MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **58** (2007) 281-299

ISSN 0340 - 6253

Discovery of Recurrent Sequence Motifs in *Saccharomyces cerevisiae* Cell Wall Proteins

Juan E. Coronado[1], Susan L. Epstein[2], Wei-Gang Qiu[1], Peter N. Lipke[1*]

[1]Department of Biological Sciences and Center for Gene Structure and Function, Hunter College of City University of New York, New York, NY 10021, USA

[2]Department of Computer Science, Hunter College of City University of New York, New York NY 10021, USA

(Received May 17, 2006)

*To whom inquiries should be addressed (current address): Peter Lipke, Dept. of Biology, Brooklyn College, 2900 Bedford Ave., Brooklyn, NY 11210 USA. (718)-951-5000 X1949; plipke@brooklyn.cuny.edu

Abstract

This paper describes a procedure for the discovery of recurrent substrings in amino acid sequences of proteins, and its application to fungal cell walls. The evolutionary origins of fungal cell walls are an open biological question. This question can be approached by studies of similarity among the sequences and sub-sequences of fungal wall proteins and by comparison to proteins in animals. We describe here how we have discovered building blocks, represented as recurrent sequence motifs (sub-sequences), within fungal cell wall proteins. These motifs have not been systematically identified before, because the low Shannon entropy of the cell wall sequences has hindered searches for local sequence similarities by sequence alignments. Nonetheless, our new, composition-based scoring matrices for local alignment searches now support statistically valid alignments for such low entropy sequences (Coronado et al. 2006. Euk. Cell 5: 628-637). We have now searched for similarities in a set of 171 known and putative cell wall proteins from baker's yeast, *Saccharomyces cerevisiae*. The aligned segments were repeatedly subdivided and catalogued to identify 217 recurrent sequence motifs of length 8 amino acids or greater. 95% of these motifs occur in more than one cell wall protein. The median length of the motifs is 22 amino acid residues, considerably shorter than protein domains. For many cell wall proteins, these motifs collectively account for more than half of their amino acids. The prevalence of these motifs supports the idea of fungal cell wall proteins as assemblies of recurrent building blocks.

Introduction

Discovery collects and organizes information in ways that are meaningful to the user. The identification of regularities in a large knowledge base is of particular interest when their existence is supported by other information. This paper describes such a discovery. We hypothesize that in unusual low-complexity protein-based sequences there are recurrent *motifs*, macros that provide building blocks for protein construction. The external support for our hypothesis is evolution. The principle results of this paper are the discovery of such motifs for yeast cell wall proteins in the fungus *Saccharomyces cerevisiae* (bakers' yeast) and the visually-oriented methods used to find them.

The evolutionary history of cell walls in fungi is an intriguing question. Fungi are a sister group to the animals, a non-walled kingdom, and both groups are postulated to descend from a common ancestor without a wall (10, 15, 18, 22). The question is therefore "How did fungal walls evolve, and what materials were used to construct this phylogenetically unique cellular structure?" In fact, anecdotal evidence suggests that recurrent sequence motifs are common in fungal wall proteins (9, 15, 16, 20). If this observation were shown to be generally true, then we could hypothesize that such motifs are "building blocks" that are replicated to make up a substantial and functionally critical portion of the proteins in the wall.

This question can be approached by comparative studies of the genes that encode the proteins in the walls, and comparisons of evolutionary history of the proteins and their component parts. Studies of molecular evolution depend upon the comparison of protein *sequences* (variable-length strings on a 20-letter alphabet of amino acid residues). Comparisons of sequence similarities and differences allow the inference of gene divergence and re-arrangements, and therefore of evolutionary history. The

occurrence of similar sequences in two different organisms or in multiple copies in one organism results from *homology,* mutual inheritance from a common ancestor. Homologous sequences diverge at a rate dependent upon the mutation rate and the strength of selection for or against changes in the sequence. Sequences tend to be more strongly conserved if they have a beneficial function, and in such cases there is evolutionary pressure to preserve the inherited sequences unchanged. If a sequence is not beneficial it may be neutral and allowed to mutate freely, or a sequence may be harmful, in which case mutations that abrogate its function are positively selected.

