29. Luo, J. *et al.* Negative control of p53 by Sir2α promotes cell survival under stress. *Cell* **107,** 137–148 (2001).
30. Rodriguez, M. S., Desterro, J. M. P., Lain, S., Lane, D. P. & Hay, R. T. Multiple C-terminal lysine residues target p53 for ubiquitin-proteasome-mediated degradation. *Mol. Cell. Biol.* **20,** 8458–8467 (2000).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We thank R. Baer, R. Dalla-Favera, B. Tycko and T. Ludwig for critical discussions; we also thank many colleges in the field for providing antibodies, cell lines and plasmids, and other members of W.G.'s laboratory for sharing unpublished data and critical comments. This work was supported in part by grants from Avon Foundation, the Stewart Trust, the Irma T. Hirschl Trust and NIH/NCI to W.G., who is also a Leukemia and Lymphoma Society Scholar.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for material should be addressed to W.G. (e-mail: wg8@columbia.edu).

Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays

Finny G. Kuruvilla*, Alykhan F. Shamji*†‡, Scott M. Sternson*‡, Paul J. Hergenrother*§ & Stuart L. Schreiber*

* *Howard Hughes Medical Institute, Institute for Chemistry and Cell Biology, Bauer Center for Genomics Research, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA*
† *Department of Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA*

Small molecules that alter protein function provide a means to modulate biological networks with temporal resolution. Here we demonstrate a potentially general and scalable method of identifying such molecules by application to a particular protein, Ure2p, which represses the transcription factors Gln3p and Nil1p[1–3]. By probing a high-density microarray of small molecules generated by diversity-oriented synthesis with fluorescently labelled Ure2p, we performed 3,780 protein-binding assays in parallel and identified several compounds that bind Ure2p. One compound, which we call uretupamine, specifically activates a glucose-sensitive transcriptional pathway downstream of Ure2p. Whole-genome transcription profiling and chemical epistasis demonstrate the remarkable Ure2p specificity of uretupamine and its ability to modulate the glucose-sensitive subset of genes downstream of Ure2p. These results demonstrate that diversity-oriented synthesis and small-molecule microarrays can be used to identify small molecules that bind to a protein of interest, and that these small molecules can regulate specific functions of the protein.

The progress in identifying and expressing all human proteins[4] presents an opportunity to develop a small-molecule modulator for every protein function. Small-molecule approaches to study protein function have illuminated diverse fields of biology. Examples include tetrodotoxin, which enabled the dissection of the action potential[5], and agonists of peroxisome-proliferator-activated receptor-γ such as rosiglitazone, which illuminated the regulation of adipogenesis[6]. However, in most cases no small molecule that can modulate the function of a protein of interest is known, and there is currently no efficient method of identifying these biological probes. Using the example of the yeast protein Ure2p, we demonstrate a general two-step method that does not require a high-resolution structure or a previously characterized small molecule known to bind the protein. First, diversity-oriented synthesis is used to produce structurally complex and diverse small molecules efficiently. Second, the resulting compounds are screened for their ability to bind a protein of interest by using small-molecule microarrays, a technique for extremely high-throughput parallel-binding assays. Cell-based studies can subsequently determine which functions of the protein are modulated by each small molecule.

The yeast protein Ure2p has been widely studied in several different contexts. Ure2p is the central repressor of genes involved

