



## Yeast Functional Analysis Report

# Associating protein activities with their genes: rapid identification of a gene encoding a methylglyoxal reductase in the yeast *Saccharomyces cerevisiae*

Ching-Nen Chen<sup>1</sup>, Larysa Porubleva<sup>2</sup>, Georgia Shearer<sup>1</sup>, Maja Svrakic<sup>1</sup>, Lauren G. Holden<sup>1</sup>, James L. Dover<sup>3</sup>, Mark Johnston<sup>3</sup>, Parag R. Chitnis<sup>2</sup> and Daniel H. Kohl<sup>1\*</sup>

<sup>1</sup>Department of Biology, Washington University, St. Louis, MO 63130, USA

<sup>2</sup>Department of Biochemistry and Biophysics, Iowa State University, Ames, IA 50011, USA

<sup>3</sup>Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA

\*Correspondence to:

Daniel H. Kohl, Biology  
Department, Washington  
University, St. Louis, MO 63130,  
USA.

E-mail: Kohl@biology.wustl.edu

## Abstract

Methylglyoxal is associated with a broad spectrum of biological effects, including cytostatic and cytotoxic activities. It is detoxified by the glyoxylase system or by its reduction to lactaldehyde by methylglyoxal reductase. We show that methylglyoxal reductase (NADPH-dependent) is encoded by *GRE2* (*YOL151w*). We associated this activity with its gene by partially purifying the enzyme and identifying by MALDI-TOF the proteins in candidate bands on SDS-PAGE gels whose relative intensities correlated with specific activity through three purification steps. The candidate proteins were then purified using a glutathione-S-transferase tag that was fused to them, and tested for methylglyoxal reductase activity. The advantage of this approach is that only modest protein purification is required. Our approach should be useful for identifying many of the genes that encode the metabolic pathway enzymes that have not been associated with a gene (about 275 in *S. cerevisiae*, by our estimate). Copyright © 2003 John Wiley & Sons, Ltd.

Received: 25 July 2002  
Accepted: 22 December 2002

**Keywords:** methylglyoxal reductase; MALDI-TOF mass spectrometry; *GRE2*; *YOL151w*; GST fusions

## Introduction

Keto-aldehydes, such as methylglyoxal, are associated with a broad spectrum of biological effects, including regulation of the cell cycle in diverse organisms (Murata *et al.*, 1985, and references therein), anti-IgE induced histamine release, microtubule assembly, inhibition of protein synthesis, and carcinostatic, as well as cytotoxic, activity (Carrington and Douglas, 1986). Interest in the biochemistry of methylglyoxal dates to the early part of the last century (Carrington and Douglas, 1986, and references therein).

Methylglyoxal is removed from the cell by two routes. In the first, the glyoxylase system (glyoxylase I and II; E.C. 4.4.1.5 and 3.1.2.5, respectively)

adds, and then eliminates, glutathione, converting methylglyoxal into D-lactate, a molecule at the same redox level. The second route is by the reduction of methylglyoxal to lactaldehyde (with either NADH or NADPH as electron donor), followed by the oxidation of lactaldehyde with NAD<sup>+</sup>. Since methylglyoxal is synthesized from the 3-carbon glycolytic intermediate, dihydroxyacetone phosphate, methylglyoxal conversion to lactate via either route results in a shunt around the glycolytic pathway, with a net loss of 2ATP/glucose.

Most of the interest in methylglyoxal metabolism has involved study of the glyoxylase system. Five enzymes have been reported to catalyse methylglyoxal reduction to lactaldehyde (E.C. 1.1.1.1,

1.1.1.2, 1.1.1.21, 1.1.1.72, 1.1.1.78; BRENDA website: <http://www.brenda.uni-koeln.de/homepage.html>) although not all are necessarily found in *Saccharomyces cerevisiae*, the organism used in this study. Here we report the identification of a gene that encodes NADPH-dependent reduction of methylglyoxal.

Our approach employed a combination of well-established methodologies (protein purification, MALDI-mass spectrometry and the use of GST-ORF fusions) to associate a protein of known function with its previously unidentified gene. Combining these methodologies relieved us of the necessity of obtaining a high degree of protein purification. As a result, we were able to employ simple and rapid protein purification methods, even if they may not result in maximum purification. Our approach should be useful for linking some of the protein activities that *S. cerevisiae* possesses that are not yet associated with a gene, even in well-studied metabolic pathways.

## Materials and methods

### Materials

All chromatographic resins and polybuffer (containing ampholytes) were purchased from Amersham Pharmacia (Uppsala, Sweden). The Bradford dye used for protein determinations was obtained from Bio-Rad (Hercules, CA). All other chemicals were obtained from Sigma (St. Louis, MO). Dried baker's yeast (*S. cerevisiae*) was from Sigma (YSC-1). Strains of yeast containing glutathione-S-transferase (GST) fusions to individual genes were obtained from Eric Phizicky (University of Rochester). All strains used were sequence-verified. In one case with an incorrect sequence (GST/YGL157w), the GST-fusion was constructed as described by Martzen *et al.* (1999).