Some sequences or *fragments* (substrings) occur multiple times in the genome of an organism. *Paralogs* are fragment recurrences within a single organism due to duplications of the DNA during replication within this organism or in its ancestors. Duplicated copies can be recombined into other parts of the genome by transposition (19). Such duplications may persist in the genomes unless they are selected against. Like other homologous sequences, paralogous sequences can be beneficial, neutral, or harmful. The rate of accumulation of mutational substitution in paralogs is an indicator of the evolutionary pressure for or against mutation and of the time since the paralogs' creation by duplication.

The origin and evolution of fungal cell walls are problems whose solutions have been hampered by the lack of good methods to identify and compare the glycoproteins that predominate in fungal walls (5-7). Although 103 of the proteins in the *S. cerevisiae* genome are known or predicted to be cell wall proteins (3, 6, 8), only a few of the proteins in the genome have known biological function (e.g., see Table 1). Since sequence similarity suggests functional similarity, our knowledge base should also include sequences similar to known cell wall proteins. Similarity between two sequences

can be measured by the quality of the best alignment between them. (An alignment creates a one-to-one mapping between the sequences, and permits the insertion of *gaps*, sub-sequences of blanks.) The *score* of an alignment measures its quality; identical or functionally similar residues should be paired, and the number and length of the gaps minimized. The *e-value* (labeled “Expect” in Figure 1) is a statistical estimate of the probability of an alignment score that is the same or greater between two random sequences (2, 11). Thus we seek proteins in the genome that have high-scoring (i.e., low *e*-value) alignments with known cell wall proteins.

Gene[a]	ORF[b]	Fragment Number	Sequence of Fragment	GO Annotation[c]		
				Location	Process	Function
Motif 12						
	YPL282C	1	VTRVITGVPWYSTRL	u	U	u
DAN2	YLR037C	2	VTRVITGVPWYSTRL	cw	U	u
	YMR325W	3	VTRVITGVPWYSTRL	u	U	u
	YOR394W	4	VTRVITGVPWYSTRL	u	U	u
	YLL025W	5	VTRVITGVPWYSTRL	u	U	u
	YIR041W	6	VTRVITGVPWYSTRL	u	U	u
	YKL224C	7	VTRVITGVPWYSTRL	u	U	u
PAU3	YCR104W	8	VTRVITGVPWYSTRL	u	U	u
PAU6	YNR076W	9	VTRMITGVPWYSTRL	u	U	u
DAN4	YJR151C	10	---MITGVPWYSTRL	cw	U	u
	YLL064C	11	VTRMITGVPWYSTRL	u	U	u
DAN1	YJR150C	12	VTRMITGVPWYSTRL	cw	St	u
	YOL161C	13	VTRMITGVPWYSTRL	u	U	u
PAU5	YFL020C	14	VTRVITGVPWYSSRL	u	U	u
PAU1	YJL223C	15	VTRMITGVPWYSSRL	u	U	u
	YIL176C	16	VTRMITGVPWYSSRL	u	U	u
	YGR294W	17	VTRMITGVPWYSSRL	u	U	u
PAU2	YEL049W	18	VTRMITGVPWYSSRL	u	U	u
	YBL108C-A	19	VTRMITGVPWYSSRL	u	U	u
	YAL068C	20	---MITGVPWYSSRL	u	U	u
PAU4	YLR461W	21	---MITGVPWYSSRL	u	U	u
	YDR542W	22	---MITGVPWYSSRL	u	U	u
	YGL261C	23	---MITGVPWYSSRL	u	U	u
	YHL046C	24	---MITGVPWYSSRL	u	U	u
DAN3	YBR301W	25	---MITGVPWYSSRL	u	U	u
TIR2	YOR010C	26	VSKMLTMVPWYSSRL	cw	Sr	u
TIR1	YER011W	27	VSKMLTMVPWYSSRL	cw	Sr	cwc
TIR4	YOR009W	28	---MLTMVPWYSSRL	cw	U	u