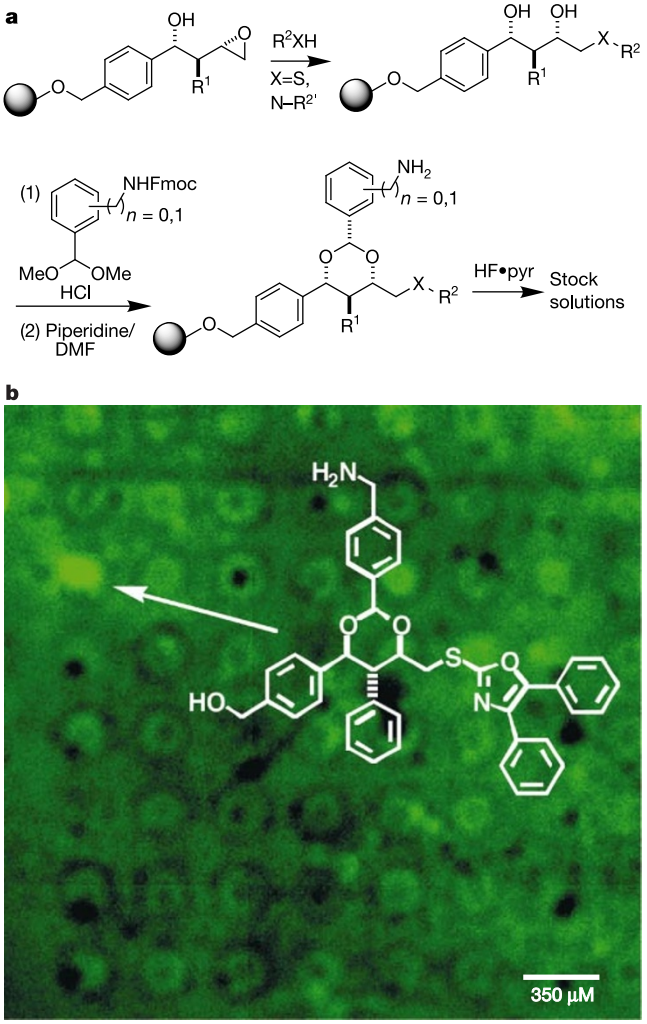


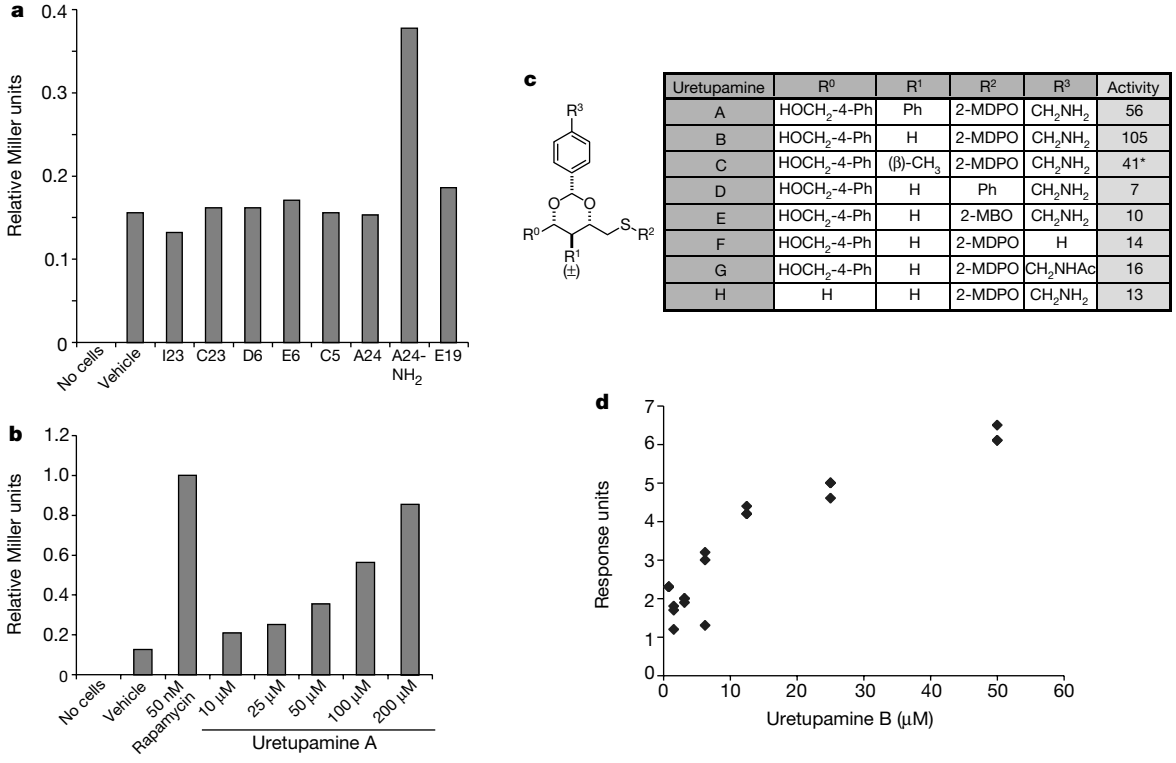
Figure 1 The library synthesis and identification of uretupamine. **a**, Outline of the diversity-oriented synthesis leading to uretupamine and other library members[11]. **b**, An expanded view of 64 compound spots on the 3,780-member small-molecule microarray (~800 spots cm^{-2}). Cy5-labelled Ure2p was passed over a microarray of the 1,3-dioxane small-molecule library, and the resulting slide was washed three times and scanned for fluorescence. The spot corresponding to uretupamine A is shown.

‡ These authors contributed equally to this work.
§ Present address: Department of Chemistry, University of Illinois, Urbana, Illinois 61801, USA.

