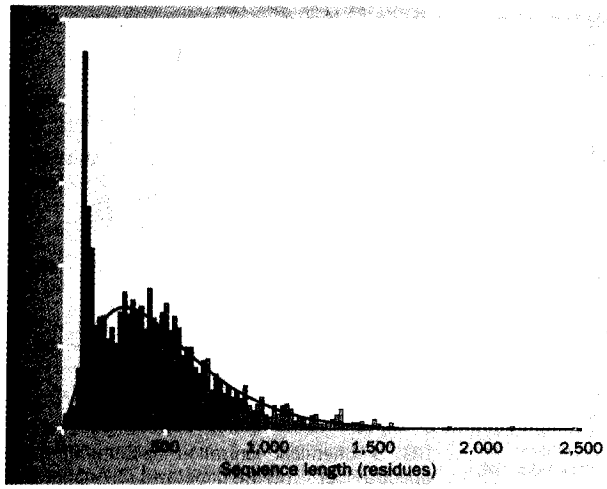
Biology's new Rosetta stone

SIR — Completion of the sequencing of the entire yeast (*Saccharomyces cerevisiae*) genome is a major event in the history of biology: all those involved are to be congratulated. Roughly 6,200 proteins from an estimated 4,000 homologous protein families are encoded on the 16 yeast chromosomes, undoubtedly representing most eukaryotic protein families. The analysis of these data will provide powerful tools for understanding the genomes of other eukaryotes, as well as the more recently completed bacterial sequence, that of *Methanococcus jannaschii*[1]. There is concern, however, that, in the competitive rush to sequence and analyse genomes, the required investment of time and money to organize and integrate this wealth of data will not be made.

Four groups have already begun initial efforts to organize these data: the *Saccharomyces* genome database project at the Department of Genetics, Stanford University Medical School; the Martinsried Institute for Protein Sequences (MIPS); the GeneQuiz[2] consortium (involving the European Molecular Biology Laboratory, the CNB–UAM in Spain and the AIC–SRI in the United States); the European Bioinformatics Institute GeneCrunch[3] resource and the Yeast protein database[4].



Sequence length distributions of *S. cerevisiae* proteins listed in GeneCrunch. ORFs that have 'clear' identifiable homologues are shown in green and those that do not match previously assigned proteins are shown in red. The probable artefact in the latter distribution contains more than 400 sequences. The former distribution seems to fit a χ^2 distribution with five degrees of freedom rather than, for example, a Poisson distribution, although we have no obvious explanation for this.

These are independent efforts encompassing the following: a physical map of genes; their assignment of probable (biochemical) function; some pattern-match information from the program PROSITE[5]; and pointers to protein structures[6]. They do not, however, have a common nomenclature or a controlled vocabulary, and contain no direct references to various known relationships between yeast genes, such as those that are on the same metabolic or regulatory pathway.

The distribution of yeast gene lengths in these databases suggests a high proportion of complex multidomain proteins and a systematic error in the assignment of open reading frames (ORFs). The figure plots the length distributions of the ORFs for which the GeneCrunch[3] resource had 'clear' matches to earlier determined sequences and for those with no significant match. The peak at 100 to 110 amino acids, three times higher than at other lengths, is seen only in the unmatched distribution, and clearly indicates an artefact in the definition of potential ORFs.

It is disturbing that this problem was not noted in any of the initial data analyses. Athough the MIPS and GeneCrunch resources are now aware of this anomaly, it still exists in the newly released *Bacillus subtilis* sequences (available at http://bacillus.genome.ad.jp) and among the many GenBank-listed mouse ORFs.

An examination of the yeast genome as presented in these resources suggests that nearly 60% of ORFs have been assigned a probable function, primarily by BLAST, FASTA or PROSITE software matches. We have confirmed most of these initial assignments, plus an additional few, using a more rigorous local dynamic programming[7] comparison with SwissProt, a new functional pattern database[8], and with *S. cerevisiae* itself. In our survey, if one requires confirmation of functional assignment by agreement between two different methods, only about 45% of the yeast ORFs have a cleanly assigned regional function.

As might be expected, the most common enzymatic function (by nearly a factor of two) is the transfer of phosphate groups within the metabolic, signalling and regulation pathways. Next are the hydrolases (acting on ester and peptide bonds), followed by the oxidoreductases, the ligases, lyases and the isomerases.