Motif 25[d]						
PST1	YDR055W	1	IYISDTSLQSVDGFSALKKVNVFNVNNNKKLTSIKSPVETVSDSLQFSFNGNQTKITFDD	cw	Cwob	u
ECM33	YBR078W	2	IIVSDTTLESVEGFSTLKKVNVFNINNNRYLNSFQSSLESVSDSLQFSSNGDNTTLAFDN	pm/cw	Cwob	u
	YCL048W	3	IVISDTSLANIENFNKVQEIDTFNINNNRFLETIHSNVKTIRGQFSVHANAKELELEMPH	pm	cwob/swa	u
SPS2	YDR522C	4	--ISDTALTSIDYFNNVKKVDIFNINNNRFLENLFASLESVTKQLTVHSNAKELELDLSN	cw	Swa	u

[a] Name given where assigned

[b] From *Saccharomyces* Genome Database

[c] Gene Ontology Annotation: u, unknown; cw, cell wall; pm, plasma membrane; st, sterol transport; sr, stress response; cwob, cell wall organization and biogenesis; swa, spore wall assembly; cwc, cell wall constituent.

[d] The first 60 amino acids of the 177-amino acid HSP

Table 1. Members of fragment families 12 and 25. The ORF name is the reference from the *Saccharomyces* Genome Database. The gene name, aligned sequences in single letter code, and Gene Ontology annotation are shown. Acronyms are listed and defined in the Appendix.

Existing powerful tools for sequence alignment (2) accept a *query* sequence and return sequences most similar to it. These tools assume that the query has high *Shannon entropy* (high diversity in sequence elements, with no individual element at greater frequency than about 15%). In the sequences for cell wall proteins in *S. cerevisiae*, however, a few letters of the alphabet are over-represented; cell wall proteins are especially rich in the amino acids serine (symbols S or Ser), threonine (T or Thr) and a few others. As a result, these sequences have low Shannon entropy and the standard search methods, BLAST and FASTA (2, 17), cannot discriminate between sequences similar to the query and dissimilar ones of similar composition (6, 23). This problem, called *low-complexity corruption*, is also present in other low-entropy proteins, including mammalian mucins and other glycoproteins. Low-complexity corruption is caused by alignment scores that are based on high scores from matrices appropriate for high-entropy sequences.

Throughout this work, we have enhanced standard search tools with our *gtQ matrices.* described in (6). These composition-modified scoring matrices define high-scoring residue pairs, and are calculated for each query sequence. They reduce the score for aligned residues *i* and *j* in proportion to the likelihood of a random *ij* alignment in sequences of similar composition to the query sequence. Two important criteria for good alignments are *discrimination* (the ability to distinguish strings with similar sequence from those with similar amino acids composition but different sequence) and *sensitivity* (the ability to identify a maximal number of similar sequences). BLAST searches with gtQ matrices (*BLAST-gtQ searches*) have improved discrimination and do not sacrifice sensitivity for cell-wall protein queries against fungal genome databases (6).

To interpret a genome, DNA sequences on the successive 3-element substrings of the 4-letter nucleic acid alphabet are rewritten as successive single elements in the 20-letter amino acid alphabet. We use the term *Open Reading Frame (ORF)* here to denote a potential protein sequence over the 20-letter alphabet. An ORF is always delimited by pre-specified start and stop signals. For statistical

reasons, ORFs are defined to be at least 75 amino acids long; shorter ORFs are biologically present but rare. When an ORF has been demonstrated to exist biochemically or genetically, it is also called a *protein sequence.* (As a result, each gene in a genome that could encode a protein has an ORF name, and many also have a gene name and a protein name.) The set of all protein sequences and other ORFs from baker's yeast (*Saccharomyces cerevisiae*) was our knowledge base. In *S. cerevisiae*, gene names are italicized with three capital letters and a numeral, e.g. *DAN1* and *ECM33.*

Motifs are recurring sub-sequences found in one or more ORFs. We used BLAST-gtQ searches to find and align homologs of cell wall proteins and ORFs in *S. cerevisiae*. BLAST reports High Scoring Pairs (*HSPs*), well-aligned pairs of sequences consisting of whole sequences or sub-sequences that have high alignment scores and low *e*-values. We then devised a strategy to define and compare the sequence motifs that are paralogous within the *S. cerevisiae* wall proteome, the set of proteins that are located in the cell wall or are homologous to known cell wall proteins (6, 7). Our results greatly constrain the possible models for evolution of fungal cell walls.