in nitrogen metabolism[7], is capable of switching to a prion form[8], and is part of a signalling cascade downstream of the Tor proteins[9,10]. Because there is no known small molecule that binds to Ure2p, we screened a collection of 3,780 structurally complex 1,3-dioxane small molecules resulting from a diversity-oriented synthesis[11,12] (Fig. 1a). The molecules are structurally unbiased towards any particular protein target and can be used to identify specific probes for many different proteins. This collection of molecules had been prepared with a 'one bead–one stock solution' technology platform with the use of macrobeads (Fig. 1a), followed by automated compound cleavage and the generation of 5-mM stock solutions in 5 μl of *N*,*N*-dimethylformamide[11,13,14]. The small molecules were arrayed in high-density on glass slides (~800 spots cm^{-2}, 1 nl of each compound per slide) with a quill-pin contact-printing robot[15,16]. These microarrays were probed with fluorescently labelled Ure2p, enabling the protein-binding properties of each molecule to be tested in parallel with minimal protein consumption (protein solution: 20 μg ml^{-1}, 0.2 ml). This method has been used to detect known interactions such as that between FKBP12 and a synthetic pipecolyl α-ketoamide[15,16] and is applied here to the identification of novel small-molecule–protein interactions with uncharacterized compounds. Eight compound spots on the microarrays showed reproducible binding to labelled Ure2p (see Fig. 1b for one such spot in an 8 × 8 spot array where each spot is derived from a single-compound stock solution derived from the diversity-oriented synthesis).

To determine cellular activity, the molecules comprising these spots were resynthesized and tested for the modulation of endogenous Ure2p function with a *PUT1-lacZ* reporter system because *PUT1* expression is known to be repressed by *URE2* (ref. 17). In addition to the positive control of rapamycin[9], one of the eight compounds activated this reporter (Fig. 2a). The compound, which we named uretupamine A, gave a concentration-dependent dose response that at higher concentrations approached the levels of reporter gene activation induced by rapamycin (Fig. 2b).

To explore structure–activity relationships, we synthesized a series of compounds with systematic variations in the structure of uretupamine A (Fig. 2c). Specific atomic interactions are responsible for uretupamine binding because most modifications of its structure resulted in a complete loss of activity (Fig. 2c). Uretupamine A was rendered functionally inactive by acylation of the primary amine, replacement of the diphenyloxazole moiety with a phenyl group or a benzoxazole group or modification of the benzyl alcohol moiety (Fig. 2c). However, the C-5 position of the dioxane ring was tolerant to modification. Because the potency of uretupamine A was attenuated by poor solubility at higher concentrations in yeast medium, we synthesized a more soluble derivative, uretupamine B, which lacked the C-5 phenyl group on the dioxane ring



Uretupamine	R^0	R^1	R^2	R^3	Activity
A	$HOCH_2$-4-Ph	Ph	2-MDPO	CH_2NH_2	56
B	$HOCH_2$-4-Ph	H	2-MDPO	CH_2NH_2	105
C	$HOCH_2$-4-Ph	(β)-CH_3	2-MDPO	CH_2NH_2	41*
D	$HOCH_2$-4-Ph	H	Ph	CH_2NH_2	7
E	$HOCH_2$-4-Ph	H	2-MBO	CH_2NH_2	10
F	$HOCH_2$-4-Ph	H	2-MDPO	H	14
G	$HOCH_2$-4-Ph	H	2-MDPO	CH_2NHAc	16
H	H	H	2-MDPO	CH_2NH_2	13

Figure 2 Studies *in vivo*, dose–response and structure–activity relationships of uretupamine. **a**, A yeast strain (DB26-3A) growing in YPD medium expressing a *PUT1-lacZ* reporter was treated with 50 nM rapamycin or with a compound that had been detected to bind to labelled Ure2p on a small-molecule microarray. After 90 min of treatment at 30 °C, a standard liquid β-galactosidase assay was performed. Data are expressed in fold Miller units compared with treatment with 50 nM rapamycin for 90 min. DB26-3A (*MATα ura3-52 ade2 PUT1-lacZ*) was a gift from M. Brandriss. Vehicle: samples treated with N,N-dimethylformamide (DMF), the vehicle into which library compounds were dissolved. **b**, Uretupamine A was resynthesized and tested in the β-galactosidase assay by using the *PUT1-lacZ* reporter at the concentrations indicated for 60 min at 30 °C in YPD medium. **c**, Compounds derived from the uretupamine A structure were synthesized to explore structure–activity relationships. Listed are β-galactosidase assay results of treatment with each compound at 100 μM (asterisk designates 50 μM) for 60 min at 30 °C in YPD medium. Data are in percentage Miller units compared with treating with 50 nM rapamycin for 60 min. Ac, acetate; MDPO, 2-mercapto-4,5-diphenyloxazole; MBO, 2-mercaptobenzoxazole; Ph, phenyl. **d**, Binding of uretupamine B to Ure2p was determined by using surface plasmon resonance (BIAcore 3000) to have a dissociation constant of 7.5 μM. Data points were acquired in triplicate. Ure2p was immobilized to CM5 sensor chips by injection of 100 μg ml^{-1} Ure2p in 10 mM sodium acetate pH 4.5 in accordance with the manufacturer's procedures. The reference cell was derivatized with antibodies against glutathione *S*-transferase (GST) followed by GST capture. Small-molecule binding measurements and dissociation were in PBS/Tween-20 containing 10% DMF flow rate 5 μl min^{-1}).

