Synergistic Computational and Experimental Proteomics Approaches for More Accurate Detection of Active Serine Hydrolases in Yeast

An analysis of the structurally and catalytically diverse serine hydrolase protein family in the *Saccharomyces cerevisiae* proteome was undertaken using two independent but complementary, large-scale approaches. The first approach is based on computational analysis of serine hydrolase active site structures; the second utilizes the chemical reactivity of the serine hydrolase active site in complex mixtures. These proteomics approaches share the ability to fractionate the complex proteome into functional subsets. Each method identified a significant number of sequences, but 15 proteins were identified by both methods. Eight of these were unannotated in the *Saccharomyces* Genome Database at the time of this study and are thus novel serine hydrolase identifications. Three of the previously uncharacterized proteins are members of a eukaryotic serine hydrolase family, designated as Fsh (family of serine hydrolase), identified here for the first time. OVCA2, a potential human tumor suppressor, and DYR SCHPO, a dihydrofolate reductase from *Schizosaccharomyces pombe*, are members of this family. Comparing the combined results to results of other proteomic methods showed that only four of the 15 proteins were identified in a recent large-scale, “shotgun” proteomic analysis and eight were identified using a related, but similar, approach (neither identifies function). Only 10 of the 15 were annotated using alternate motif-based computational tools. The results demonstrate the precision derived from combining complementary, function-based approaches to extract biological information from complex proteomes. The chemical proteomics technology indicates that a functional protein is being expressed in the cell, while the computational proteomics technology adds details about the specific type of function and residue that is likely being labeled. The combination of synergistic methods facilitates analysis, enriches true positive results, and increases confidence in novel identifications. This work also highlights the risks inherent in annotation transfer and the use of scoring functions for determination of correct annotations. *Molecular & Cellular Proteomics* 3:209–225, 2004.

Development of large-scale proteomics technologies for analysis of genes and proteins and their functions is a major focus of post-genomic biology. mRNA expression monitoring using gene chips and protein expression analysis using two-dimensional (2D) PAGE are powerful and widely used technologies for characterizing biological systems and pathways. The power of these techniques is demonstrated, for example, by the use of transcript profiling to classify cancer subtypes (1–4). However, these technologies also exhibit some limitations. Even in the relatively simple yeast system, the issue of mRNA transcript and protein level correlation is actively debated (5, 6). 2D PAGE experiments, which identify proteins directly, are limited by resolving power, both in extremes in mass (greater than 100 kDa or smaller than 15 kDa) and isoelectric point (greater than 10 or lower than 3). 2D PAGE also suffers from the general inability to resolve useful quantities of low-abundance proteins (7). To overcome some of these issues, large-scale analysis linking 2D liquid chromatography with tandem mass spectrometry (LC-MS/MS) (8–10) and immunodetection methods have been developed (11). Such technologies represent a significant improvement in proteomic analysis on a large scale.

Proteome analysis must be followed by the function annotation or characterization of each expressed protein, information that is at the core of biological understanding and is essential in the pharmaceutical industry for development of small molecule inhibitors. Many proteins identified by large-scale proteomics are likely to have functions previously unknown or uncharacterized. The assignment of function to a protein cannot be made until further analysis and characterization of the protein by a variety of methods has been performed. The ability to fractionate the complex proteome into functional subsets is crucial for these purposes.

© 2004 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org
Synergistic Proteomics Approaches Identify Serine Hydrolases

scale proteomics methods cannot be assigned a biochemical function. For example, a recent proteomics analysis of the rice proteome identified 2528 unique proteins in leaf, root, and seed tissue. Basic sequence-based approaches to functional classification of these proteins showed that the most abundant group (31.8%) belonged to a protein family of unknown function or exhibited low sequence identity to proteins of known function (12). Additional approaches are necessary to further determine the functional class and functional state of relevant components of the proteome.

Approaches aimed at functional analysis of proteomes are being developed. These include, for example, computational methods utilizing sequence comparison (13, 14), methods focused on functional site analysis (15–17), methods identifying protein–protein interactions (18), chemical proteomics approaches aimed at tagging functional sites on a large scale (19–22), and metabolomic methods (23). To overcome limitations of individual analyses and to provide a more accurate and precise functional analysis, we have combined synergistic computational and chemical proteomics approaches to fractionate the well-studied yeast proteome into functional subsets with high confidence.

In this work, we focused on the identification of serine hydrolases in yeast. Serine hydrolases are of interest because of their range of biological activities and because they are targets of several pharmaceutical agents. Serine hydrolases are present in all organisms and are active in diverse cellular compartments and functions. This class of enzymes includes proteases involved in the coagulation cascade (24); amidases responsible for the metabolism of endogenous signaling molecules (25); penicillin-binding proteins responsible for antibiotic sensitivity (26); and carboxylesterases involved in the metabolism of pharmaceuticals (27). Current drugs targeted against specific human serine hydrolases include Angiomax® for cardiovascular disease, Xenical® for obesity, and Aricept® and Cognex® for Alzheimer’s disease, as well as drugs in development for diabetes, arthritis, and cancer. Serine hydrolases are highly regulated and often present in low abundance, characteristics that present significant challenges to current methods of proteomic analysis. In addition, serine hydrolase activity is exhibited by enzymes distributed across most International Union of Biochemistry and Molecular Biology Enzyme Classification (EC) classes and is found in a wide variety of tertiary structures (Fig. 1), enzymatic functions (Table 1), and mechanisms. Active site diversity gives rise to several mechanisms that lower the pK_a of the key catalytic, nucleophilic serine. Both Ser-His-Asp catalytic triads and Ser-Lys catalytic dyads (28, 29), arranged in a particular three-dimensional configuration, can carry out serine hydrolase activity. Any method for proteome-scale analysis of serine hydrolases must adequately handle this mechanistic and structural diversity, thus this system was chosen as a difficult challenge for our combined proteomics methods.

The yeast Saccharomyces cerevisiae was chosen as a model organism because of the high quality of available genomic information and the relative ease of biochemical and genetic manipulation in this system. It has served as a useful model for demonstration of other proteomics methods (5, 6, 8, 10, 11, 18, 23, 30–37).

We independently applied structural and chemical proteomics technologies to identify yeast proteins that exhibit serine hydrolase function and then compared the results. Both of these function-based proteomics methods identify a large number of proteins. Fifteen serine hydrolase proteins are identified by both methods, and these are designated as high-confidence annotations. About half of these high-confidence identifications are known serine hydrolases, and about half are annotated here for the first time. Remarkably in this well-studied genome, the combined whole-proteome methodologies uncover a family of serine hydrolases in yeast not previously recognized, which we designate as Fsh (family of serine hydrolases). The results of this study demonstrate the utility of combining independently two complementary and synergistic function-based approaches to produce a more accurate analysis of complex proteomes.

**EXPERIMENTAL PROCEDURES**

**Materials**—Yeast strains were obtained from either the American Type Culture Collection (Manassas, VA) or Research Genetics (Huntsville, AL). Acid-washed glass beads were obtained from Sigma (St. Louis, MO), and most chemicals were obtained from Fisher (Pittsburgh, PA) and used without further purification. For computational analyses, 6530 full-length protein sequences from the S. cerevisiae genome were downloaded from the National Center for Biotechnology Information (NCBI) website (www.ncbi.nlm.nih.gov/). An additional 416 unique sequences were downloaded from the Saccharomyces Genome Database (SGD; www.yeastgenome.org). Sequence annotations were taken from SGD at the time the study was performed. It is not possible to identify all true positive serine hydrolases in the yeast genome to calculate absolute true and false positive identifications. Thus, all sequences identified by both methods are presented in Table II and analyzed according to the SGD annotations at the time of the study.

**Growth and Lysis of S. cerevisiae** Activity-based Probe (ABP) Labeling—For analytical scale analysis (such as in Fig. 3), yeast cultures were grown overnight in YPD (1% yeast extract, 2% peptone, 2% dextrose). One microliter was sedimented and resuspended in 100 μl phosphate-buffered saline (PBS) with 100 μl of glass beads, followed by vortexing twice for 1 min. After cell lysis, membrane proteins were solubilized with 0.05–0.1% Triton X-100. Insoluble material was sedimented by centrifugation, and soluble extract was labeled with 2 μM ABP (described in “Results” and Refs. 19 and 20) for 30 min at room temperature, followed by quenching (see below).