### Partial purification of methylglyoxal reducing activity

Dried yeast cells (10 g) were autolysed in 100 ml 0.25 M Na<sub>2</sub>HPO<sub>4</sub> and 1 ml protease cocktail inhibitor for fungal and yeast extracts (Sigma, P8215). After 2.5 h in a rotary shaker at 30 °C, the extract was centrifuged at 12 000 × *g* at 4 °C for 10 min. The supernatant (crude extract) was subjected to the following protocols, depending

upon whether NADH or NADPH was to be used as electron donor.

### NADH-dependent activity

Protein in the crude extract was precipitated with graded concentrations of (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>. The material precipitating between 30–50% saturated (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub> was found to have the majority of the methylglyoxal-reducing activity. After redissolving this protein in 20 mM sodium phosphate buffer (pH 7.4) and desalting on a 2.5 × 12.5 cm Sephadex G-25 column, the sample (465 mg protein) was applied to a 1.5 × 27 cm anion exchange column (DEAE Sepharose CL-6B). Proteins were eluted with a linear NaCl gradient (0–0.3 M in 20 mM phosphate buffer, pH 7.4). The active fraction (151 mg protein) was applied to a chromatofocusing (PBE Polybuffer exchanger) column (1 × 17 cm) and eluted with polybuffer 74, which establishes a pH gradient of 7–4. The active fraction (2.8 mg protein) eluted between pH 5 and 5.5. The active fraction from each purification step was subjected to one-dimensional SDS-PAGE. Bands were visualized by staining with Coomassie blue R-250 (Merrill, 1990). The total activity, specific activity and relative specific activity in the active fractions are recorded in Table 1.

### NADPH-dependent activity

Alcohol dehydrogenase (NADP<sup>+</sup>), which has been reported (BRENDA website) to reduce methylglyoxal with NADPH, was separated from other methylglyoxal reducing activity by precipitating the protein in a crude extract (3610 mg protein) with graded concentrations of (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>. Each fraction was assayed for both methylglyoxal and acetaldehyde reduction. These two activities separated well (data not shown). The fractions showing methylglyoxal reducing activity, but little acetaldehyde reducing activity [ $> 60\%$  saturated (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>], were pooled for further purification. After desalting on a 2.5 × 12.5 cm Sephadex G-25 column, the sample (872 mg protein) was applied to the anion exchange column, as above. The active fraction (54 mg protein) was subjected to gel filtration (Sephadex G-100, 3 × 45 cm column). The active fraction contained 1.1 mg protein. Again, active fractions from each step were subjected to one-dimensional SDS-PAGE.

### Assays

Protein concentration was estimated by the method of Bradford (1976), using bovine serum albumin as standard. Methylglyoxal reduction was measured as follows. The assay mixture contained 0.1 mM NADPH or NADH, 20  $\mu$ l *S. cerevisiae* extract or partially purified enzyme, 10 mM methylglyoxal and 50 mM sodium phosphate buffer (pH 7) in a total volume of 1 ml. The rate of methylglyoxal reduction was calculated from the change in absorbance at 340 nm, followed on a Hitachi U-3110 recording spectrophotometer, after addition of the substrate. The rate of acetaldehyde reduction was determined in exactly the same way, except acetaldehyde was substituted for methylglyoxal.

### MALDI-TOF mass spectrometry

Excision of protein bands, in-gel trypsin digestion, elution of tryptic fragments and their analysis using MALDI-TOF mass spectrometry was performed according to previously published methods (Porubleva *et al.*, 2001). Spectra from 100 shots at several different positions were obtained in a reflection-delayed extraction mode over a mass range of 600–4000 Da and were combined to generate a peptide mass fingerprint for each protein sample. The peptide ions generated by autolysis of trypsin (with  $m/z$  842.51+ and 2211.10+) were used as internal standards for calibrating the mass spectra. The search in the databases was performed using mass spectrometry utilities (MSU) software (Porubleva *et al.*, 2001), developed to automate the use of the MS-Fit and PeptIdent protein identification tools. We used the following parameters for database search: mass tolerance of 30 ppm and no missed cleavages. The identity of a protein was considered affirmative if all of the following criteria were met: a high MOWSE score ( $>10^3$  for the MSFit program and  $>0.2$  for the PeptIdent program); matching masses for more than four peptides:  $<30$  ppm mass deviation between the experimental and theoretical; and  $>25\%$  sequence coverage by the matching peptides.

### Verification

The proteins encoded by the candidate genes identified by MALDI-TOF analysis were tested for NADPH-dependent methylglyoxal reducing activity using GST fusion proteins (Martzen *et al.*,

1999). This led to the identification of a gene whose protein product had this activity. This gene had three homologues. Their GST fusions were tested for NADPH-dependent methylglyoxal activity.

Cells containing Cu<sup>2+</sup>-inducible GST-fusions were grown as previously described (Martzen *et al.*, 1999). The cells were broken with glass beads (Catley, 1988), and the enzyme activity in crude extracts prepared before and after induction was determined. Proteins from crude extracts were purified by glutathione agarose chromatography (Martzen *et al.*, 1999), and the activity in the purified proteins was measured.