(Fig. 2c). As expected, uretupamine B was more potent than uretupamine A (Fig. 2c). Surface plasmon resonance was used to obtain a binding constant for uretupamine A and B binding to purified Ure2p. This demonstrated that uretupamine A and B bound to Ure2p with equilibrium dissociation constants of 18.1 and 7.5 μM, respectively (Fig. 2d), which is consistent with their potencies in cells.

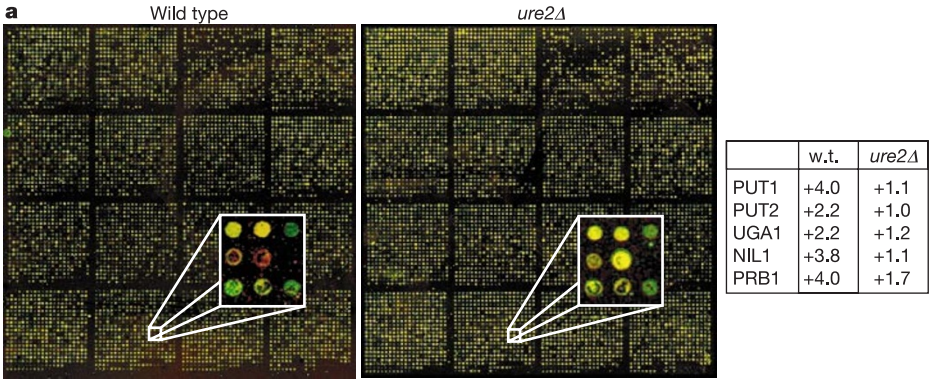
To determine the precise effects and specificity of uretupamine, we used whole-genome transcription profiling in wild-type cells as well as an otherwise isogenic *ure2*Δ strain–a 'targetless' strain[18]. (Complete transcription profiling data are publicly available in Supplementary Information and at http://www.schreiber.chem.harvard.edu.) Both uretupamine A and B upregulated only a subset of genes (including *PUT1*, *PUT2*, *PRB1*, *NIL1* and *UGA1*) known to be under the control of Ure2p (Fig. 3a, b). The expression of other genes (including *GAP1*, *MEP2*, *AGP1*, *BAT2* and *DAL5*) controlled by Ure2p was essentially unchanged (Fig. 3a, b). The compounds had little or no effect on either set of genes in a targetless *ure2*Δ strain, an otherwise identical strain lacking only the gene encoding the putative protein to which uretupamines A and B bind (Fig. 3a, b). This result suggests that a small molecule readily obtained from diversity-oriented synthesis and screening with the use of small-molecule microarrays has nearly complete cellular specificity for its screening partner, at least as judged by its global effects on the mRNA levels of treated cells.

Although Ure2p-controlled genes are normally thought of as responsive to nitrogen quality, the subset of genes induced by uretupamine (*PUT1*, *PUT2*, *PRB1*, *NIL1* and *UGA1*) has been shown to be upregulated when glucose is removed from the media[19]. The mechanism for this differential regulation of Ure2p-controlled genes in response to different nutrient signals is not understood. Ure2p represses transcription factors Gln3p and Nil1p, which might be differentially regulated to achieve this effect[19,20]. To test this hypothesis, we therefore profiled uretupamine B in *gln3*Δ and *nil1*Δ strains. Remarkably, we found that deleting *GLN3* had little effect on the actions of uretupamine, whereas deleting *NIL1* abrogated its actions (Fig. 3b). Further confirmation for this selectivity comes from whole-genome vector-based comparisons[19,21] of four profiles; these comparisons show that *URE2* and *NIL1*, but not *GLN3*, are critical for uretupamine action (Fig. 3c). Northern blot analysis confirmed the effect of uretupamine B on *PUT1* expression (normalized to *ACT1* expression) in wild-type, *gln3*Δ and *nil1*Δ strains (data not shown).

The fact that the binding of uretupamine to Ure2p induces the expression of glucose-sensitive genes in a *NIL1*-dependent manner suggests that Ure2p might itself be the target of a glucose-sensitive pathway. This is in contrast to a glucose-sensitive pathway impinging on Nil1p, bypassing Ure2p.

Because Ure2p is a phosphoprotein[9,10,22], we examined the phosphorylation state of Ure2p after different types of nutrient shift. We shifted wild-type cells from the high-quality nitrogen source, ammonium sulphate, to the low-quality nitrogen source proline. We also shifted cells from the high-quality (fermentable) energy source glucose to the low-quality (non-fermentable) energy sources acetate or glycerol. Surprisingly, Ure2p was not dephosphorylated when ammonium sulphate was removed but was dephosphorylated when glucose was removed (Fig. 4a). These data indicate that signals not previously thought to regulate Ure2p alter its phosphorylation state, whereas signals previously thought to regulate Ure2p do not alter its phosphorylation state. It remains to be seen whether there is another means by which these signals can regulate Ure2p function. Identical results were obtained for cells transferred from a glucose-containing medium to an ethanol-containing medium and for cells of a different background (W303) (data not shown). Other stresses (such as 1 M NaCl, 1 M sorbitol, pH 9.5 or heat shock) known to upregulate Ure2p-dependent genes[23,24] did not cause Ure2p dephosphorylation (Fig. 4b, data not shown). These data suggest that Ure2p is part of a signalling pathway that specifically responds to glucose.