For purification and mass spectral identification of proteins, yeast were grown under the appropriate conditions, sedimented, and resuspended in 100 mM Tris, pH 9.1, 10 mM dithiothreitol (DTT). Following incubation at 30 °C for 20 min, the cultures were sedimented and resuspended in PBS. The yeast cells were lysed by high-pressure homogenization, using two passes through an Emulsiflex C5 (Avestin, Ottawa, Canada) at 10,000–15,000 psi. The resulting extract was subjected to sequential centrifugation at 15,000 and 100,000 × g. The supernatants were labeled with 4 μM ABP for 1 h at room temperature. The pellets were resuspended in PBS and 4 μM ABP was added.
After 30 min, Triton X-100 was added to 0.05% final concentration, and the labeling was allowed to proceed for another 30 min then quenched.

Purification and LC-MS/MS of Labeled Proteins in the Yeast Proteome—To identify as many serine hydrolases as possible, yeast cultures were grown in four different media: ideal (YPD, 2% dextrose), aerobic oxidation (YP with 2% galactose and 0.5% lactate), anaerobic fermentation (YP with 3% ethanol), and sporulation (1% KOAc). After growth under each condition, the yeast cells were lysed, centrifuged first at 15,000 × g, then at 100,000 × g. For each growth condition, proteins from both fractions of the high-speed spin only were labeled as described above with either a biotin-containing or a tetramethylrhodamine-containing ABP. After the ABP labeling, the reactions were quenched by addition of solid urea (6 M final concentration), followed by sequential treatment with DTT (10 mM final concentration) and iodoacetamide (40 mM final concentration) to reduce and alkylate free cysteines, respectively. After gel filtration to remove urea, DTT, and iodoacetamide, the labeled samples were subjected to affinity chromatography using either avidin agarose (Sigma) or an anti-rhodamine monoclonal antibody-agarose (prepared at ActiX). The resins were washed with buffer containing 1% SDS. Eluted proteins were separated by one-dimensional SDS-PAGE, and labeled proteins were excised and in-gel digested with trypsin following standard protocols (38). Tryptic peptides were analyzed using a ThermoFinngan (San Jose, CA) LCQ Deca XP and either Sequest or Mascot software, essentially as previously described (39). Results from the four growth conditions were combined. To control for the appearance of abundant proteins nonspecifically bound to the affinity matrices, parallel experiments were conducted wherein the ABP-labeling step was omitted. Proteins identified in the control experiments were subtracted from those identified in the ABP-labeling experiments.

Serine Hydrolases Fuzzy Functional Forms (FFFs)—A set of serine hydrolase FFFs, structural motifs for identification of functional sites, was used to identify proteins in the S. cerevisiae proteome. As described in previous work (40, 41), physicochemical and structural data from Protein Data Bank (PDB) entries are combined with activity information from the biochemical literature to identify the key functional residues. Each FFF is defined by the following criteria: one or a small number of residue identities for each key residue, a set of geometric descriptors describing the relative orientation of the key residues, and the allowed variability (a standard deviation) for each geometric descriptor. As previously described (15, 41), a standard cross-validation training procedure creates each FFF to uniquely recognize the true positive structures. In this work, the resulting serine hydrolase FFFs were sensitive enough to discriminate between known serine hydrolase functional sites and all other proteins in a test database of 12,009 PDB structure files (PDB, release 9/2).

The set of serine hydrolase template structures and FFFs were selected based on the following criteria (20): 1) the FFF describes a function requiring a nucleophilic serine; and 2) the FFF describes protease, lipase, esterase, amidase, or transacylase enzymatic activity. In addition, the flavin adenine dinucleotide-independent (FDH) dehydrogenase, lipase, esterase, amidase, or transacylase enzymatic activity requires a nucleophilic serine; and 2) the FFF describes a function requiring a nucleophilic serine, a characteristic Ser-His-Asp catalytic triad, and an α/β fold. The catalytic triad includes an ALW motif in space (illustrated in Fig. 1) are extracted from each structure and their sequences are aligned using CLUSTALW (46, 47). This alignment of the fragments from the active site vicinity in known structures is termed an active site profile for a given function or FFF. For each predicted functional site, the local fragments around the FFF-identified active site residues are extracted, aligned with the active site profile from the structures known to exhibit the function, and scored against these active site profiles. Each local score position in the functional site profile is scored by identity, conservation, and the presence of a gap. For a gap-free alignment, the score varies from 0 to 1. When gaps are introduced into predicted functional site profiles, the score can fall below zero. High confidence function annotations have functional site profile scores greater than 0.25 (45).

Function Identification Using Motif Databases—By definition an FFF serves as a template of the underlying chemical functionality of a protein, so equivalencies can be defined between FFFs and public tool motifs that describe the same or a related function; thus, motif equivalencies were established between FFFs and Pfam, BLOCKS and PRINTS motifs. The threading/FFF results were compared with the results obtained using three sequence motif databases: PRINTS 20.0 (48, 49); Pfam version 6.0 (50, 51); and BLOCKS (52, 53). These databases receive a sequence as input and output a list of sequence motifs ranked by score that may match the function of the query sequence. The top 10 hits by PRINTS and all query sequences above cutoff scores of 10 for Pfam and 5 for BLOCKS were analyzed to determine if the motifs identified a function equivalent to the FFF- assigned function. In addition, BLAST (54, 55) was used to assign function based on annotation transfer. Function assignment is inferred from sequences with similarity to the query sequence. For this study, a cutoff value of 0.01 was used, to ensure that this analysis identified distantly related sequences.

RESULTS

FFF-based Identification of Serine Hydrolase Functional Sites—FFFs were developed to identify functional sites in protein structures (15). One of the most powerful aspects of the FFF technology is its ability to identify functional sites accurately in both experimentally determined and computationally modeled protein structures (41). Another advantage of the FFF technology is that it does not rely on function annotation transfer based on global sequence alignment. Key functional residues are specifically identified in the protein structure, regardless of overall global sequence similarity to any other protein exhibiting the same function. This feature has allowed identification of similar functional sites in proteins with different overall architectures and low overall sequence similarity (56, 57).
For this study, FFFs were created to describe and identify the diversity of serine hydrolase functional sites (such as the examples in Fig. 1). A common feature of these FFFs is the inclusion of a nucleophilic or active serine. The library utilized in this study contained 35 different serine hydrolase FFFs, including six composite FFFs (Table I). Composite FFFs are defined when more than one FFF describes a functional family or subfamily, a feature that allows identification of multiple biochemical activities within a functional site. For example, XE3.4.12 is a composite FFF composed of two individual FFFs: the serine protease/serine carboxypeptidase family catalytic site (E3.4.33) and the serine carboxypeptidase family pH regulatory site (E3.4.39) (Table I). In cross-validation studies, the FFFs in the serine hydrolase library were shown to uniquely identify serine hydrolase functional sites in experimentally determined structures (see “Experimental Procedures”).

The set of 35 FFFs describe a diverse group of 25 different serine hydrolase functions (Table I). A discrepancy in the number of FFFs, 35, versus the number of different serine hydrolase functions, 25, exists because an enzyme function, as defined by the EC system, can be described by more than one FFF. For example, lipase (serine hydrolase defined by EC 3.1.1.3) catalytic sites can differ between bacterial and fungal organisms in structure and/or sequence, as illustrated by FFFs E3.1.58.1 (a bacterial lipase catalytic site) and E3.1.17 (fungal triacylglycerol lipase catalytic site). In these instances, two FFFs are necessary to describe the structural motifs that carry out this single EC-defined function. A common fold associated with serine hydrolases is the α/β fold family (58). Some serine hydrolases assume the α/β structure; however, many do not. Further, not all α/β hydrolases are serine hydrolases. FFFs can distinguish α/β hydrolases that exhibit a serine hydrolase function from those proteins that fold similarly to an α/β hydrolase but exhibit another function altogether (56). Sixteen FFFs in the library describe serine hydrolase function in an α/β hydrolase fold, including a single “family” FFF that is designed to recognize all the serine hydrolase proteins with α/β hydrolase fold (E1.11.2) (Table I). There are 19 FFFs that describe serine hydrolases with a fold other than the α/β hydrolase fold, including four composite FFFs (Table I).