Results

We carried out BLAST-gtQ searches to identify the paralogs of known cell wall proteins in the yeast proteome. New protein sequences or ORFs identified as homologs were then used as queries in a second round, and the process was repeated until no new HSPs were identified. The HSPs from all these searches were then used to define a set of recurrent sequence motifs that make up a large part of the sequences of cell wall proteins (the cell wall *proteome*).

Identification of paralogs of cell wall proteins. The query set was the 103 *S. cerevisiae* cell wall proteins annotated as cell wall in the Gene Ontology (GO) database or identified as glycosylphosphatidyl inositol-anchored (GPI-anchored) proteins (3, 7). BLAST-gtQ searches of the *S. cerevisiae* genome identified 1597 HSPs with *e* values $\leq 10^{-5}$. Two examples of HSPs are shown in

Figure 1. The search found such HSPs in 68 other ORFs in *S. cerevisiae*. Thus a total of 171 ORFs, including the original 103 queries, were identified as cell wall components or their paralogs.

```
Query= YAL063C
>YHR213W YHR213W SGDID:S0001256, Chr VIII from 539147-539743,
     Uncharacterized ORF
     Length = 198

 Score = 403 bits (949), Expect = e-112
 Identities = 150/179 (83%), Positives = 161/179 (89%)

Query: 140 MTGYFLPPQTGSYTFKFATVDDSAILSVGGSIAFECCAQEQPPITSTNFTINGIKPWNGS 199
           MTGYFLPPQT SYTF+FA VDDSAILSVGG +AFECCAQEQPPITST+FTINGIKPW GS
Sbjct: 1   MTGYFLPPQTSSYTFRFAKVDDSAILSVGGNVAFECCAQEQPPITSTDFTINGIKPWQGS 60

Query: 200 PPDNITGTVYMYAGFYYPMKIVYSNAVAWGTLPISVTLPDGTTVSDDFEGYVYTFDNNLS 259
            PDNI G VYMYAG+YYP+K+VYSNAV+WGTLPISV LPDGTTVSDDFEGYVYTFD++LS
Sbjct: 61  LPDNIGGTVYMYAGYYYPLKVVYSNAVSWGTLPISVELPDGTTVSDDFEGYVYSFDDDLS 120

Query: 260 QPNCTIPDPSNYTVSTTITTTEPWTGTFTSTSTEMTTVTGTNGVPTDETVIVIRTPTTA 318
           Q NCTIPDPS +T S   TTTE WTGTFTSTSTEMTTVTGTNG PTDETVIV + PTTA
Sbjct: 121 QSNCTIPDPSKHTTS IVTTTTELWTGTFTSTSTEMTTVTGTNGQPTDETVIVAKAPTTA 179

-----------------------------------------------------------------

 Score = 90.2 bits (205), Expect = 2e-18
 Identities = 53/63 (84%), Positives = 56/63 (88%), Gaps = 1/63 (1%)
Query: 815 LVTTTTEPWTGTFTSTSTEMTTITGTNGQPTDETVIIVKTPTTAISSSLSSSSG-QITSF 873
           +VTTTTE WTGTFTSTSTEMTT TGTNGQPTDETVI+ K PTTA SSSLSSSS  QITS
Sbjct: 136 IVTTTTELWTGTFTSTSTEMTTVTGTNGQPTDETVIVAKAPTTATSSSLSSSSSEQITSS 195

Query: 874 ITS 876
           ITS
Sbjct: 196 ITS 198
```

Figure 1: Two low-complexity BLAST HSPs generated for the query yeast ORF YAL063c with gtQ scoring. The top HSP shows an alignment of amino acids 140-318 of ORF YAL063c with amino acids 1-179 of ORF YHR213w. The bottom HSP is residues 815-876 of YAL063c with 136-198 of YHR213w. In the middle rows of the alignments, identical amino acids are repeated and similar amino acids score a "+". Note that amino acid residues 136-179 of YHR213w (in italics) align similarly with two regions of the query sequence (positions 275-318 in the first match and 815-858 in the second).