Caution must be applied to any interpretation of this function hierarchy. The proteins with assigned functions are those that have been most studied, and they may therefore be disproportionately represented in the sequence databases, making it easier to identify their yeast homologues. In addition, many of the functional assignments are the result of second-hand, inherited annotation. There are, for example, yeast genes that have been assigned tentative functional family membership, but with no information as to how the annota-

COMPARISON OF COMPLETED GENOMES

S. cerevisiae	*Z*-score	*M. jannaschii*	
YPR118w (hypothetical)	49.138	MJ0454	(translation-initiation factor, eIF-2B, subunit α)
YPR118w (hypothetical)	27.740	MJ0122	(translation-initiation factor, eIF-2B)
YGL169w (SUA5 protein, translation-initiation protein)	23.262	MJ0062	(unknown function)
YNL040w (hypothetical)	25.772	MJ0564	(alanyl tRNA synthetase)
YKL215c (hypothetical)	91.577	MJ0963	(*N*-methylhydantoinase)
YPL175w (*N*-acetylglucosaminyltransferase)	47.874	MJ1178	(unknown function)

B. subtilis	Blast score	Matches	
YABH_PBS (unknown function)	20	MJ1104	(homoserine kinase)
YABR_PBS (unknown function)	21	MJ0117	(translation-initiation factor)
	9	YJR007w	(translation-intiation factor)
YAAD_PBS (unknown function)	111	YMR096w	(hypothetical)
Similar product in *Haemophilus influenzae*, yeast and Archaebacterium *Methanococcus vannielii*	111	YNL333w	(hypothetical)
	111	YFL059w	(hypothetical)
	126	MJ0677	(ethylene-inducible protein homologue)
	33	gi1574495	(hypothetical, in *H. influenzae*)
	34	gi297176	(expressed ORF, in *Stellaria longipes*)
	33	gi44737	(unidentified reading frame, in *M. vannielii*)

Top, Homologies between hypothetical proteins, or those of unknown function, with proteins of assigned biochemical function (given in parentheses) in *S. cerevisiae* and *M. jannaschii*. Six pairs are listed for which the similarity is better than 20 s.d. from the mean. The similarity statistics were calculated using the Smith–Waterman dynamic programming comparison[8] of all 6,200 *S. cerevisiae* ORFs with all 1,738 *M. jannaschii* ORFs. Bottom, Hypothetical *B. subtilis* proteins that are homologous to proteins in *M. jannaschii* and/or *S. cerevisiae*. These were identified using BLASTp and confirmed by multi-alignment pattern analysis[9]. The blast scores are presented in equivalent numbers of aligned amino-acid identities. The YAAD *B. subtilis* protein, which is homologous to three yeast genes, one *M. jannaschii* and three others, all of unknown function, is highly conserved across the three kingdoms. It contains, for example, 37 shared amino-acid identities in a run of 40, along with many other conserved positions.

tion was originally assigned to the matched sequence. Finally, and more important, such assignments of homology are generally only at the biochemical level and provide very limited information about 'biological' function — similar biochemistry does not necessarily imply similar biology.

Caution must also be applied when considering the homologous functional assignments given to multidomain proteins. Among the yeast proteins longer than 600 amino acids, there are at least 170 with newly assigned functions where the annotation was inherited from a match to a sequence or pattern matching only a small fraction of yeast protein's length. Even when yeast genes are compared to other yeast genes, the potential exists for such confusion in defining homologous families.

A comparison of all yeast genes to themselves identified more than 500 ORFs that cannot simply be resolved into separate homologous clusters, together with about 1,800 that resolved into nearly 700 clusters defined by a common, similar amino-acid sequence. On average, the members of these clusters are randomly distributed over the 16 chromosomes, that is, the number of distinct chromosomes containing members from a particular cluster is proportional to cluster size. In addition, members of different clusters have similar neighbours on different chromosomes. These facts suggest that many transposon or other recombination events in the history of *S. cerevisiae* are possibly related to its domestication.

At least 1,867 of the yeast ORFs are preceded by two or more previously recognized yeast regulatory elements[9,10], excluding the TATA box. As described in the Transfac database, MIG1 is a yeast regulatory DNA-binding protein involved in glucose utilization. YBL086c is the only hypothetical protein matched in a set of 10 other genes by the gal-10 MIG1 binding site (R04145 in Transfac). The 14 genes originally used to define the MIG1 binding site additionally contain the site for the repressor of CAR1 expression. This site is also found in the upstream region of the hypothetical protein YBL086c, supporting its involvement in sugar metabolism.

Only six of the 6,200 promoter regions examined contain the long MATα1 sites. One of these, the upstream region of the hypothetical YBL028c gene, is additionally matched by an RAP1 binding site involved in repressing the silent mating loci. Surely this gene has a biological function associated with mating-type regulation in *S. cerevisiae*? Such analyses support the need to integrate fully our various yeast resources.