To predict how successful the computational functional identification or annotation might be, we wanted to estimate the FFF coverage of the total currently defined structural space and the total serine hydrolase biological functional space. “Known serine hydrolase structural space” is based only on structures available in the Research Collaboratory for Structural Bioinformatics (RCSB) PDB (version 092). Approximately 63% of the total serine hydrolase structural space available at the initiation of the study was covered by this set of FFFs (Fig. 2). The FFFs used in this study describe several serine hydrolase subclasses, including serine proteases, serine lipases, serine esterases, serine transacylases, and (S) hydroxynitrile lyases. The FFF coverage of each of these subclasses, based on known structural space, ranges from 55 to 100%, with the exception of the serine amidases (Fig. 2). The serine amidase subclass FFF is conspicuously missing because of the limited structural coverage of serine hydrolase functional space at the time of this study.

Although the coverage of known structural space is relatively high, the coverage of biological functional space is not nearly as complete, because the libraries of experimentally determined structures, on which the FFFs are built, are still quite limited. To estimate coverage of biological functional space, we focused on the serine protease subclass of serine hydrolases because this subclass is relatively well studied. FFFs belonging to the serine protease subclass are estimated to cover explicitly only 23% of the total serine protease functional space. To calculate this crude approximation, total functional space is defined as the number of serine protease functions currently classified or defined by EC classes. The estimate is approximate because it contains the underlying, and inaccurate, assumption that there is a one-to-one corre-
lation between the number of EC classes and the number of biological functions. The number of 23% represents a lower bound, because family FFFs can identify members of subfamilies that are not explicitly represented in the PDB structure database.

Activity-based Labeling of Yeast Serine Hydrolases—ABPs have been developed that are able to interact specifically with active serine hydrolases in complex protein mixtures, including whole cells (for reviews, see Refs. 59 and 60). One of the most powerful aspects of the ABP technology is that it efficiently fractionates the proteome based on chemical reactivity, not on protein abundance. Because of the ABP's ability to label the functional subset of the proteome, simple separation methods, such as one-dimensional gel electrophoresis, are able to resolve the bulk of the proteins of interest (Fig. 3, for example). In general, ABPs contain three subunits: a) a react-

| Table I | FFFs used in this study |
|-----------------|-----------------------|
| Serine hydrolase proteins with an α/β hydrolase fold |
| E1.11.2αβ | Serine hydrolase/α/β hydrolase subset catalytic site |
| E3.1.13α | Carboxylesterase catalytic site |
| E3.1.17α | Fungal triacylglycerol lipase catalytic site |
| E3.1.18αβ | Acetylcholinesterase catalytic site |
| E3.1.35α | Bile activated lipase catalytic site |
| E3.1.44α | Fungal lipase B catalytic site |
| E3.1.45α | Culinase catalytic site |
| E3.1.55α | Pancreatic lipase catalytic site |
| E3.1.56αβ | Gastric lipase catalytic site |
| E3.1.58.1α | Pseudomonas sp. and Chromobacterium viscosum lipase catalytic site |
| E3.1.58.2αβ | Streptomyces exfoliates lipase catalytic site |
| E3.4.3αβ | Prolyl oligopeptidase family catalytic site |
| E3.4.8αβ | Prolyl aminopeptidase catalytic site |
| E3.4.33αβ | Serine protease-serine carboxypeptidase family catalytic site |
| E4.1.3αβ | Hydroxynitrile lyase catalytic site |
| XE3.4.12αβ | Serine protease—serine carboxypeptidase family |
| E3.4.33αβ | Serine carboxypeptidase family catalytic site |
| E3.4.39αβ | Serine carboxypeptidase family pH regulation site |
| Serine hydrolase functions with a fold other than an α/β hydrolase fold |
| E2.3.5αβ | Malonyl-CoA acyl carrier protein transacylase catalytic site |
| E3.1.7αβ | Platelet-activating factor acetylhydrolase catalytic site |
| E3.1.38α | Bacterial outer membrane phospholipase A catalytic site |
| E3.1.51αβ | CheB methyltransferase catalytic site |
| E3.4.17αβ | Dengue virus NS3 serine protease catalytic site |
| E3.4.24αβ | Viral serine protease catalytic site |
| E3.4.25αβ | Tonin catalytic site (inactive form) |
| E3.4.38αβ | Serine protease—subtilisin family catalytic site |
| E3.4.42αβ | Exfoliative toxins A and B catalytic site |
| E3.4.44αβ | Complement factor D catalytic site |
| E3.4.55αβ | Herpesvirus family serine protease catalytic site |
| E3.5.7αβ | Class A α-lactamase catalytic site |
| E3.5.8αβ | Class C α-lactamase catalytic site |
| E3.9.1αβ | Signal peptidase I catalytic site |
| XE3.4.14αβ | Ak.1 protease |
| E3.4.38α | Serine protease—subtilisin family catalytic site |
| M1.5.7αβ | Subtilisin enzyme family Ca2 binding site |
| M1.5.8αβ | Subtilisin enzyme family Ca3 binding site |
| M1.99.7αβ | Subtilisin enzyme family Na1 or Ca4 binding site |
| XE3.4.15αβ | Thermitase |
| E3.4.38α | Serine protease—subtilisin family catalytic site |
| M1.5.7αβ | Subtilisin enzyme family Ca2 binding site |
| M1.99.7αβ | Subtilisin enzyme family Na1 or Ca4 binding site |
| XE3.4.16αβ | Proteinase K |
| E3.4.38α | Serine protease—subtilisin family catalytic site |
| M1.5.10αβ | Proteinase K Ca2 binding site |
| M1.5.9αβ | Proteinase K Ca1 binding site |
| XE3.4.18αβ | Hepatitis C virus NS3 protease |
| E3.4.24αβ | Viral serine protease catalytic site |
| M1.10.31αβ | Hepatitis C virus NS3 protease Zn binding site |
| XE3.4.19αβ | Tonin catalytic site (inactive form) |
| E3.4.25αβ | Tonin catalytic site (inactive form) |
| M1.10.32αβ | Tonin Zn binding site |

α FFFs that identified serine hydrolases in S. cerevisiae proteins.

β FFF numbering conventions as follows: E identifies an enzyme function; M identifies a metal or other cofactor binding function; X identifies a composite FFF made of two or more component FFFs.
tive moiety specific for an amino acid in the active site of enzymes of a particular class, b) a linker, and c) a tag that enables visualization or purification of probe-modified proteins. For the experiments performed here to identify the active site serine hydrolases from the yeast proteome, the reactive moiety was a fluorophosphonate, related to the broad specificity serine hydrolase inhibitor diisopropyl fluorophosphonate, and the tag was either biotin or tetramethylrhodamine (19, 20). An example of the activity profile obtained from the different centrifugation fractions is shown in Fig. 3. The pellet from the low-speed spin contains mostly unbroken cells and large fragments. The proteins identified in the supernatant from the low-speed spin were identical to proteins found in the fractions obtained from the high-speed spin. Thus, only the fractions from the high-speed spin were used in subsequent experiments. Comparison of the proteins identified using the avidin affinity column and the antibody affinity column shows that avidin affinity column binds more proteins. Abundant or sticky proteins, i.e. those that bind readily to either avidin or antibody columns without the ABP treatment, are also readily identified by other methods (8) (Fig. 3). Output from the two affinity chromatography methods were combined to generate the results described here.

To demonstrate that these ABPs label yeast proteins in an activity-dependent manner, whole-cell yeast extracts were crudely fractionated by differential centrifugation, first at 15,000 x g, then at 100,000 x g. Portions of the pellet (P) and supernatant (S) fractions from these centrifugations were labeled with two different ABPs and analyzed by LC-MS/MS. A, Example profiles from different fractions separated using one-dimensional SDS-PAGE. Lanes 1 and 2 are the ABP-labeled profiles of the pellet and supernatant fractions from the 15,000 x g spin, lanes 3 and 4 are the ABP-labeled profiles of the pellet and supernatant fractions from the 100,000 x g spin. B, Comparison of numbers of proteins identified by affinity chromatography using avidin agarose or an anti-rhodamine monoclonal antibody-agarose. Black bars are those proteins identified in the work described here; gray bars are those proteins identified by 2D LC-MS/MS (8) and by the ABP labeling work described here.