Identification and alignment of cell wall sequence motifs. The next challenge was to determine whether the HSPs contained recurrent sequence motifs. Traditionally, such motifs are found by sequence similarity within functional regions of proteins or by searches for recurrent sub-sequences within an ORF (motif searches). For cell wall proteins, however, few functional regions are known, and their low-complexity regions make searches for repeats slow, insensitive to variations within motifs, and unable to discriminate against sub-sequences consisting of a single amino acid

(repeats of a single letter) (1, 4, 14). We therefore adopted an approach that divides HSPs into unique sub-sequences (those fragments without homologs) and recurrent sub-sequences (those with at least one homolog). This approach is illustrated in Figure 2.

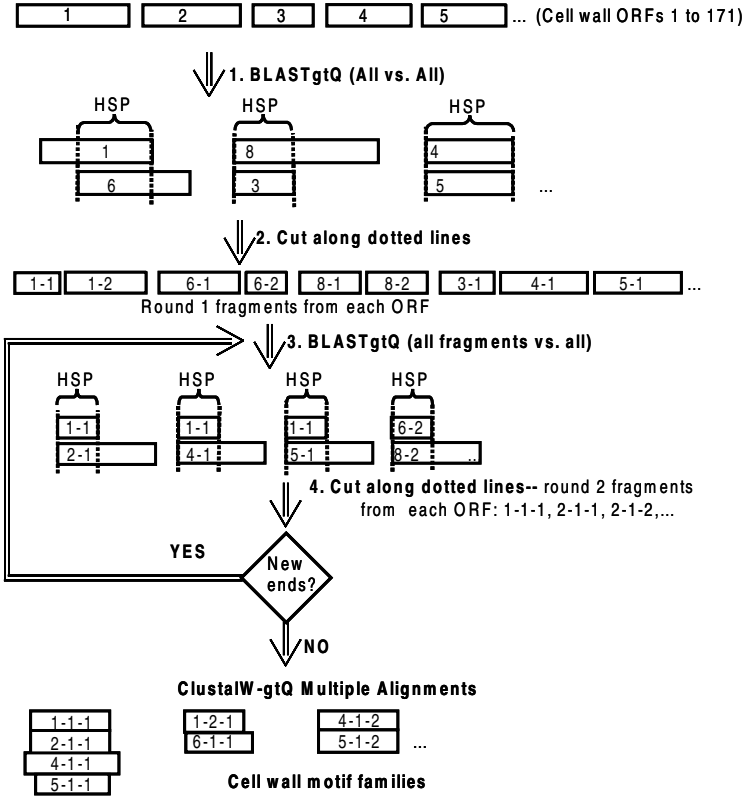

Fig. 2. Fragmentation process. (1) All 171 putative cell wall proteins were compared with BLAST-gtQ searches. Pairwise sequence alignments with $e \leq 10^{-5}$ detected homologies, denoted as vertically aligned segments. Note that more than one good alignment from different portions of a query is possible. (2) The boundaries of these alignments were treated as cuts to produce fragments with lengths $\geq n$. (3) All fragments were compared with BLAST-gtQ searches, and all new fragment alignments with length $\geq n$ and $e \leq 10^{-3}$ were used to cut the fragments as in step (2). Because the gtQ scoring matrices are modified based on the composition of each query string, statistically significant alignments were produced even with short fragments like 1-1. (4) Fragment alignment and cutting continued until there were no new fragments with length $\geq n$. Finally, fragments with sequence

homology were aligned with CLUSTALW-gtQ to produce cell wall motif families.

Both ORFs in a BLAST HSP with $e \leq 10^{-5}$ were partitioned into fragments at the boundaries of the match. Each resulting fragment of length at least n was then used as the query in a new BLAST-gtQ search against all other wall protein fragments of length $\geq n$. The newly-aligned fragments were again cut at their boundaries as long as both resulting fragments would have length $\geq n$, and the process was repeated until no new fragments were identified with length $\geq n$ and $e \leq 10^{-3}$. (Because of the reduced length of the query strings, searches after the first round used a cut-off of $e \leq 10^{-3}$ rather than $e \leq 10^{-5}$.) The result was a set of sequence fragments that were either unique or were similar to as many as 41 other sub-sequences in the 171 protein database. Those fragments that had at least one other similar sequence constitute the set of recurrent *cell wall motifs*. The number of cell wall motifs identified depended on the motif minimum length n. We investigated n values from 8 to 20; $n = 20$ identified 156 motifs, while $n = 8$ identified the most, 217. Because $n = 8$ gave the maximum number of cell wall motifs, we chose this set for further analysis.