Kornberg and Krebs first proposed that on energy sources such as



	w.t.	*ure2Δ*
PUT1	+4.0	+1.1
PUT2	+2.2	+1.0
UGA1	+2.2	+1.2
NIL1	+3.8	+1.1
PRB1	+4.0	+1.7

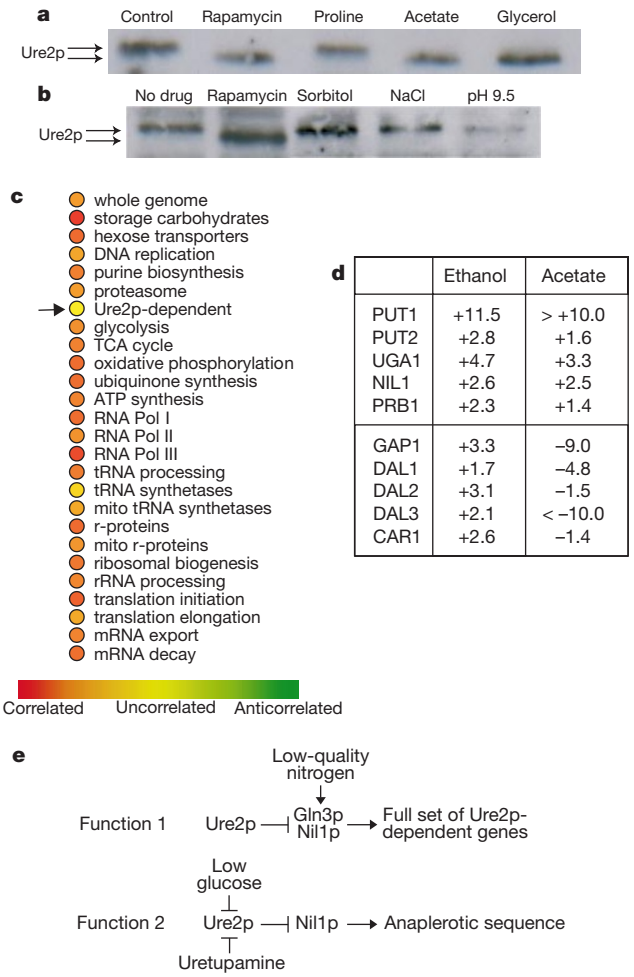
b

Gene sets	w.t.	*gln3Δ*	*nil1Δ*	*ure2Δ*
GAP1, MEP2, DAL5, BAT2, AGP1	+1.1	+1.1	–1.0	–1.0
PUT1, PUT2, UGA1, NIL1, PRB1	+2.3	+2.5	+1.6	+1.2

c

Gene sets	w.t.	*gln3Δ*	*nil1Δ*	*ure2Δ*
Whole genome	100%	89%	56%	52%
URE2-dependent genes	100%	115%	51%	59%

Figure 3 Transcription profiling of treatment with uretupamine. **a**, The left microarray corresponds to wild-type cells (PM38) treated with vehicle (DMF) versus wild-type (w.t.) cells treated for 30 min with uretupamine A at 100 μM. The right microarray corresponds to *ure2Δ* cells (PH2) treated with vehicle versus *ure2Δ* cells treated for 30 min with uretupamine A at 100 μM. Profiles were obtained as described[9]. At the right are shown specific gene inductions of some *URE2*-dependent genes from the microarrays. PM38 (*MATα leu2-3,112 ura3-52*), PM71 (*MATα leu2-3,112 ura3-52 gln3Δ5::LEU2*), MS221 (*MATα ura3-52 nil1::hisG*), PH2 (*MATα leu2-3,112 ura3-52 ure2Δ12::URA3*) were gifts from B. Magasanik and M. Brandriss. **b**, The transcription profiles of wild-type cells (PM38), *ure2*Δ cells (PH2), *gln3*Δ cells (PM71) and *nil1*Δ cells (MS221) grown in YPD medium treated for 30 min with 100 μM uretupamine B were obtained. The geometric means of gene inductions of the sets listed are shown. **c**, An analysis was performed based on treating individual profiles as high-dimensional vectors and then examining the ratios of their magnitudes as a measure of relative activity[19,21].