We performed another set of experiments to demonstrate that the ABP labels serine hydrolases specifically. In these experiments, deletion strains of S. cerevisiae were used to show that labeled bands in the one-dimensional gel experiment correspond to unique, active serine hydrolases. After
The same protein was identified in multiple experiments. Proteins belonging to the other population, though not identified in the control experiments, gave weaker mass spectrometry results (only one peptide from the protein or identified in only one experiment) and may not have been modified with the ABP. The 57 proteins identified by this lower-quality data are listed in a footnote to Table II.

Of the 23 proteins identified by ABP labeling with high-quality mass spectrometry data, eight were previously annotated as hydrolases (Dap2, Kex1, Ppe1, Prb1, Prc1, Ste13, Yhl068c, and Amd2; Table II). One additional protein, Fas2, was previously annotated as a 2-oxo-acyl carrier protein reductase/synthase and encodes the α subunit of yeast fatty acid synthase. (This function, fatty acid synthase, was also identified computationally by the Prospector threading algorithm, see results below.) The function annotation in SGD (at the time of this study) for the other 14 ABP-identified proteins is “function unknown” (Table II), and these experimental results alone now suggest the presence of a nucleophilic serine at the functional site in these 14 proteins.

Computational Identification of Serine Hydrolases in the Yeast Proteome—To analyze the yeast proteome computationally, the set of serine hydrolase FFFs were applied to the proteins encoded by the S. cerevisiae genome, as described in “Experimental Procedures.” Briefly, threading alignments for each yeast amino acid sequence were generated using the Prospector threading algorithm (44), and confidence in each threading alignment was determined by a Z-score. FFFs were then applied to the top 20 threading alignments to identify the function(s) and active site(s) of each protein. The combination of a structure prediction and an FFF-based functional assignment for any sequence identified a putative serine hydrolase. Confidence in this function assignment was determined by calculating an active site profile score (45).

Overall, 19 individual hydrolase FFFs (Table I) identified 146 serine hydrolase protein sequences in the yeast genome (Table II and footnotes). Ten serine hydrolase FFFs did not hit any yeast sequences (Table I). Both component parts of one composite serine hydrolase FFF, XE3.4.12, hit two S. cerevisiae sequences, while the other five composites did not hit any sequences. Fifty-two of the 146 sequences were hit by more than one FFF. In these cases, both the protein family FFF and a more specific serine hydrolase FFF identified the functional site. For example, Ybr139w was identified by FFFs E1.11.2, E3.4.33, and E3.4.39 (serine hydrolase/α/β family, serine protease/serine carboxypeptidase family catalytic site, and serine carboxypeptidase family pH regulatory site, respectively; Table II). These multiple hits add confidence in the function assignment because the FFF technology recognizes active site structural and chemical features of both the family and a subclass of proteins.

Z-scores and active site profile scores were calculated for each sequence annotated by a serine hydrolase FFF (Figs. 5 and 6). Z-scores are a quantitative measure of the confidence...
| Gene name | M<sub>r</sub> (kDa) | Codon bias | FFF number<sup>a</sup> | FFF function<sup>a</sup> | Threaded PDB<sup>b</sup> | Z-score<sup>c</sup> | ASP score<sup>d</sup> | SGD function<sup>h</sup> |
|-----------|------------------|------------|------------------|-------------------------|-------------------|-----------------|-----------------|------------------|
| **Proteins identified by ABP labeling (high confidence) and by serine hydrolase FFFs** | | | | | | | | |
| Dap2 (Yhr028c) | 93 | 0.07 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1qfmA | 2.5 | 0.32 | Dipeptidyl aminopeptidase B |
| Kex1 (Ygl203c) | 82 | 0.04 | E3.4.3 | Prolyl oligopeptidase | 1qfmA | 2.8 | 0.32 | |
| E3.4.33 | Serine protease/serine carboxypeptidase family catalytic site | 1vyA | 13 | 0.85 | Carboxypeptidase D |
| E3.4.39 | Serine carboxypeptidase family pH regulatory site | 1whA | 14 | 0.91 | |
| Ppe1 (Yhr075c) | 45 | 0.03 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1a8q | 3.5 | 0.23 | Carboxy methyl esterase |
| Prb1 (Yel060c) | 30 | 0.45 | E3.4.38 | Serine protease/subtilisin family catalytic site | 2pkc | 8.6 | 0.82 | Cerevisin (Vacuolar protease B) |
| Proc1 (Ymr297w) | 60 | 0.38 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1ysc | 21.5 | 0.98 | Carboxypeptidase C |
| Ste13 (Yor219c) | 107 | 0.07 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1qfmA | 6.1 | 0.28 | Dipeptidyl aminopeptidase B |
| Yjl068c | 34 | 0.23 | E3.4.3 | Prolyl oligopeptidase | 1qfmA | 6.1 | 0.29 | |
| E3.4.33 | Serine protease/serine carboxypeptidase family catalytic site | 1ysc | 21.5 | 0.98 | Carboxypeptidase C |
| E3.4.39 | Serine carboxypeptidase family pH regulatory site | 1ysc | 21.5 | 0.98 | |
| Ynu3 (Ykl094w) | 36 | 0.09 | E3.4.24 | Viral serine protease | 6.9 | 0.21 | Unknown |
| Ybr139w | 58 | 0.09 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1qfmA | 21.4 | 0.86 | Unknown |
| E3.4.39 | Serine carboxypeptidase family pH regulatory site | 1qfmA | 21.4 | 0.86 ||
| Ybr204c | 43 | 0.00 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1c4xA | 1.9 | 0.28 | Unknown |
| Yhr049w | 27 | 0.52 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1auA | 4.8 | 0.45 | Unknown |
| E3.1.13 | Carboxylesterase catalytic site | 1auA | 4.7 | 0.46 | |
| Ylr118c | 25 | 0.05 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1auA | 8.2 | 0.59 | Unknown |
| E3.1.13 | Carboxylesterase catalytic site | 1auA | 8.2 | 0.59 | |
| Ymr222c | 25 | 0.05 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1auA | 3.7 | 0.49 | Unknown |
| E3.1.13 | Carboxylesterase catalytic site | 1auA | 3.7 | 0.49 | |
| Yor280c | 30 | 0.12 | E1.11.2 | Serine hydrolase/α/β hydrolase subset | 1frA | 1.8 | 0.36 | Unknown |
| E3.1.58.2 | Bacterial lipase | 1.8 | 0.36 | |
| **Proteins identified by ABP labeling (high confidence) and by FFFs other than serine hydrolase FFFs** | | | | | | | | |
| Ygl039w | 38 | 0.17 | E5.1.3 | UDP-galactose-4-epimerase | 1xf | 3.8 | 0.25 | Unknown |
| E1.1.3 | Estradiol-17-β dehydrogenase catalytic site | 1bhs | 2.1 | 0.38 | |
| E1.1.16 | 3α, 20-β-hydroxysteroid dehydrogenase catalytic site | 1hdcA | 1.6 | 0.27 | |
| Ygl157w | 38 | 0.18 | E5.1.3 | UDP-galactose-4-epimerase catalytic site | 1xf | 3.7 | 0.17 | Unknown |
| E1.1.13 | Carbonyl reductase catalytic site | 1cydA | 2.6 | 0.12 | |
| E1.1.3 | Estradiol-17-β dehydrogenase catalytic site | 1bhs | 1.8 | 0.39 | |
| E1.1.15 | Sepiapterin reductase catalytic site | 1sep | 1.4 | 0.29 | |
in a global sequence alignment between a yeast sequence and a serine hydrolase whose structure has been determined. The active site profile score, on the other hand, quantifies the similarity between the sequence and a structurally determined serine hydrolase only in the region of the functional site (45). Without an FFF annotation, a Z-score greater than or equal to 5.0 is considered significant for a threading alignment profile. Of eight proteins previously annotated as serine hydrolases by SGD (and identified by ABP labeling), only four (Kex1, Prb1, Prc1, and Ste13) align to known serine hydrolase structures with a Z-score greater than 5. Three others (Dap2, Ppe1, and Yjc068c) exhibit insignificant Z-scores of 2.8, 3.5, and 1.9, respectively (Table II). Amd2 did not align to a known serine hydrolase structure using this threading algorithm.