Characteristics of cell wall motifs. Mutually paralogous motifs were aligned by CLUSTALW, again using composition-dependent gtQ scoring. CLUSTALW aligns a set of previously identified similar sequences in parallel, rather than pairwise, the way BLAST does (13). Each set of mutually aligned motifs was called a *motif family*. These motif families are available at http://diverge.hunter.cuny.edu:8080/modmat/misc.do?action=ShowCutDirs. The cell wall motifs ranged in length from 8 to 507 amino acids for $n = 8$. Ninety percent (90%) of the 217 motifs occurred once in each of multiple ORFs. The other fragments (10%) were present as two or more repeats in at least one ORF, and half of these (5%) occurred only as repeats in a single ORF. Thus 95% of the motifs were sub-sequences present in multiple ORFs.

Two representative alignments of cell wall motifs are shown in Table 1. Cell wall motif 12 is a highly similar group of 28 short fragments, each occurring exactly once in an ORF. Such

conservation of sequence is characteristic of recently duplicated sequences or of strong function conservation (and therefore sequence conservation) after more ancient origin of the copies. The former interpretation is unlikely for this motif family, given its widespread occurrence in sequences that are otherwise not homologous, because it would be unusual for all copies to have entered the genome at the same recent time. Because many recombination mechanisms (e.g., transposition) tend to insert multiple copies of sequences in one locus, it is also interesting that these motifs occur only once per ORF. It is possible that multiple insertions would be detrimental and therefore selected against. There are other cell wall motifs which do display a tendency to multiple insertions in the same ORF (19-21).

There is great diversity in size, membership, and evolutionary rate of cell wall motifs. Cell wall motif 25, the other example in Table 1, is longer (177 amino acids), present in 4 ORFs and has more sequence divergence. Motifs 12 and 25 illustrate the range in sequence length, frequency of occurrence, and divergence among HSPs. Cell wall motif 12 is highly conserved: there are 7 apparent amino acid polymorphisms (differences in amino acid sequence), while the other 8 residues remain constant. In contrast, fragment family 25 shows 73 polymorphic sites in the first 240 amino acids, 18-fold more substitutions per position in the alignment.

Most of the motifs are short, although a few are as long as ORFs and represent entire gene sequences that have been duplicated. For motifs of length $n \geq 8$, the median length is 22 amino acids; and 74% of them are of length 30 or less. These lengths are much shorter than protein domains, the longer sequences that define functional units of globular proteins. Such domains typically fold with discrete topologies, and have a median length near 100 amino acids.

We determined the prevalence of motifs in the cell wall proteins, by asking what fractions of the ORF sequences consisted of motifs. Twenty-eight of the original 171 query proteins contained no motifs (Figure 3). The other 143 wall proteins fell into three populations. About 34 proteins had some

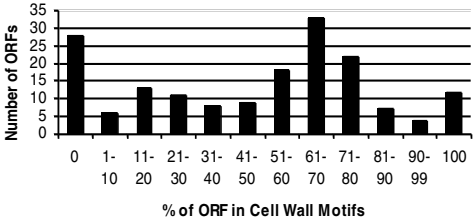

Figure 3. Fraction of cell wall protein sequences that are covered by motifs. The lengths of the fragments derived from a single cell wall protein were added and divided by the length of the protein.

motifs, with a mean motif content of about 20% of their sequences. The largest population (97 proteins) was composed mostly of motifs, with a mean motif content of about 65%. Surprisingly, 12 protein sequences were composed wholly of motifs. Thus, almost 70% of the wall proteins had motif content over 50%.

Cluster graphs. The relationships among the members of individual cell wall motif families can be represented as graphs (Fig. 4). These graphs reveal sequence characteristics less accessible from traditional BLAST and CLUSTALW alignments. Motifs with many edges (similarities) cluster more closely; those with fewer similarities protrude from the cluster. Thus it is possible to distinguish quickly between close and more distant relationships.