	Ethanol	Acetate
PUT1	+11.5	> +10.0
PUT2	+2.8	+1.6
UGA1	+4.7	+3.3
NIL1	+2.6	+2.5
PRB1	+2.3	+1.4
GAP1	+3.3	−9.0
DAL1	+1.7	−4.8
DAL2	+3.1	−1.5
DAL3	+2.1	< −10.0
CAR1	+2.6	−1.4

Figure 4 Glucose-sensitive signalling and a model of Ure2p function. **a**, A wild-type strain (PM38) was grown to mid-exponential phase in synthetic glucose (dextrose) (SD)–ammonium sulphate (AS) medium, washed with PBS and split into either SD–AS medium, SD–AS medium containing 50 nM rapamycin, SD-proline medium, synthetic acetate–AS medium or synthetic glycerol–AS medium, then incubated with shaking for 1 h at 30 °C. Control: the sample split into SD–AS medium. Synthetic medium consisted of 1.7 g of YNB medium, without amino acids and without AS, 2% carbon source, 0.1% nitrogen source, and auxotrophic supplements when needed (leucine 120 mg l^{-1}, uracil 20 mg l^{-1}). Whole-cell lysates were blotted with anti-Ure2p antibodies kindly provided by R. Wickner as described previously[19]. A previous report claimed that Ure2p-dephosphorylation does occur upon nitrogen limitation but those authors examined cells shifted from a rich medium to a synthetic-nitrogen-limited medium, thus changing many variables of the medium simultaneously[10]. In our experiments, cells were shifted from a synthetic medium containing 0.1% ammonium sulphate as a high-quality nitrogen source to an otherwise identical medium containing 0.1% proline as a low-quality nitrogen source. Inspection of published Ure2p immunoblots under similar conditions also supports the lack of a mobility shift[27,28]. **b**, A wild-type strain (PM38) was treated for 1 h with no drug, 50 nM rapamycin, salt stress (1 M NaCl), high osmolarity (1 M sorbitol) or high pH (pH 9.5). **c**, A wild-type strain (PM38) was shifted from an SD–AS medium to a synthetic acetate–-AS medium and incubated with shaking for 30 min at 30 °C. This transcription profile was compared to the profile of the same strain shifted from SD–AS to synthetic ethanol–AS medium[19]. The expression of various subsets of genes was compared between profiles by using vector algebra and the results are presented in a colorimetric comparison array[19,21]. **d**, Inspection of Ure2p-dependent genes reveals that uretupamine-sensitive genes behave similarly when cells are shifted to ethanol or acetate, whereas uretupamine-insensitive genes behave differently. **e**, Nitrogen quality regulates the Ure2p–Gln3p/Nil1p complex by signalling to Gln3p/Nil1p, whereas low glucose concentration regulates the complex by signalling to Ure2p. Low glucose concentration leads to dephosphorylation of Ure2p to induce a specific set of genes involved in an anaplerotic sequence, a state mimicked by uretupamine.

acetate, metabolic sequences called anaplerotic are activated to replenish tricarboxylic-acid-cycle intermediates[25]. Yeast cells growing in acetate-containing media have been shown to accumulate ammonia[26], which leads to the following paradox. Ammonia would repress the expression of Ure2p-dependent genes, including those thought to promote survival on acetate as part of an anaplerotic sequence (*PUT1*, *PUT2* and *UGA1*)[20]. It is possible that acetate-induced Ure2p-dephosphorylation protects the anaplerotic sequence from this repression by ammonia. We performed the transcription profile of yeast shifted from glucose to acetate and compared it with that of cells shifted from glucose to ethanol. Genome-wide analysis showed that Ure2p-dependent genes were differentially affected by the two transitions (Fig. 4c). Unlike ethanol, acetate caused the downregulation of some Ure2p-dependent genes but, like ethanol, acetate induced those genes activated by uretupamine (Fig. 4d). Taken together, these data suggest that Ure2p-dephosphorylation stabilizes the induction of genes for an anaplerotic sequence when other cellular forces might repress the expression of these same genes (Fig. 4e).

Our approach to uncovering the role of Ure2p in glucose signalling is rooted in the principles of reverse genetics. We desired a method of modulating Ure2p function selectively to examine the resulting phenotype. Because uretupamine modulates only a subset of Ure2p function, its effects are more specific than those resulting from deletion of the *URE2* gene. This property of uretupamine highlights the multifunctionality of individual proteins and addresses the challenge in proteomics to identify and control all possible inputs and outputs of each protein. With uretupamine, we demonstrated a functional connection between Ure2p, Nil1p and glucose levels. We additionally have a means to control this system more selectively than any physiological stimulus or genetic deletion. Diversity-oriented synthesis and small-molecule microarrays provide a potentially systematic method for acquiring powerful probes, where different small molecules can modulate different aspects of a protein's function, preceding the discovery of a genetic allele of a similar phenotype. □

Received 13 December 2001; accepted 5 February 2002.