Fifty-two of the 146 FFF assignments exhibit active site profile scores greater than 0.25 (Fig. 6), and these are considered significant. Of the same eight proteins previously annotated as serine hydrolases at SGD (and identified by ABP labeling), six exhibit profile scores greater than 0.25 (Dap2, Kex1, Prb1, Prc1, Ste13, and Yjl068c; Table II). One of the eight serine hydrolases, Ppe1, exhibits a slightly less significant active site profile score of 0.23. Thus, Z-scores and active site profile scores identified six and four known serine hydrolases, respectively, with confidence. This comparison demonstrates the advantage of using active site profiling, in addition to the threading Z-score, because more known serine hydrolases can be confidently identified computationally.

Thirty-three of 52 FFF-identified proteins with significant active site profile scores (Table II and footnotes) were annotated as active site hydrolases by SGD or sequence databases. The following proteins were identified by FFF annotations as low confidence (profile score < 0.25) and were not labeled by an ABP: Bem3p, Cdc40p, Chl12p, Dat1p, HOP1, Lph1p, Lhp12p, Mak10p, Mih1p, Sam2p, Snf1, SSM4, Vps3p, Yal001c, Ypl095c, Ypr158w-B, Ypr140wp, Ypr147cp, Lpg13p, and gi887585. The following proteins were identified by FFF annotations as high confidence (profile score ≥ 0.25), but not by ABP labeling: Kex2 (Ynl238w), Ysp3 (Yor003w), Rnh1p, Yar009cp, Ybl089w, Ybl100c, Yol019W, Ycl061C(Mrc1), Ycr045c, Ydr303cp, Ydr422c, Ydr434wp, Ydr435cp, Ydr506cp, Yer049wp, Yer083cp, Yer127wp, Yfl007w, Yfr012w, Yfr054c, Ygl027c, Ygr031w, Yhl009w-b, Yhr159wp, Yil037c, Yil068c, Yil079c, Yil159w, Yjl157c, Yjr078w, Ykl014c, Ykr030w, Yll012w, Yll030c, Yll072w, Ylr020c, Ylr087c, Ylr098c, Yor343c-B, Ypl095c, Ypr158w-B, Ypr140wp, Ypr147cp, Lpg13p, and gi887585. The following proteins were identified by ABP labeling, mass spectrometry with low confidence: Acs1, Ade3, Ade5, Adk1, Ala1, Dal7, Dld3, Gal1, Lap3, Gar1, Gla1, Gin1, Gsp1, Gut2, Imd1, Nmd5, Kgd1, Leu1, Lys20, Lys4, Lys9, Mak16, Mes1, Nam9, Ncf1, Om45, Osh2, Pet9, Pfs2, Pmt2, Pre2, Pre5, Prs1, Ptk2, Rhr2, Rhr1, Rhr2, Rnr4, Sac1, Sam1, Scs2, Thr4, Tps1, Tps2, Tsa1, Ypt1, Ydi086w, Ygr017w, Yli041w, Yji107c, Yir149c, Yir179c, Ymr289w, Ynl092w, Ynl134c, Ypl044c, and Ypl105c.

The following proteins were identified by ABP labeling (high confidence), but not by FFFs:

- **Amd2** (Ydr242w)
- **Fas2** (Ypl231w)
- **Ydr428c**
- **Ynl123w**
- **MYor084w**

**TABLE II—continued**

| Gene namea | M (kDa) | Codon biasb | FFF numberc | FFF functiond | Threaded PDBd | Z-scoree | ASP scoref | SGD functiong |
|------------|---------|-------------|-------------|---------------|---------------|-----------|------------|---------------|
| Yml059c    | 187     | 0.16        | M3.1.4      | UDP-galactose 4-epimerase NAD binding site | 1xel          | 1.7       | −0.12      | Unknown       |

**Notes:**

- a Sequence name from SGD.
- b Measure of biased usage of synonymous codons.
- c Number used to designate the FFF.
- d Function that the FFF descriptor represents.
- e Z-score calculated for the alignment of the yeast sequence to the PDB structure.
- f Active site profile score.
- g Function annotation taken from SGD or sequence databases.
quences with significant profile scores, including Yar009cp, Ycl019w, and Ydr034c-d, have an SGD annotation of “protease” so the current computational analysis provides some additional information to support that annotation. A small number had previously been annotated, albeit not as serine hydrolases. Two of these, Ynl277w and Yfr027w, were annotated as acetyl transferases. Although, acetyl transferases were not specifically covered by the FFFs used in this study, the malonyl-CoA acyl carrier protein transacylase function was covered (E2.3.5; Table I), as this function is suggested to be a serine hydrolase (20). It is possible that the FFFs are correctly identifying an active site serine in these proteins. Four other proteins were annotated in SGD with other functions: Yjl045w, succinate dehydrogenase; Ynd055c, voltage-dependent ion-selective channel; Yor191w, DNA-dependent adenosine triphosphatase; and Ymr234w, Rnh1 or ribonuclease HI. The relationship between these annotations and the FFF-based annotations, if any, remains to be determined.

Comparison of Computational and Experimental Proteomics Methods: Serine Hydrolases Identified by ABP Labeling and FFFs—Under four expression conditions, ABP labeling identified 23 proteins with high-quality mass spectrometry data (Table II). If all of these are correct identifications, FFF analysis identified over 65% (15 of 23) as serine hydrolases (Table II). Based on estimates of FFF coverage of structural space (63%, Fig. 2) and functional space (23%), this is the expected result. Of the 15 proteins identified both by ABP labeling (high-quality mass spectrometry data) and by serine hydrolase FFFs, seven had been annotated at SGD prior to this work (Dap2, Kex1, Ppe1, Prb1, Prc1, Ste13, and Yjl068c; Table II). Eight sequences (over 50%) were previously annotated as “function unknown” or “hypothetical protein” at the time of this work (Eht1, Yju3, Ybr139w, Ybr204c, Yhr049c, Ylr118c, Ymr222c, and Yor280c; Table II). The combination of independently applied computational and experimental proteomics methods described in this paper allows confident assignment of serine hydrolase function to these proteins. The chemical proteomics technology indicates that a functional protein with an active serine is expressed in the cell. The computational proteomics technology adds details about the type of function, the structure of the functional site, and the specific residue that is likely labeled by the ABP. This combination of technologies adds significant knowledge about the family of serine hydrolases in the well-studied yeast organism.
Of the eight ABP-labeled proteins that were not annotated by serine hydrolase FFFs, one, Amd2, is annotated as a putative amidase in SGD. The lack of identification by an FFF is not surprising because no amidase FFF was available at the time of this study. Three of these eight ABP-labeled proteins (Ygl039w, Ygl157w, and Yml059c) were annotated by another common set of FFFs (Table II). These include FFFs covering the functions UDP-galactose-4-epimerase, estradiol-17β-dehydrogenase, and 3α, 20β-hydroxysteroid dehydrogenase. The FFFs for these three functions have a common active site tyrosine and serine, but these functions were not included in the serine hydrolase FFF library. The mass spectrometric method used here does not report the amino acid labeled by the ABP. Thus, computational FFF analysis serves to clarify the function of these ABP-identified proteins and suggests the specific residues that may be labeled.

Five proteins identified by ABP labeling and high-quality mass spectrometry data were not annotated by any FFFs (Table II). Three of these, however, did thread to proteins whose structures had previously been determined: Yor084w threaded to 1a8uA, Fas2 threaded to 1kas, and Ynl123w threaded to 1pysB, all with significant Z-scores. 1a8uA is the structure of cofactor-free chloroperoxidase T, a known serine hydrolase. In the threading alignment, the active site serine aligns with a serine in Yor084w, but neither the active site aspartic acid nor the histidine of 1a8uA are aligned with similar residues in Yor084w (data not shown); thus the FFF was unable to recognize this alignment. This protein is identified by the ABP, Prospector recognizes an overall similarity to chloroperoxidase T, and a potential active serine is recognized, so this protein may be a serine hydrolase. However, the alignment does not include a properly aligned active site that could be recognized by the complete FFF, so further experimentation would be required to understand the function of Yor084w. 1kas, to which Fas2 aligned, is the structure for β keto acyl ACP synthase. Fas2 is a known 3-oxo-ACP (acyl carrier protein) reductase/synthase (annotation provided by SGD), so Prospector easily recognized this homolog. No FFF has been constructed to recognize active sites in this protein. This protein is known to have active serines, one of which binds the pantetheine prosthetic group (62, 63), and the active serines may be the ABP binding site in this protein. Again, both methods individually identified these proteins, but the methods are synergistic and together provide additional information to aid in the interpretation of the results.