### Estimating the number of enzymes in *S. cerevisiae* with unassigned ORFs

The number of enzymes in yeast not yet associated with their gene was estimated using the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000; <http://www.genome.ad.jp/KEGG/kegg3.html>). Each metabolic pathway map has a generic and an organism-specific form. In the reference pathway, any reaction that has been reported for any organism is included in the map. When the organism is limited to *Saccharomyces cerevisiae* (from the pull-down menu), all of the E.C. numbers for enzymes whose *S. cerevisiae* gene is considered to be characterized are indicated by colour. The remaining E.C. numbers have not been associated with a gene for one of two reasons: (1) that the enzyme is not found in this yeast; or (2) the enzyme is in this yeast but its corresponding ORF has not been identified. We are interested in the latter class, which we call 'foundlings'.

When the information in all of the pathways is summed, there are 1942 enzymes. The genes for 793 of these have been characterized in *S. cerevisiae*, leaving 1149 potential yeast 'foundlings'. The number of 'foundling' enzymes (i.e. those known to be present in yeast, but whose gene has not yet been identified) is a subset of the 1149 potential 'foundlings'. The remainder are not present in yeast. To estimate the number of 'foundlings' in yeast, we selected 50 (see Table 5 for a list of the 50 enzymes) of the 1149 enzymes at random. (A sample size of 50 will give a 4–7% standard error (SE) of the estimate ( $SE = [PQ/n]^{1/2}$ , where  $n$  = sample size,  $P$  = fraction of  $n$  which was found in yeast and  $Q$  = fraction of  $n$  which was not found in yeast. Note that the

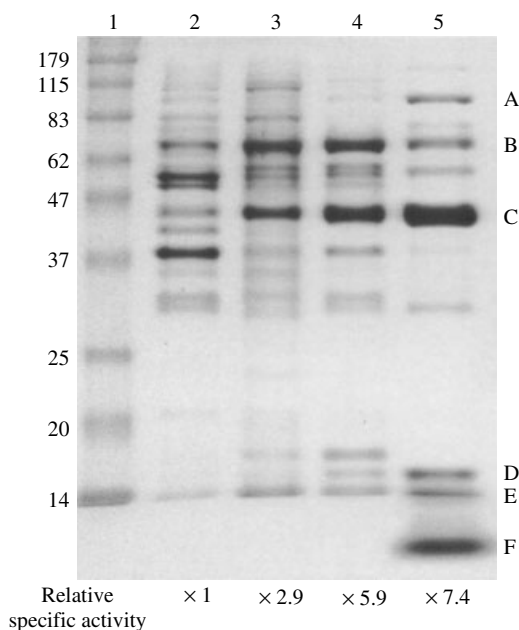
standard error of the estimate is independent of the size of the universe ( $N$ ) from which the sample is drawn, as long as  $n \ll N$ ). We used five sources to determine whether an enzyme had been reported in yeast. First, the 50 E.C. numbers were searched in *Chemical Abstracts* (<http://stnweb.cas.org> for [bakers yeast or *Saccharomyces*] but not [complementation or disease or promoter or tissue culture or two-hybrid or YAC or yeast artificial chromosome or yeast vector]). We identified the two earliest citations because we thought that a significant amount of early enzymology employed yeast, and thus we expected this literature to be a rich source of potential 'foundlings' (unfortunately, even though *Chemical Abstracts* has a very complete bibliography starting in 1907, it can only be searched by computer back to 1966). Second, we used *Enzyme Nomenclature* (International Union of Biochemistry, 1984 <http://www.chem.qmw.ac.uk/iubmb/enzyme/>), since the references in this source often predate 1966. We then checked the enzyme in BRENDA (see above), a comprehensive enzyme database developed and maintained by the Institute of Biochemistry at the University of Cologne. Finally we searched the *Saccharomyces* Genome Database (<http://genome-www.stanford.edu/Saccharomyces/>) and NCBI (<http://www.ncbi.nlm.nih.gov/>) to verify that no sequences had been reported in other databases.

## Results

Using the procedures described in Materials and methods, we estimated that *S. cerevisiae* possesses approximately 275 pathway activities that are not associated with a gene. Before attempting to identify the gene associated with a founding protein, we first used the strategy described in the introduction to identify a yeast gene encoding a protein with known NADH-dependent methylglyoxal-reducing activity.

### Identification of Adh1 as a methylglyoxal reductase

The degree of purification of NADH-dependent methylglyoxal reducing activity (Table 1) was modest: only 7.4-fold after three purification steps. Based on the proportional increase in relative band



**Figure 1.** SDS-PAGE gel of samples obtained at different purification steps of NADH-dependent methylglyoxal-reducing activity. Lane 1 contains a ladder of molecular weight standards. Lanes 2–5 were loaded with a 5  $\mu$ g protein (estimated by Bradford) aliquot of crude extract and the active fractions for the reduction of methylglyoxal by NADH from ammonium sulphate precipitation, anion exchange and chromatofocusing chromatography, respectively. The small molecular weight bands at E and F are associated with components of the polybuffer (data not shown). Relative specific activity refers to the relative rate of reduction of methylglyoxal by NADH per mg protein. See Table 1

intensity and specific activity (Figure 1), band C seemed most likely to contain a protein with methylglyoxal-reducing activity, but for completeness we subjected bands A, B, and D to peptide mass fingerprinting as well.