1. Blinder, D., Coschigano, P. W. & Magasanik, B. Interaction of the GATA factor Gln3p with the nitrogen regulator Ure2p in *Saccharomyces cerevisiae*. *J. Bacteriol.* **178,** 4734–4736 (1996).
2. Beck, T. & Hall, M. N. The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. *Nature* **402,** 689–692 (1999).
3. Cunningham, T. S., Andhare, R. & Cooper, T. G. Nitrogen catabolite repression of DAL80 expression depends on the relative levels of Gat1p and Ure2p production in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **275,** 14408–14414 (2000).
4. Wiemann, S. *et al.* Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* **11,** 422–435 (2001).
5. Narahashi, T., Moore, J. W. & Scott, W. R. Tetrodotoxin blockage of sodium conductance increase in lobster giant axons. *J. Gen. Physiol.* **47,** 965–974 (1964).
6. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor γ (PPARγ). *J. Biol. Chem.* **270,** 12953–12956 (1995).
7. Coschigano, P. W. & Magasanik, B. The URE2 gene product of *Saccharomyces cerevisiae* plays an important role in the cellular response to the nitrogen source and has homology to glutathione S-transferases. *Mol. Cell. Biol.* **11,** 822–832 (1991).
8. Wickner, R. B. [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. *Science* **264,** 566–569 (1994).
9. Hardwick, J. S., Kuruvilla, F. G., Tong, J. K., Shamji, A. F. & Schreiber, S. L. Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proc. Natl Acad. Sci. USA* **96,** 14866–14870 (1999).
10. Cardenas, M. E., Cutler, N. S., Lorenz, M. C., Di Como, C. J. & Heitman, J. The TOR signaling cascade regulates gene expression in response to nutrients. *Genes Dev.* **13,** 3271–3279 (1999).
11. Sternson, S. M., Louca, J. B., Wong, J. C. & Schreiber, S. L. Split-pool synthesis of 1,3-dioxanes leading to arrayed stock solutions of single compounds sufficient for multiple phenotypic and protein-binding assays. *J. Am. Chem. Soc.* **123,** 1740–1747 (2001).
12. Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287,** 1964–1969 (2000).
13. Blackwell, H. E. *et al.* A one-bead, one-stock solution approach to chemical genetics: part 1. *Chem. Biol.* **8,** 1167–1182 (2001).
14. Clemons, P. A. *et al.* A one-bead, one-stock solution approach to chemical genetics: part 2. *Chem. Biol.* **8,** 1183–1195 (2001).
15. MacBeath, G., Koehler, A. N. & Schreiber, S. L. Printing small molecules as microarrays and detecting protein–ligand interactions en masse. *J. Am. Chem. Soc.* **121,** 7967–7968 (1999).
16. Hergenrother, P. J., Depew, K. M. & Schreiber, S. L. Small-molecule microarrays: covalent attachment and screening of alcohol-containing small molecules on glass slides. *J. Am. Chem. Soc.* **122,** 7849–7850 (2000).

17. Xu, S., Falvey, D. A. & Brandriss, M. C. Roles of URE2 and GLN3 in the proline utilization pathway in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **15,** 2321–2330 (1995).
18. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4,** 1293–1301 (1998).
19. Shamji, A. F., Kuruvilla, F. G. & Schreiber, S. L. Partitioning the transcriptional program induced by rapamycin among the effectors of the Tor proteins. *Curr. Biol.* **10,** 1574–1581 (2000).
20. Kuruvilla, F. G., Shamji, A. F. & Schreiber, S. L. Carbon- and nitrogen-quality signaling to translation are mediated by distinct GATA-type transcription factors. *Proc. Natl Acad. Sci. USA* **98,** 7283–7288 (2001).
21. Kuruvilla, F. G., Park P. J. & Schreiber, S. L. Vector algebra in the analysis of genome-wide expression data. *Genome Biol.* **3(3),** 0011.1–0011.11 (2002).
22. Bertram, P. G. *et al.* Tripartite regulation of Gln3p by TOR, Ure2p and phosphatases. *J. Biol. Chem.* **275,** 35727–35733 (2000).
23. Causton, H. C. *et al.* Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12,** 323–337 (2001).
24. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257 (2000).
25. Kornberg, H. L. & Krebs, H. A. Synthesis of cell constituents from C_2 units by a modified tricarboxylic acid cycle. *Nature* **179,** 988–991 (1957).
26. Bogonez, E., Machado, A. & Satrustegui, J. Ammonia accumulation in acetate-growing yeast. *Biochim. Biophys. Acta* **733,** 234–241 (1983).
27. Edskes, H. K., Hanover, J. A. & Wickner, R. B. Mks1p is a regulator of nitrogen catabolism upstream of Ure2p in *Saccharomyces cerevisiae*. *Genetics* **153,** 585–594 (1999).
28. Edskes, H. K. & Wickner, R. B. A protein required for prion generation: [URE3] induction requires the ras-regulated mks1 protein. *Proc. Natl Acad. Sci. USA* **97,** 6625–6629 (2000).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We thank R. Melki for providing bacterially expressed Ure2p protein. F.G.K. was supported by the NIH Medical Scientist Training Program, A.F.S. by the Howard Hughes Medical Institute predoctoral fellowship, S.M.S. by the Roche and NSF predoctoral fellowships, and P.J.H. by the American Cancer Society. S.L.S. is an Investigator at the Howard Hughes Medical Institute. This research was funded by a grant from the NIGMS (GM-38627).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.L.S. (e-mail: sls@slsiris.harvard.edu).