A New Family of Eukaryotic Serine Hydrolases Identified by a Combination of Chemical and Computational Proteomics Methods—Of the 15 proteins identified both by the ABP lab-
belonging and FFF analysis, eight were previously of unknown function (Table II). Three of these, Yhr049w, Ymr222c, and Yor280c, are related to each other by sequence similarity (Fig. 7) and appear to constitute a novel family of serine hydrolases found only in eukaryotic proteins. We propose to call these proteins Fsh1–3 (Yhr049w/Fsh1, Ymr222c/Fsh2, and Yor280c/Fsh3). To compare how other computational proteomics methods annotate these proteins, Pfam, BLOCKS, and PRINTS sequence motif methods were applied to Fsh1–3. None of these sequence motif tools annotated Yor280c/Fsh3 as any type of serine hydrolase.

BLAST, a sequence comparison tool (54, 55), was used to assess similarity among Yhr049w/Fsh1, Ymr222c/Fsh2, Yor280c/Fsh3, and related sequences. According to BLAST, the three Fsh sequences are related to each other with E-values more significant than 10^{-9}, but have less than 15.6% pairwise sequence identity between them. In addition, comparison of the three Fsh sequences to genomic sequences available at NCBI revealed other closely related (E-values less than 10^{-10}) proteins from several organisms, including proteins from Mus musculus, Caenorhabditis elegans, Arabidopsis thaliana, Drosophila melanogaster, S. pombe, and the malaria vector A. gambiae. Two additional sequences were identified as similar to OVCA2, but are not included in this alignment. The S. pombe sequence with Protein Information Resource accession number T43248 differs from the DYR_SCHPO sequence at only two positions and was therefore excluded. The A. thaliana sequence with GenBank accession number AAC24078.1 is identical to NP_563840.1 in the Fsh domain and was also excluded.

Synergistic Proteomics Approaches Identify Serine Hydrolases
Synergistic Proteomics Approaches Identify Serine Hydrolases

DISCUSSION

Synergies Provided by Complementary Proteomics Methods Is Key to Confident and Accurate Proteomic Analysis—To fully exploit the information provided by genomic analysis, large-scale, high throughput proteomics technologies and methods for data analysis must be developed. Currently, most large-scale methods generate interesting data, within a large amount of irrelevant data and a significant number of false positive findings. These shortcomings can be addressed using a combination of parallel and synergistic large-scale proteomic methods to facilitate analysis, enrich the true positive results, and increase confidence in the results.

In this study, a unique combination of computational, structural, and chemical proteomics methods was independently applied to identify active serine hydrolases from *S. cerevisiae*. The computational method utilizes structural information to identify functional sites in sequences. This method has the advantages of identifying specific functional sites and not relying on global sequence or structure alignment, but suffers from false positives resulting from inaccurate threading alignments and spurious alignment of putative functional residues. Computational methods also suffer limitations due to the use of scoring cutoffs, causing the loss of true positive results. The chemical proteomics method has the advantages of experimentally identifying functional sites in whole cells and distinguishing between functional and nonfunctional proteins. This ABP-based method, however, provides results that are specific to the conditions of the experiment. Additionally, the ABPs used in this study react with nucleophilic hydroxyl groups, whether they are a part of serines in serine hydrolases or reactive serines or tyrosines in other enzymes.

Independent application of these methods and comparison of the results provides unique insight into these advantages and disadvantages. Both methods generate a significant number of results that could not be confirmed by the other method (footnotes, Table II). These other identifications are not necessarily incorrect. For instance, several of the FFF-identified proteins with significant profile scores are annotated as peptidases or peptide hydrolases in SGD, including Kex2 and Ysp3. A protein may be identified by the computational method and not by the chemical proteomics method because the correct condition for expression of active protein was not probed or tested. Alternatively, a protein may be identified by the chemical proteomics method and not the computational method because the protein is a novel serine hydrolase whose structure or functional site has not previously been described. Thus, the proteins identified by only one method await confirmation by other methods.

Fifteen proteins, however, were identified by both computational and chemical proteomics methods, and these are designated as high-confidence identifications. Seven of the 15 proteins were previously identified as serine hydrolases by other methods and are confirmed by the current analysis. Eight of the proteins were previously unannotated in the SGD database, thus the combined methods add a significant amount of knowledge regarding the function of these proteins. Because of the combined approach, confidence in these designations is high. Within these eight previously unrecognized proteins, we have discovered a novel family of eukaryotic serine hydrolases, which we designate as Fsh. Three related members of this family, Fsh1, Fsh2, and Fsh3, were identified in *S. cerevisiae*. Surprisingly, the Fsh family member protein found in *S. pombe* is fused to DHFR. The fusion of a serine hydrolase domain to DHFR indicates a possible novel pathway in folate metabolism that requires coordinated function of a serine hydrolase with DHFR in *S. pombe* and perhaps other organisms, even though the domains are not covalently fused in these other organisms.
Synergistic Proteomics Approaches Identify Serine Hydrolases

The results are remarkable for the overlap, given the low coverage of serine hydrolase space by FFFs for computational proteomics and the inability to exhaustively test expression conditions for chemical proteomics. Given these practical limitations, the results demonstrate that both methods worked well and the synergies obtained from independent application of the two methods are significant.

Comparison of Combined Proteomics Analysis with Results from Other Proteomics Methods—Comparison of our combined proteomics methods with other experimental proteomics methods is difficult because experimental conditions and test sets are not the same. In addition, most technologies do not identify functions specifically. Yates and colleagues developed the MUDPIT technology linking 2D LC-MS/MS and performed a large-scale analysis of the yeast proteome (8, 9). This technology identified 1484 proteins expressed in yeast under one expression condition. Gygi and colleagues utilized multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS), with which they identified 7537 unique peptides and 1504 proteins under one expression condition (10). While these methods represent significant advances in analysis of complex proteomes, neither addresses the question of protein functionality. We can, however, determine how many of the serine hydrolases identified by the combined ABP and FFF technologies were also identified by these 2D LC-MS/MS technologies (Fig. 8). Of the fifteen sequences identified by the ABP/FFF technology, four were identified by Yates and colleagues (Kex1, Prc1, Eht1, and Yhr049w) and eight were identified by Gygi and colleagues (Prb1, Prc1, Yj1068c, Eht1, Ybr139w, Yhr049w, Ylr118c, and Yor280c).

Protein abundance continues to be an issue. A recent mass spectroscopic effort demonstrated excellent coverage for the most abundant proteins (>50,000 molecules per cell; coverage ~60%); however, for the 75% present at fewer than 5000 molecules per cell, only 8% of the proteins were observed (37). A recent report notes the issues with identifying low-abundance proteins in the cell and proposes an immunodepletion method to identify these proteins (11). The codon bias data presented in Fig. 8 demonstrate the advantage of using the ABP technology for identifying low-abundance proteins in the proteome. In addition, the combined ABP/FFF technologies provide critical information on the functionality and active site structure of the identified proteins.

We also compared the ability of the computational methods, FFFs and Pfam, to correctly annotate the proteins identified by ABP labeling. Of the 23 proteins identified by ABP labeling, we have already shown that FFFs identified 15. Pfam identified 10 of the 23 as serine hydrolases. Thus, the number of high-confidence identifications would be fewer if Pfam was used as the computational method. As described above, 14 of the 23 ABP-identified proteins were previously annotated as molecular function unknown. Of these 14 novel identifications by ABP labeling, FFFs identified eight and Pfam identified four as serine hydrolases. These results emphasize the synergies between the FFF and the ABP labeling technologies.