The tryptic peptide fragments from the predominant protein in band C matched alcohol dehydrogenase (Adh1, E.C.1.1.1.1) (Table 2). Adh1 has been reported (see BRENDA Website) to be able to reduce methylglyoxal with NADH. Thus, by our approach we quickly associated a protein of known function with its gene.

### Identification of GRE2 as a gene encoding methylglyoxal reductase

We used the same strategy to identify the (previously unknown) gene encoding NADPH-dependent methylglyoxal reducing activity. Since alcohol

dehydrogenase (EC 1.1.1.2) has been reported (see BRENDA Website) to have NADPH-dependent methylglyoxal-reducing activity in several organisms (in addition to reducing its usual substrate, acetaldehyde), methylglyoxal-reducing activity was separated from acetaldehyde-reducing activity by  $(\text{NH}_4)_2\text{SO}_4$  precipitation, as described in Materials and methods. The majority of the NADPH-dependent methylglyoxal-reducing activity, but virtually no acetaldehyde-reducing activity, remained in the supernatant at > 60% saturation of ammonium sulphate (data not shown). This supernatant served as the substrate for further purification, as detailed in Table 3. The degree of purification (120-fold) after three steps, although not nearly enough to yield a pure protein, was substantially greater than for NADH-dependent methylglyoxal reductase.

Band D (Figure 2, lane 5) is the best candidate for NADPH-dependent methylglyoxal reductase because the increase in its intensity seems best

correlated with the increase in specific activity in successive purification steps. This band was excised and subjected to *in situ* digestion with trypsin, followed by MALDI-mass spectrometry (for completeness, bands A, B, C and E were also subjected to the same treatment). The ORFs identified in this analysis are listed in Table 4. The peptide mass fingerprint of band D indicated that it contains Gre2, a putative reductase.

Three other *S. cerevisiae* ORFs (YDR541c, YGL039w and YGL157w) encode proteins that are similar to Gre2 (56–59% identity; 71–75% similarity; <http://www.proteome.com/databases/YPD/YPDsearch-long.html>). All three proteins are similar to the plant enzyme dihydroflavonol-4-reductase. Crude extracts of strains expressing Gre2 and three of its homologue as GST fusion proteins from the  $\text{Cu}^{2+}$ -inducible *CUPI* promoter were tested for NADPH-dependent methylglyoxal-reducing activity. Methylglyoxal-reducing activity is clearly induced by  $\text{Cu}^{2+}$  in the strain expressing

**Table 1.** Purification steps for NADH-dependent methylglyoxal-reducing activity

Sample	Total protein (mg) in active fraction	Total activity ( $\mu\text{mol}/\text{min}$ ) in active fraction	Specific activity ( $\mu\text{mol}/\text{min}/\text{mg}$ protein)	Relative specific activity
Crude extract	1435	297	0.207	1
$(\text{NH}_4)_2\text{SO}_4$ ppt	465	279	0.600	2.9
Anion exchange	151	184	1.22	5.9
Chromatofocusing	2.8	4.3	1.53	7.4

**Table 2.** Summary of MALDI-TOF results for proteins off SDS-PAGE gels from final purification step for NADH-dependent methylglyoxal-reducing activity

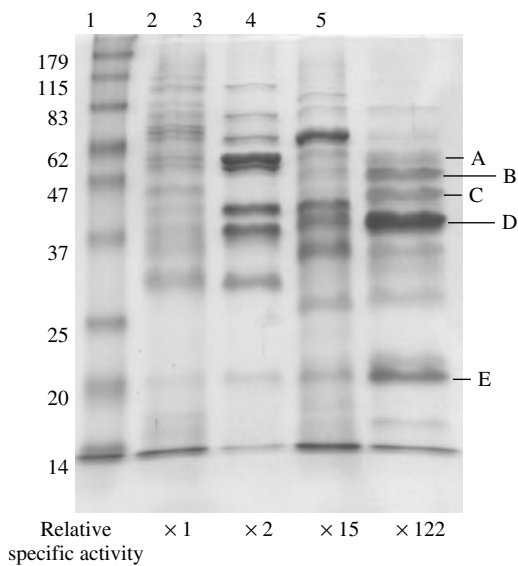
Band	MOWSE score	Masses matched	Sequence covered (%)	Protein [MW(Da)/pI]	Protein accession no.	Gene location	Gene name	Protein name or function
A	2.98E + 10	18/31	36	85860/6.07	6320936	YERO91c	MET6	Methionine synthase ( $\text{B}_{12}$ independent)*
B upper	1.27E + 08	17/43	42	61496/5.80	6323073	YLR044c	PDC1	Pyruvate decarboxylase I
	1.26E + 05	8/43	30	61581/5.80	6321524	YGR087c	PDC6	Third minor isozyme of PDC
B lower	1.14E + 08	19/47	49	61300/6.00	6319673	YBR196B	PGI1	G6P isomerase
C upper	9.95E + 04	11/21	43	36849/6.21	6324486	YOL086c	ADH1	Alcohol dehydrogenase I
C lower	8.54E + 08	13/36	44	36849/6.21	6324486	YOL086c	ADH1	Alcohol dehydrogenase I
D	2.35E + 05	9/36	43	31518/6.38	223142	YMR303c	ADH2	Alcohol dehydrogenase II
	0.3588**			11693/5.22	P22943	YFL014w	HSP12	12 kDa HSP

