isolation of motile and non-motile insertional mutants of *Campylobacter jejuni*: the role of motility in adherence and invasion of eukaryotic cells. *Mol. Microbiol.* 14:883-93.

336. Ye K, Malinina L, Patel DJ. (2003) Recognition of small interfering RNA by a viral suppressor of RNA silencing. *Nature* 426:874-8.

337. Yu S. (2010) Antifreeze proteins: activity comparisons and *de novo* design of an ice-binding protein. MSc. Thesis. Queen's University, Kingston, ON, Canada.

338. Yu Y, et al. (2009) Effects of *Mycoplasma pneumoniae* infection on sphingolipid metabolism in human lung carcinoma A549 cells. *Microb. Pathog.* 46:63-72.

339. Zevenhuizen LPTM, Bartnicki-Garcia S. (1969) Structure of the insoluble hyphal wall glucan of *Phytophthora cinnamomi*. *Biochemistry* 8:1496-1502.

340. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W. (2004) GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. *Plant Physiol.* 136:2621-32.

341. Zhang, J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18:292-298.

342. Zuo J, et al. (2000) KORRIGAN, an *Arabidopsis* endo-1,4-β-glucanase, localizes to the cell plate by polarized targeting and is essential for cytokinesis. *Plant Cell* 12:1137-52.