The results obtained by FFF and ABP labeling were compared with results obtained by other computational methods, including the local sequence signature databases and sequence-based function annotation tools (BLOCKS (52, 53), PRINTS (48, 49), and Pfam (50, 51)). A BLAST (54, 55) analysis was also used to assign function by sequence similarity to other annotated proteins in the NCBI GenBank nonredundant sequence database. Of the 146 FFF assignments, 87 were identified only by the FFF technology and not by any other computational tool. Thirteen of these novel hits (Yor280c, Ycl019w, Yol007c, Yhr134wp, Rnhlp, Yj1045w, Ynl182c, Ylr103c, Yor191w, Ybl089w, Ylr345w, Ypr147cp, and Yfr027w) have functional site profile scores greater than 0.25 (Fig. 6A, gray bars; Table II footnotes). One, Yor280c, was identified by ABP labeling. None of these novel hits have significant Z-scores (Fig. 5A, gray bars). This result emphasizes the similarity between what can be identified by public tools and by threading algorithms and also emphasizes the difference between these global alignment methods and what is identified by FFF analysis and active site profiling; however,
further experimentation is required to understand its implications.

**Analysis of High-confidence Identifications Provides Insight into the Limitations of Computational Scoring Methods**—The Z-score distributions for the threading alignments on the 23 ABP-identified proteins are shown in Fig. 5B. There is no correlation between Z-score and confidence in ABP-labeled proteins—the scores range from 1 to >20. About half of the high-confidence proteins have Z-scores greater than 5, and about half have Z-scores less than 5 (Fig. 5B). Z-scores less than five indicate statistically insignificant alignments—using only threading and this scoring statistic, there would be no confidence in the function identification for these proteins. Because Z-scores do not correlate with the high-confidence ABP hits, threading and similar methods that rely on the global alignment of sequences or structures are inadequate for assigning function between distantly related sequences.

On the other hand, methods that focus on functional sites themselves, such as FFF and active site profile analysis, correlate much better with the ABP-labeled proteins. Eighty percent of the high-confidence ABP-identified proteins have significant (greater than 0.25) active site profile scores (Fig. 6B). Use of a scoring function that focuses on the active site improves the correlation between computation and experiment and provides a better computational function annotation.

**A Cautionary Tale Involving Annotation Transfer Based on Sequence Alignment**—Function assignment is often based on annotation transfer when experimental evidence is unavailable; furthermore, annotation transfer based on sequence similarity is often applied in a high-throughput fashion, without manual curation. Two findings reported here highlight the risks of this approach, which has been pointed out by several other researchers (68, 69).

Using annotation transfer, the Fsh proteins might be assigned DHFR function, or at least be placed in the same protein family as DHFR, because of sequence similarity between the Fsh proteins and the N-terminal domain of DYR_SCHPO, a known DHFR. The data presented here suggests that DYR_SCHPO is a multifunctional protein, both an Fsh and DHFR, and the S. cerevisiae Fsh proteins presumably contain only one domain exhibiting serine hydrolase function. At the Yeast Proteome Database (www.proteome.com), Ymr222c/Fsh2 is predicted to have oxidoreductase function, presumably due to annotation transfer. Thus, the assignment of oxidoreductase function to Ymr222c/Fsh2 may be a faulty hypothesis and provides a cautionary example for post-genomic analyses.

A second such example can be found in another result from this study in which the candidate human tumor suppressor OVCA2 was found to have serine hydrolase function. Although the physiologic role of OVCA2 has not been determined, a recent study by Prowse et al. suggests a role in retinoid-induced growth arrest, differentiation, and apoptosis, and identifies homology with DHFRs (67). We show here that OVCA2 likely contains a serine hydrolase domain, but not a DHFR domain.

**CONCLUSION**

These results demonstrate the precision derived from combining independent, but complementary and synergistic, proteome-wide, function-based approaches to extract valuable biological information from complex proteomes. The chemical proteomics technology indicates that a functional protein with an active site serine is being expressed in the cell. The computational proteomics technology adds details about the specific type of function and the specific residue that is likely being labeled. About half of the proteins identified by the combination of methods had not been previously identified as serine hydrolases, thus the use of synergistic proteomics technologies confidently enhances knowledge about protein function in this well-studied organism. In addition, a previously unrecognized family of eukaryotic serine hydrolases was identified. The study emphasizes the risks of using an isolated analysis technique for protein function determination. In particular, it demonstrates that important information may be missed when using only protein sequence or structure similarity methods for function annotation. Instead, a combination of complementary, large-scale methods that provide different types of functional information can be used to extract valuable biological information that will help decipher protein function in complex pathways.

**Acknowledgments**—We thank John Kozarich for insightful conversations and encouragement, Matt Patricelli and Jane Wu for experimental facilitation, Dan Giang for help with early experiments, and Gabriela Tobal and Ruth Feldblum for assistance with the manuscript.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ Current address: National Center for Genome Resources, Santa Fe, NM 87505.

|| Current address: Departments of Physics and Computer Science, Wake Forest University, Winston-Salem, NC 27109.

** Current address: Science Applications International Corp., Biomedical Information Solutions Division, 10210 Campus Point Drive MS A2F, San Diego, CA 92121.

‡‡ Current address: Wadsworth Center for Laboratory Research, 120 New Scotland Avenue, Room 5046, Albany, NY 12201.

§§ Current address: The Scripps Research Institute, Department of Molecular Biology, 10550 North Torrey Pines Road, La Jolla, CA 92037.

¶¶ Current address: Wake Forest University, Departments of Physics and Computer Science, Winston-Salem, NC 27109.

¶ To whom correspondence should be addressed: Departments of Physics and Computer Science, 100 Olin Physical Laboratory, Wake Forest University, Winston-Salem, NC 27109-7507. Tel.: 336-758-4957; Fax: 336-758-6142; E-mail: fetrowjs@wfu.edu.
Synergistic Proteomics Approaches Identify Serine Hydrolases