\* MSFit gives the protein NCBI Accession No. and identifies a functional name when known. NCBI (Entrez) and the *Saccharomyces* Genome Database was searched for gene location and gene name.

\*\* The protein was identified searching the SwissProt and TrEMBL databases with the PeptIdent program (<http://www.expasy.ch/tools/peptident.html>), which has a different algorithm of the MOWSE score calculation. Note the difference in scoring between MOWSE and PeptIdent. MOWSE scores  $< 10^3$  and PeptIdent  $< 0.2$  are not included in the table.

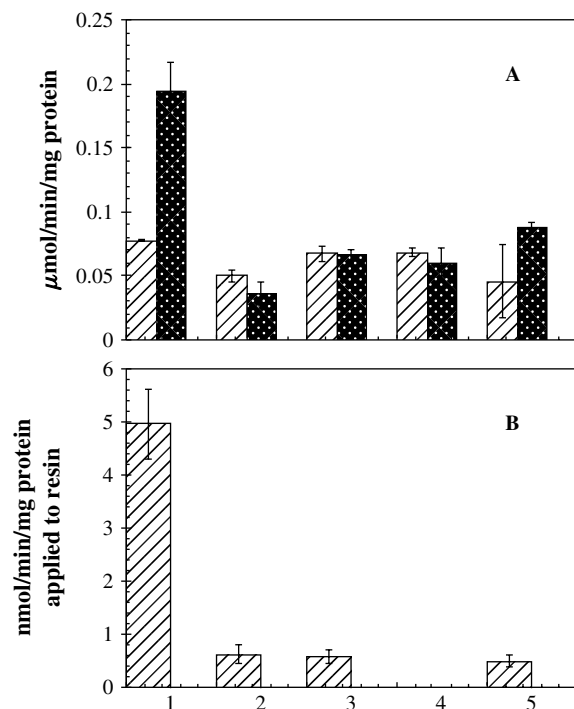
**Table 3.** Purification steps for NADPH-dependent methylglyoxal reducing activity

Sample	Total protein (mg) in active fraction	Total activity ( $\mu\text{mol}/\text{min}$ ) in active fraction	Specific activity ( $\mu\text{mol}/\text{min}/\text{mg}$ protein)	Relative specific activity
Crude extract	3610	247	0.0685	1
$(\text{NH}_4)_2\text{SO}_4$ ppt.	872	112	0.128	2
Anion exchange	54	56	1.03	15
Gel filtration	1.1	9	8.35	122

**Figure 2.** SDS-PAGE gel of samples obtained at different purification steps of NADPH-dependent methylglyoxal-reducing activity. Lane 1 contains a ladder of molecular weight standards. Lanes 2–5 were loaded with a 5  $\mu\text{g}$  protein (estimated by Bradford) aliquot of crude extract and the active fractions for the reduction of methylglyoxal by NADPH from ammonium sulphate precipitation, anion exchange chromatography and gel filtration, respectively

GST-Gre2/Yol151w (strain 1, Figure 3A; hatched vs. solid bars), but not in strains expressing the Gre2 orthologues Ydr541c, Ygl039w and Ygl157w. A strain expressing GST-Zwf1 was included as a control. It exhibited no  $\text{Cu}^{2+}$ -inducible NADPH-dependent methylglyoxal reducing activity, but it was the only strain exhibiting G6PDH activity (data not shown).

GST fusion proteins of Gre2 and two of its orthologues were enriched by glutathione affinity chromatography. Gre2 had the greatest methylglyoxal-reducing activity (Figure 3B). We conclude that GRE2/YOL151w, but not YDR541c, YGL039w or

**Figure 3.** NADPH-dependent methylglyoxal-reducing activity in cells containing GST fusions to the indicated ORFs (strain 1, YOL151w; strain 2, YDR541c; strain 3, YGL039w; strain 4, YGL157w; strain 5, YNL241c). (A) Activity in crude extracts of uninduced (hatched bars) and  $\text{Cu}^{2+}$ -induced (solid bars) cells. (B) Activity in purified GST fusion proteins. Purification was by glutathione agarose chromatography (Martzen *et al.*, 1999). A strain containing a GST/YNL241c (Zwf1) was included as a negative control. These strains were also assayed for G6PDH activity. Only the strain containing the GST-YNL241c fusion protein showed induction of this activity in the crude extract or any activity in the purified proteins

YGL157w, encodes a protein with methylglyoxal-reducing activity. Ynl134c was also a candidate for having this activity (band C, lower), but its GST fusion protein did not have NADPH-dependent methylglyoxal-reducing activity (data not shown).