In Figure 4, the vertices represent the motif sequences in Table 1, and the edges represent a BLAST-gtQ alignment for each pair below the scoring threshold. The "`prefuse`" JAVA graph library was used to construct a graph with ORFs as vertices and BLAST similarities as edges (12). Although omitted here for clarity, the thicknesses of the edges can represent the BLAST *e* value, with lower *e* values producing thicker edges. The `prefuse` library generates the diagram by simulating an environment where vertices repel each other while edges attract based on their thickness. Motif 25 family is a *clique* (every possible edge appears among its vertices), showing close relationships

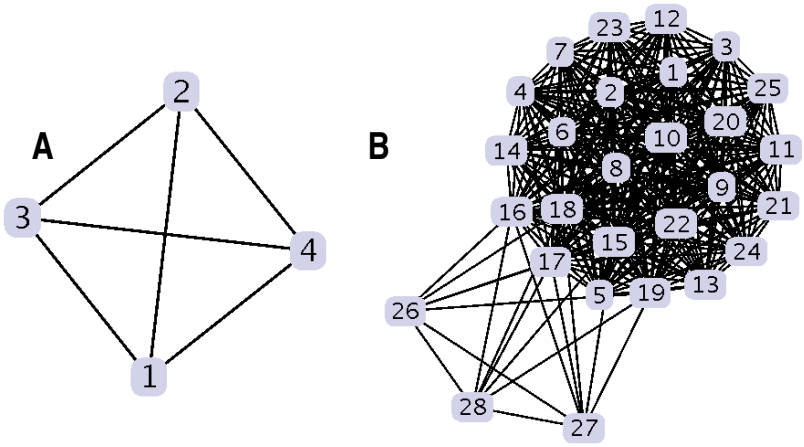

Figure 4. Graphs of motif families 25 (A) and 12 (B). The numbers indicate the order of the vertices in the multiple alignments in Table 1. For clarity all edges are the same thickness, but the graph can also be shown with thicker lines corresponding to smaller BLAST statistical *e*-values. Each edge represents a high-scoring alignment between a pair of motifs.

between the sequences. Motif family 12 is nearly a clique; 25 of the 28 members are completely interconnected. The three remaining sequences, however, form a clique (at the lower left in Figure 4B) with multiple edges (significant similarity) to 8 members of the larger clique Thus, the cluster graph demonstrates the inter-relationships among the individual members of the motif families. Others families are less clique-like (not shown).

Summary.

The shared functional and structural components of fungal cell wall proteins were largely unknown due to a lack of computational methods for detecting significant evolutionary relatedness among low-complexity sequences (6). Eukaryotic proteins are mosaics of domains, which are typically sequences of length 100 or more. Many of the sequences we identify here, however, are considerably shorter,

and therefore their detection is more demanding computationally. We have used compositionally modified gtQ matrices to identify homologs of known yeast cell wall proteins, and to carry out a search for recurrent sub-sequences among them. The resultant aligned pairs were the basis for division into those subsequences that occurred only once and those that occurred more often, the latter called *motifs* in analysis of protein sequences. We have calculated the fraction of these proteins that are covered by motifs, and used graphs to illustrate relationships among them. To our knowledge, this is the first systematic search and analysis of motifs in fungal cell wall proteins. We conclude that there is a relatively limited set of between 156 and 217 sequence motifs present in the set of 171 putative wall proteins. A large proportion of the total sequence length represented by these proteins is part of one or more recurrent motifs. Some of these motifs have been noticed anecdotally before (such as the *DAN/PAU* and *PIR/TIR* gene families and some of the cysteine-rich motifs; (9, 16, 19-21)) but the majority of motifs were newly identified in our analysis. The motifs have large ranges in length, frequency of occurrence, and rate of evolution. Many of the motifs occur in many ORFs, and a few are multiply repeated within individual ORFs.