A 'periodic table' for protein structures

William R. Taylor

Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Current structural genomics programs aim systematically to determine the structures of all proteins coded in both human and other genomes, providing a complete picture of the number and variety of protein structures that exist. In the past, estimates have been made on the basis of the incomplete sample of structures currently known. These estimates have varied greatly (between 1,000 and 10,000; see for example refs 1 and 2), partly because of limited sample size but also owing to the difficulties of distinguishing one structure from another. This distinction is usually topological, based on the fold of the protein; however, in strict topological terms (neglecting to consider intra-chain cross-links), protein chains are open strings and hence are all identical. To avoid this trivial result, topologies are determined by considering secondary links in the form of intra-chain hydrogen bonds (secondary structure) and tertiary links formed by the packing of secondary structures. However, small additions to or loss of structure can make large changes to these perceived topologies and such subjective solutions are neither robust nor amenable to automation. Here I formalize both secondary and tertiary links to allow the rigorous and automatic definition of protein topology.

The organization and classification of the bewildering variety of protein structure has been approached using clustering methods. Various computer programs have been devised to measure the three-dimensional similarity of one protein coordinate set to another[3]. From these measures, similar proteins can be grouped together, given a name, and arranged in a hierarchical clustering with others that share some partial or overall similarity. Depending on the method of comparison, or the extent of expert judgements needed, differing classifications of protein structure have emerged, ranging from one that is almost completely expert-based[4], through partially automated methods[5,6] to an almost fully automatic method[7]. The drawback of these hierarchical approaches is that, although the close relationships between similar proteins are reasonably well defined, the more tentative relationships that give the large-scale structure to the hierarchy are usually beyond the ability of the computer programs to recognize and are subject to variation when defined by experts. As a result, the classifications tend to have numerous small clusters (families of super families) all roughly grouped into just a few categories that are based on overall secondary structure content and arrangement.

An important secondary problem in the classification of proteins is how large proteins should be divided into pieces (domains) that can be classified more easily. The current approach among the more automated methods is first to divide and then to classify. However, it is clear from the expert-based approach that the initial process of classification can affect how the protein is then broken into domains and so differences in domain definition are a major source of inconsistency among the current classification systems[8].

To avoid the problems associated with a hierarchy, the method outlined here is based on a set of idealized structures that are compared with all known structures. (The programs and data described in this work can be found at: http://mathbio.nimr.mrc.ac.uk/ftp/wtaylor.) The domain definition problem is less directly solved, although as the ideal structures are all of domain size, the best match can be used to define (or bias) the definition of the domains. This approach is unusual in that it shifts the classification from a clustering problem to that of finding the best set of ideal structures that can account for as much protein structure as possible. As the ideal structures will be generated from rules applied to basic 'Forms', this can be viewed as finding a minimum basis set of generating Forms. These Forms were derived from a model in

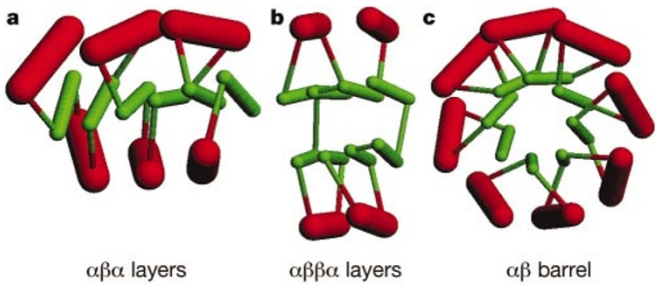


Figure 1 Stick-figure representations of the basic Forms. Each of the basic generating Forms is represented by 'stick' models in which α-helices are red and drawn thicker than the green β-strands. **a**, αβα layers. Six strands are shown, but the sheet can extend indefinitely. **b**, αββα layers. As in **a**, the sheets can be extended. (Removal of the α-layers leaves the common β-'sandwich'). **c**, Eight-fold αβ barrel. Similar barrels with 5–9 strands were constructed. (See Supplementary Information A.1 for construction details). By deleting helices and strands from these models, almost all known globular protein domains of β and βα types can be generated. Figures 1 and 4 were prepared using the program RasMol (http://www.umass.edu/microbio/rasmol).