**Table 4.** Summary of MALDI–TOF results for proteins off SDS–PAGE gels from final purification step for NADPH-dependent methylglyoxal reducing activity

Band	MOWSE score*	Masses matched	Sequence covered (%)	Protein [MW(Da)/pI]	Protein accession no.	Gene location**	Gene name**	Protein name or function**
A	757	5/29	10	68095/5.47	6321731	YGR292w	MAL12	Maltase
B	3.62E + 05	10/36	25	57474/5.46	6319901	YCR053w	THR4	Threonine synthase
C upper	3.16E + 08	12/39	37	44739/7.11	129930	YCR012w	PGK1	Phosphoglycerate kinase
	2.37E + 04	8/39	22	48452/8.98	609258	YOR209c	NPT1	Putative nicotinate phosphorybosyl transferase
C lower	6.82E + 05	8/31	30	41165/5.83	6324195	YNL134c	Unassigned	Unknown function
	7.73E + 04	7/31	37	37445/5.25	172713	YAL005c	SSA1	Heat shock protein
D	6.83E + 08	18/45	60	38170/5.78	6324421	YOL151w	GRE2	Putative reductase
E	3.28E + 04	8/26	44	19115/5.01	6233138	YLR109w	AHP1	alkyl hydroperoxide reductase

\* Mowse scores  $<10^3$  and Peptident  $<0.2$  are not included in the table.

\*\* Gene location and gene name were obtained as follows: amino acid sequence was obtained from NCBI (Entrez) (<http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?SUBMIT=y>) using protein Accession No. This sequence was applied to the SDG BLAST database (<http://genome-www.stanford.edu/cgi-bin/SGD/nph-blast2sgd>) to give gene location and name.

Of the purified GST–ORF fusion proteins, only GST–Zwf1 showed G6PDH activity (data not shown).

## Discussion

We have identified a previously uncharacterized gene, *GRE2*, whose protein product can reduce methylglyoxal with NADPH. Gre2 is similar to plant dihydroflavonol 4-reductase. The three Gre2 orthologues (Ydr541c, Ygl039w and Ygl157w) do not exhibit NADPH-dependent methylglyoxal-reducing activity.

Because our approach combines a verification method (purification and assay of a GST fusion protein) with traditional protein purification, the protein need not be highly purified in order to identify good candidates for the gene encoding the activity of interest. The criterion for a candidate protein is that the increase in its relative amount (as visualized by the relative intensity of the band on an SDS–PAGE gel) should parallel the increase in the specific activity being measured. Band C in Figure 1, lane 5, was the best candidate and, indeed, it proved to contain Adh1, a protein known to be capable of catalysing this reaction (see BRENDA website).

The degree of purification required to identify a gene depends on the abundance of the enzyme. Of

course, the more abundant the protein is, the easier it is to accomplish the task. The relatively small degree of purification required to identify the gene encoding Adh1 (7.4-fold) reflects the fact that it is very abundant. Futcher *et al.* (1999) documented a correlation in yeast between the Codon Adaptation Index (CAI; <http://www.proteome.com/databases/YPD/YPDsearch-long.html>) of a gene in yeast and the amount of its product ( $R = 0.8$ ). The CAI of *ADH1* (0.811) suggests that Adh1 is about 0.7% of total cell protein. Only 0.6% of all yeast proteins are predicted to be more abundant than Adh1.

The SDS–PAGE band containing Gre2 was not visible until after anion exchange chromatography resulted in 15-fold purification. At this level of purification, there was no obvious candidate band, but one additional purification step increased the intensity of one band approximately in proportion to the increase in specific activity. This band contained the likely candidate protein, which turned out to be encoded by *GRE2*.

Based on its CAI value (0.244) Gre2 is predicted to be much less abundant than is Adh1 — about 0.036% of cell protein — so it required somewhat more purification for its gene to be identified. Most (89%) of the 6080 yeast ORFs are predicted to be less highly expressed than Gre2. For the 927 ORFs that the KEGG database associates with pathway enzymes, 26% have CAIs corresponding to 0.036% abundance or greater. We infer from this

**Table 5.** List of pathway enzymes for which no ORF is assigned. We randomly selected 300 such enzymes of the 1149 given in KEGG. Below are listed the first 50 which met the criteria given in the text. They are arranged by E.C. number rather than by the randomized set of 300