REFERENCES

1. Balmain, A., Gray, J., and Ponder, B. (2003) The genetics and genomics of cancer. Nat. Genet. 33, (suppl) 238–244
2. Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloo, M., D’Amico, M., Pestell, R. G., West, M., and Nevins, J. R. (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat. Genet. 34, 226–230
3. Dressman, M. A., Baras, A., Malinowski, R., Alvis, L. B., Kwon, I., Walz, T. M., and Polymeropoulos, M. H. (2003) Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer. Cancer Res. 63, 2194–2199
4. Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. 24, 227–235
5. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. Mol. Cell Biol. 19, 1720–1730
6. Futschik, B., Latter, G. I., Monardo, P., McLaurin, C. S., and Garrels, J. I. (1999) A sampling of the yeast proteome. Mol. Cell. Biol. 19, 7357–7368
7. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc. Natl. Acad. Sci. U. S. A. 97, 9390–9395
8. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247
9. Washburn, M. P., Ulaszek, R., Deciu, C., Schietz, D. M., and Yates, J. R., 3rd (2002) Large-scale identification of abundant proteins. Mol. Cell. Proteomics. 1, 199–207
10. Lynch, M., and Aebersold, R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247
11. Isseroff, R. R., and Aebersold, R. (2001) Large-scale identification of abundant proteins. Mol. Cell. Proteomics. 1, 199–207
12. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using iso- tope-coded affinity tags. Nat. Biotechnol. 17, 984–999
13. Greenbaum, D., Baruch, A., Hayrapetian, L., Darula, Z., Burlingame, A., Medzhiradsy, K. F., and Bogoy, M. (2002) Chemical approaches for functionally probing the proteome. Mol. Cell. Proteomics. 1, 60–68
14. Allen, J., Davey, H. M., Broadhurst, D., Head, J. K., Rowland, J. J., Oliver, S. G., and Kel, D. B. (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. Nat. Biotechnol. 21, 692–696
15. Davey, E. W., Fujikawa, K., Kurachi, K., and Kiesel, W. (1979) The role of serine proteases in the blood coagulation cascade. Adv. Enzymol. Relat. Areas Mol. Biol. 48, 277–318
16. Patricelli, M. P., and Gravatt, B. F. (2001) Characterization and manipulation of the acyl chain selectivity of fatty acid amidase hydrolase. Biochemistry 40, 6107–6115
17. Pinho, M. G., de Lencastre, H., and Tomasz, A. (2001) An acquired and a native penicillin-binding protein cooperate in building the cell wall of drug-resistant staphylococci. Proc. Natl. Acad. Sci. U. S. A. 98, 10896–10901
18. Satoh, T., and Hosokawa, M. (1998) The mammalian carboxylesterases: From molecules to functions. Annu. Rev. Pharmacol. Toxicol. 38, 257–288
19. Dodson, G., and Wlodawer, A. (1998) Catalytic triads and their relatives. Trends Biochem. Sci. 23, 347–352
20. Paetzel, M., and Dalby, R. E. (1997) Catalytic hydroxyl/ammine dyads within serine proteases. Trends Biochem. Sci. 22, 28–31
21. Perrot, M., Saiag, F., Mini, T., Moniot, C., Schneider, U., Shevchenko, A., Mann, M., Jeno, P., and Boucherie, H. (1999) Two-dimensional gel protein database of Saccharomyces cerevisiae (update 1999). Electrophoresis 20, 2280–2288
22. Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Saiag, F., Wilms, M., Vorm, O., Mortensen, P., Boucherie, H., and Mann, M. (1996) Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two-dimensional gels. Proc. Natl. Acad. Sci. U. S. A. 93, 14440–14445
23. Garrels, J. I., McLaurin, C. S., Warner, J., Futschik, B., Latter, G. I., Kobayashi, R., Schwender, B., Wolpe, T., Anderson, D. S., Mesquita-Fuentes, R., and Payne, W. E. (1997) Proteome studies of Saccharomyces cerevisiae: Identification and characterization of abundant proteins. Electrophoresis 18, 1347–1360
24. Giaever, G., Chu, A. M., Nas, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lu, A., Dun-Rana, A., Anderson, K., Andre, B., Arkin, A. P., Astronoff, A., El-Bakkoury, M., Bangham, R., Benito, R., et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418, 387–391
25. Lu, L., Araikaki, A. K., Lu, H., and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the Saccharomyces cerevisiae proteome. Genome Res. 13, 1146–1154
26. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423, 241–254
27. Bader, G. D., and Hogue, C. W. (2001) Analyzing yeast protein-protein interaction data obtained from different sources. Nat. Biotechnol. 20, 991–997
28. Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R., Plouffe, D., Deciu, C., Winzeler, E. A., and Yates, J. R., 3rd (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. 100, 3107–3112
29. Hellman, U., Wernstedt, C., Gornez, J., and Heldin, C. H. (1995) Improvement of an “In-Gel” digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. Anal. Biochem. 220, 451–455
30. Stone, K. L., DeAngelis, R., LoPresti, M., Jones, J., Papov, V. V., and Williams, K. R. (1998) Use of liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) for routine identification of enzymatically digested proteins separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. Electrophoresis 19, 1046–1052
31. Di Cerbo, A. J., Siew, N., Hoffman, B. T., Zhang, L., Skolnick, J., Neilson, L. I., and Fetrow, J. S. (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. J. Struct. Biol. 134, 232–245
41. Fetrow, J. S., Godzik, A., and Skolnick, J. (1998) Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J. Mol. Biol. 282, 703–711
42. Zuegg, J., Gruber, K., Gugganig, M., Wagner, U. G., and Kratky, C. (1999) Three-dimensional structures of enzyme-substrate complexes of the hydroxynitrile lyase from Hevea brasiliensis. Protein Sci. 8, 1990–2000
43. Gruber, K., Gugganig, M., Wagner, U. G., and Kratky, C. (1999) Atomic resolution crystal structure of hydroxynitrile lyase from Hevea brasiliensis. Bioll. Chem. 380, 993–1000
44. Skolnick, J., and Khara, D. (2001) Defrosting the frozen approximation: PROSPECTOR—A new approach to threading. Proteins 42, 319–331
45. Cammer, S. A., Hoffman, B. T., Speir, J. A., Canady, M. A., Nelson, M. R., Knutson, S., Gallina, M., Baxter, S. M., and Fetrow, J. S. (2003) Structure-based active site profiles for genome analysis and functional family subclassification. J. Mol. Biol. 334, 387–401
46. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680
47. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gieben, T. J., Higgins, D. G., and Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31, 3497–3500
48. Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P., and Selly, J. (1998) The PRINTS protein fingerprints database in its fifth year. Nucleic Acids Res. 26, 304–308
49. Attwood, T. K., Beck, M. E., Bleasby, A. J., and Parry-Smith, D. J. (1994) PRINTS—A database of protein motif fingerprints. Nucleic Acids Res. 22, 3550–3562
50. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002) The Pfam protein families database. Nucleic Acids Res. 30, 276–280
51. Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins Struct. Funct. Gen. 28, 405–420
52. Henikoff, J. G., and Henikoff, S. (1996) Blocks database and its applications. Methods Enzymol. 266, 88–105
53. Henikoff, J. G., Pietrokovski, S., McCullum, C. M., and Henikoff, S. (2000) Blocks-based methods for detecting protein homology. Electrophoresis 21, 1700–1706
54. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410
55. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Shang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402
56. Zhang, L., Godzik, A., Skolnick, J., and Fetrow, J. S. (1998) Functional analysis of the Escherichia coli genome for members of the alpha/beta hydrolase family. Fold Des. 3, 535–548
57. Fetrow, J. S., Siew, N., and Skolnick, J. (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. FASEB J. 13, 1866–1874
58. Cousin, X., Hotelier, T., Giles, K., Toutant, J. P., and Chatonnet, A. (1998) aCheIDb: The database system for ESTHER, the alpha/beta fold family of proteins and the Cholinersterase gene server. Nucleic Acids Res. 26, 226–229
59. Campbell, D. A., and Szardening, A. K. (2003) Functional profiling of the proteome with affinity labels. Curr. Opin. Chem. Biol. 7, 296–303
60. Kozarich, J. W. (2003) Activity-based proteomics: Enzyme chemistry redux. Curr. Opin. Chem. Biol. 7, 78–83
61. Ogris, E., Du, X., Nelson, K. C., Mak, E. K., Yu, X. X., Lane, W. S., and Pallas, D. C. (1999) A protein phosphatase methylesterase (PME-1) is one of several novel proteins stably associating with two inactive mutants of protein phosphatase 2A. J. Biol. Chem. 274, 14382–14391
62. Schuster, H., Rautenstrauss, B., Mittag, M., Stratmann, D., and Schweizer, E. (1995) Substrate and product binding sites of yeast fatty acid synthase. Stoichiometry and binding kinetics of wild-type and in vitro mutated enzymes. Eur. J. Biochem. 228, 417–424
63. Mohamed, A. H., Chirala, S. S., Mody, N. H., Huang, W. Y., and Wakis, S. J. (1988) Primary structure of the multifunctional alpha subunit protein of yeast fatty acid synthase derived from FAS2 gene sequence. J. Biol. Chem. 263, 12315–12325
64. Bertani, L. E., and Campbell, J. L. (1994) The isolation and characterization of the gene (dfr1) encoding dihydrofolate reductase (DHFR) in Schizosaccharomyces pombe. Gene 147, 131–135
65. Schultz, D. C., Vanderveer, L., Berman, D. B., Hamilton, T. C., Wong, A. J., and Godwin, A. K. (1996) Identification of two candidate tumor suppressor genes on chromosome 17p13.3. Cancer Res. 56, 1997–2002
66. Huang, J., Hu, N., Goldstein, A. M., Emmert-Buck, M. R., Tang, Z. Z., Roth, M. J., Wang, Q. H., Dawsey, S. M., Han, Y. Y., Ding, T., Li, G., Giffen, C., and Taylor, P. R. (2000) High frequency allelic loss on chromosome 17p13.3-p11.1 in esophageal squamous cell carcinomas from a high incidence area in northern China. Carcinogenesis 21, 2029–2032
67. Prowse, A. H., Vanderveer, L., Milling, S. W., Pan, Z. Z., Dunbrack, R. L., Xu, X. X., and Godwin, A. K. (2002) OVCA2 is downregulated and degraded during retinoid-induced apoptosis. Int. J. Cancer 99, 185–192
68. Rost, B. (2002) Enzyme function less conserved than anticipated. J. Mol. Biol. 318, 595–608
69. Hegyi, H., and Gerstein, M. (2001) Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. Genome Res. 11, 1632–1640