This paper provides the first computational method to identify evolutionarily-conserved sequence motifs in the low-complexity part of a proteome. The result is the discovery of a set of relatively short sequence motifs that comprise a large fraction of the total length of the genes, consistent with the idea that these short motifs may be fundamental “building blocks” for fungal cell wall proteins. This is supported by how closely related the sequences in each motif family are, and by the prevalence of motifs in the sequences of fungal cell wall proteins. Thus, any theory for the evolutionary origin of cell wall proteins must account for the prevalence of these motifs. Detailed analyses of the structure and function of specific motifs will lead to further insights into structure and evolution of the wall proteins.

Acknowledgements

This work was supported by NIH program grants from RCMI (G12 RR030307) and NIGMS-SCORE (S06 GM060654) to Hunter College.

Literature Cited

1. **Alba, M. M., R. A. Laskowski, and J. M. Hancock.** 2002. Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics **18:**672-8.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-402.
3. **Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock.** 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:**25-9.
4. **Brendel, V., P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, and S. Karlin.** 1992. Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci U S A **89:**2002-6.
5. **Caro, L. H., H. Tettelin, J. H. Vossen, A. F. Ram, H. van den Ende, and F. M. Klis.** 1997. In silicio identification of glycosyl-phosphatidylinositol-anchored plasma-membrane and cell wall proteins of Saccharomyces cerevisiae. Yeast **13:**1477-89.
6. **Coronado, J. E., O. Attie, S. L. Epstein, W. G. Qiu, and P. N. Lipke.** 2006. Composition-Modified Matrices Improve Identification of Homologs of Saccharomyces cerevisiae Low-Complexity Glycoproteins. Eukaryotic Cell **5:**628-37.
7. **De Groot, P. W., K. J. Hellingwerf, and F. M. Klis.** 2003. Genome-wide identification of fungal GPI proteins. Yeast **20:**781-96.
8. **Dwight, S. S., M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry.** 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res **30:**69-72.
9. **Ecker, M., R. Deutzmann, L. Lehle, V. Mrsa, and W. Tanner.** 2006. Pir Proteins of Saccharomyces cerevisiae Are Attached to beta-1,3-Glucan by a New Protein-Carbohydrate Linkage. J Biol Chem **281:**11523-9.
10. **Embley, T. M., and W. Martin.** 2006. Eukaryotic evolution, changes and challenges. Nature **440:**623-30.
11. **Gumbel, E. J.** 1958. Statistics of extremes. Columbia University Press, New York,.
12. **Heer, J., S. K. Card, and J. A. Landay.** 2005. Presented at the Conference on human factors in computing. Portland, OR.

13. **Higgins, D. G., J. D. Thompson, and T. J. Gibson.** 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol **266:**383-402.
14. **Huang, J. Y., and D. L. Brutlag.** 2001. The EMOTIF database. Nucleic Acids Res **29:**202-4.
15. **Lipke, P. N., and J. Kurjan.** 1992. Sexual agglutination in budding yeasts: structure, function, and regulation of adhesion glycoproteins. Microbiol Rev **56:**180-94.
16. **Mrsa, V., T. Seidl, M. Gentzsch, and W. Tanner.** 1997. Specific labeling of cell wall proteins by biotinylation. Identification of four covalently linked O-mannosylated proteins of Saccharomyces cerevisiae. Yeast **13:**1145-54.
17. **Pearson, W. R.** 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol **132:**185-219.
18. **Steenkamp, E. T., J. Wright, and S. L. Baldauf.** 2006. The protistan origins of animals and fungi. Mol Biol Evol **23:**93-106.
19. **Verstrepen, K. J., A. Jansen, F. Lewitter, and G. R. Fink.** 2005. Intragenic tandem repeats generate functional variability. Nat Genet **37:**986-90.
20. **Verstrepen, K. J., and F. M. Klis.** 2006. Flocculation, adhesion and biofilm formation in yeasts. Mol Microbiol **60:**5-15.
21. **Verstrepen, K. J., T. B. Reynolds, and G. R. Fink.** 2004. Origins of variation in the fungal cell surface. Nat Rev Microbiol **2:**533-40.
22. **Wainright, P. O., G. Hinkle, M. L. Sogin, and S. K. Stickel.** 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. Science **260:**340-2.
23. **Yu, Y. K., and S. F. Altschul.** 2005. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics **21:**902-11.