E.C. No.		Name	Is there evidence for presence in yeast?	
1	1.1.1.64	Testosterone 17 $\beta$ -dehydrogenase (NADP)		No
2	1.1.1.79	Glyoxylate reductase (NADP)	Yes	
3	1.1.1.110	Indolelactate dehydrogenase		No
4	1.1.1.178	3-Hydroxy-2-methylbutyryl-CoA dehydrogenase		No
5	1.1.3.15	(S)-2-Hydroxy-acid oxidase	Yes	
6	1.3.1.24	Biliverdin reductase		No
7	1.3.2.3	Galactonolactone dehydrogenase		No
8	1.4.1.16	Diaminopimelate dehydrogenase		No
9	1.6.6.4	Nitrite reductase		No
10	1.6.99.8	Aquocobalamin reductase		No
11	1.14.99.11	Estradiol-6 $\beta$ -hydroxylase		No
12	2.1.1.5	Betaine-homocysteine-S-methyltransferase		No
13	2.1.1.47	Indolepyruvate-C-methyltransferase		No
14	2.1.2.4	Glycine formiminotransferase		No
15	2.3.1.47	8-Amino-7-oxononanoate synthase		No
16	2.3.1.62	2-Acylglycerophosphocholine-O-acyltransferase	Yes	
17	2.4.1.5	Dextranucrase		No
18	2.4.1.12	Cellulose synthase (UDP-forming)	Yes	
19	2.4.1.20	Cellobiose phosphorylase		No
20	2.4.2.23	Deoxyuridine phosphorylase		No
21	2.5.1.18	Glutathione transferase	Yes	
22	2.6.1.23	4-Hydroxyglutamate transaminase		No
23	2.6.1.37	2-Aminoethylphosphonate-pyruvate transaminase		No
24	2.6.1.39	2-Aminoadipate transaminase	Yes	
25	2.7.1.5	Rhamnulokinase		No
26	2.7.1.33	Pantothenate kinase		No
27	2.7.1.49	Hydroxymethylpyrimidine kinase	Yes	
28	2.7.1.69	Protein-N-p-phosphohistidine-sugar phosphotransferase		No
29	2.7.1.94	Acylglycerol kinase		No
30	2.7.4.1	Polyphosphate kinase	Yes	
31	2.7.9.1	Pyruvate, phosphate dikinase		No
32	3.1.3.24	Sucrose-phosphatase		No
33	3.2.1.39	Glucan endo-1,3- $\beta$ -D-glucosidase	Yes	
34	3.4.11.4	Tripeptide aminopeptidase	Yes	
35	3.5.1.3	$\omega$ amylase		No
36	3.5.1.16	Acetylmithine deacetylase	Yes	
37	3.5.3.10	D-arginase		No
38	3.5.5.2	Ricinine nitrilase		No
39	3.6.1.6	Nucleoside-diphosphatase	Yes	
40	3.6.2.1	Adenylylsulphatase		No
41	4.1.1.19	Arginine decarboxylase		No
42	4.1.1.36	Phosphopantothenoylcysteine decarboxylase		No
43	4.1.1.41	Methylmalonyl-CoA decarboxylase		No
44	4.1.3.20	N-Acylneuraminate-9-phosphate synthase		No
45	4.5.1.2	3-Chloro-D-alanine dehydrochlorinase		No
46	5.1.3.4	L-Ribulose-phosphate 4-epimerase		No
47	5.3.3.6	Methylitaconate $\Delta$ -isomerase		No
48	5.5.1.6	Chalcone isomerase		No
49	6.2.1.11	Biotin-CoA ligase		No
50	6.4.1.4	Methylcrotonoyl-CoA carboxylase		No

Table 5. Continued

E.C. No.	Name	Is there evidence for presence in yeast?
Enzymes rejected for failing to meet criteria		Failed criterion
1.1.1.187	$\gamma$ -guanidinolbuteraldehyde dehydrogenase	Not in Chemical Abstracts
1.2.1.54	$\gamma$ -guanidinol dehydrogenase	Not in Chemical Abstracts
1.3.99.1	Succinate dehydrogenase	Sequence given in GenBank
1.4.99.4	Aralkylamine dehydrogenase	Not in Chemical Abstracts
1.6.4.9	bis-Glutamylcystine reductase (NADPH)	Not in Chemical Abstracts
2.3.1.99	Quinate <i>O</i> -hydroxycinnamoyltransferase	Not in Chemical Abstracts
2.4.1.111	Coniferyl-alcohol glucosyltransferase	Not in Chemical Abstracts
2.7.7.1	Nicotinamide-nucleotide adenyltransferase	Sequence given in GenBank
3.5.1.62	Acetylputrescine deacetylase	Not in Chemical Abstracts
3.5.4-	(In a class which acts on cyclic amidines)	Does not have four numbers
3.6.1.43	Dolichyldiphosphatase	Accidentally omitted from search
4.99.1	(A miscellaneous subclass of lyases)	Does not have four numbers
6.3.1.-	[Acid-ammonia (or amine) ligases (amide synthases)]	Does not have four numbers

that pathway enzymes tend to be more abundant than the complete complement of yeast proteins. Thus, without further refinement, about a quarter of the pathway enzymes should be able to be associated with their gene as easily as was Gre2. Nevertheless, even if the abundance of a protein is as low as 0.01%, (about 3.5 times less abundant than Gre2), only about a 200-fold purification should be required to visualize the protein (200 ng) with Coomassie brilliant blue R-250 when 10  $\mu$ g total protein is loaded onto the SDS-PAGE gel [it is not possible to estimate the fraction of proteins that are less abundant than 0.01%, since the correlation of Futcher *et al.* (1999) is based on genes with a CAI  $\geq$  0.18, corresponding to 0.026% protein abundance]. An approximately 10-fold greater sensitivity could be achieved by using more sensitive protein dyes (such as Sypro<sup>®</sup> ruby), or using larger gels with longer stackers to enable loading of substantially more protein without significant loss of resolution.

We estimate that there are approximately 275 'foundling' pathway enzymes in yeast that have not been associated with a gene (see Materials and methods for details of how this was estimated). This is probably an underestimate, because the primary database we used, Chemical Abstracts, only allowed references back to 1966 to be searched by computer, possibly causing us to miss some earlier reports. In addition, some of the enzymes listed in the KEGG database as having been associated with a gene have been assigned on the basis of

homology rather than experimental demonstration. Our estimate of 275 foundlings, in yeast metabolic pathways is consistent with an estimate based on data available in the Integrated Genomics database ([http://wit.integratedgenomics.com/IGwit/CGI/org.cgi?org=SC&user=&where=summary\\_statistics](http://wit.integratedgenomics.com/IGwit/CGI/org.cgi?org=SC&user=&where=summary_statistics)) for *S. cerevisiae*, where it is reported that even with the use of sophisticated genome annotation tools, functions of putative metabolic pathway enzymes cannot be assigned or predicted for 249 yeast ORFs, and 87 known physiological functions cannot be assigned to ORFs. If the number of 'foundling' pathway enzymes is so substantial with an organism as well-studied and with such a small genome as yeast, their number is surely much higher for organisms with more complex proteomes.

The method we have described can be used to associate proteins of known function with their genes in any organism for which there is enough genomic sequence information to allow identification of ORFs by MALDI-TOF mass spectrometry. Since the activity of the protein product is verified using GST-ORF fusions, there is no need for high confidence in nominating candidate genes. The availability of GST-ORF fusions for virtually all yeast ORFs (Martzen *et al.*, 1999; Zhu *et al.*, 2001) greatly facilitates verification of tentative gene assignments. It seems unlikely that similar arrays will be constructed for many organisms with larger genomes, but it would be possible to construct the few GST-fusions necessary to evaluate

the handful of candidate ORFs identified as a result of the correlation between the increase in gel band intensity and the increase in the specific activity during the partial purification. Since the list of complete genome sequences is increasing, the usefulness of our approach should also increase. While this increase in the amount of available genome sequence has the tendency to stimulate the development of high-throughput methods, often with large up-front costs, the approach we employed can be applied on a small scale and should enable many scientists to identify the genes encoding 'foundling' proteins with which they have been working.

### Acknowledgements

PRC acknowledges grants from Iowa Corn Promotion Board and Plant Sciences Institute at Iowa State University. Work in MJ's laboratory was supported by funds provided by the James S. McDonnell Foundation. We thank Phillip Debnam for setting up the protein purification columns and establishing the protocols for their use. We thank Robert T. McFarland, Washington University Chemistry Librarian, for doing the searches in Chemical Abstracts. We thank Maria C. Costanzo, Senior Editor, Fungal Databases, ProteomeInc., for calling methylglyoxal reductase to our attention.

### References

Bradford M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing

- the principle of protein-dye binding. *Anal Biochem* **72**: 248–254.
- BRENDA Website. <http://www.brenda.uni-koeln.de/homepage.html>.
- Catley BJ. 1988. *Yeast: A Practical Approach*. IRL Press: Washington, DC.
- Carrington S, Douglas KT. 1986. The glyoxylase enigma — the biological consequences of a ubiquitous enzyme. *IRCS Med Sci* **14**: 763–768.
- Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. 1999. A sampling of the yeast proteome. *Mol Cell Biol* **19**: 7357–7368.
- Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 29–34.
- International Union of Biochemistry, Nomenclature Committee. 1984. *Enzyme Nomenclature*. Academic Press: Orlando, FL.
- Martzen MR, McCraith SM, Spinelli SL, et al. 1999. A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**: 1153–1155.
- Merrill CR. 1990. Gel staining techniques. *Methods Enzymol* **182**: 477–488.
- Murata K, Fukuda Y, Shimosaka M, et al. 1985. Phenotypic character of the methylglyoxal resistance gene in *Saccharomyces cerevisiae*: expression in *Escherichia coli* and application to breeding wild-type yeast strains. *Appl Environ Microbiol* **50**: 1200–1207.
- Porubleva L, Vander Velden K, Kothari S, Oliver DJ, Chitnis P. 2001. The proteome of maize leaves: use of gene sequences and EST data for identification of proteins with peptide mass fingerprints. *Electrophoresis* **22**: 1724–1738.
- Zhu H, Bilgin M, Bangham R, et al. 2001. Global analysis of protein activities using proteome chips. *Science* **293**: 2101–2105.