Sequence-comparison methods between the various recently sequenced genomes is a powerful tool for cross-species identification of function (table, top section), raising provocative new questions. In one case (table, bottom section), there is a very highly conserved gene product, of unknown function, shared among the three representative genomes of the three terrestrial kingdoms, *M. jannaschii*, *B. subtilis* and *S. cerevisiae*. Given its conservation, its fuction must be rather fundamental. But because of the complexities of multiple-domain proteins, and the missing regulatory and metabolic database pointers, much work remains to be done to exploit fully such comparative analyses. It is to be hoped that a major collaborative effort will be undertaken to provide a fully integrated yeast genome database that will provide biology with a new 'Rosetta stone'.

Sudeshna Das, Lihua Yu
Chrysanthe Galtatzes, Robert Rogers
James Freeman, Jadwiga Blenkowska
R. Mark Adams, Temple F. Smith
BioMolecular Engineering Research Center,
College of Engineering,
Boston University,
Boston, Massachusetts 02215, USA
James Lindellen
Time Logic Inc.,
Incline Village, Nevada 89451, USA

1. Bult, C. J. *et al. Science* **273,** 1058–1073 (1996).
2. Scharf, M. *et al.* in *Proc. 2nd Int. Conf. Intelligent Systems Mol. Biol.* (eds Altman, R. *et al.*) 348–353 (AAAI, Menlo Park, 1994).
3. Schneider, R. *et al.* in *Proc. Supercomput. '96 Semin.* (Saur, Mannheim, in the press).
4. Garrels, J. I. *Nucleic Acids Res.* **24,** 46–49 (1995).
5. Bairoch, A. *Nucleic Acids Res.* **20,** 2013–2018 (1992).
6. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. in *Crystallographic Databases — Information Content, Software Systems, Scientific Applications* (eds Allen, F. H., Bergerhoff, G. & Sievers, R.) 107–132 (Int. Union Crystallog., Bonn, 1987).
7. Smith, T. F. & Waterman, M. S. *J. Mol. Biol.* **147,** 195–197 (1981).
8. Adams, R. M., Das, S. & Smith, T. F. *Protein Sci.* **5,** 1240–1249 (1996).
9. Ghosh, D. *Nucleic Acids Res.* **20,** (suppl.) 2091–2093 (1992).
10. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. *Transcript. Factors* **24,** 238–241 (1996).

Induced synthesis of plant volatiles

SIR — Wounding of plant tissue by insect feeding triggers the release of volatile compounds, which have been implicated in the attraction of natural enemies of the insect herbivores[1–8], and are derived from several biosynthetic pathways. The release of indole and several terpenes is delayed, and is induced only by interaction of damaged plant tissues with a substance (or substances) in the oral secretions of insects[9,10]. We have used $^{13}CO_2$ to show that cotton plants (*Gossypium hirsutum* L.) damaged by feeding beet armyworms (*Spodoptera exigua* Hübner) synthesize these induced volatile compounds *de novo*, as distinct from the conversion of some previously prepared precursor. Furthermore, ^{13}C-labelled volatiles are released from the leaves soon after synthesis and are rapidly replaced by unlabelled compounds when $^{13}CO_2$ is replaced by $^{12}CO_2$.

This is the first demonstration of *de novo* synthesis of volatiles induced by insect damage to a plant, showing the dynamic role of plants in interactions with the second (insect herbivores) and third (parasitoids and predators) trophic levels.

For several of the volatiles that are released, the delay in emission after herbivore attack suggests that chemical changes

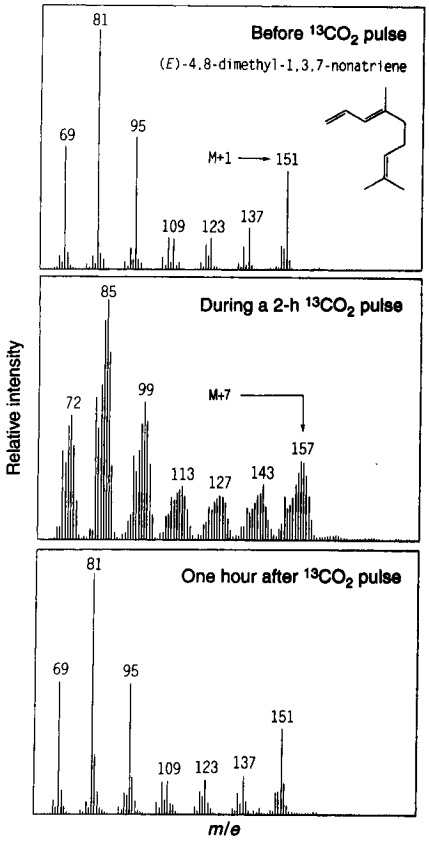


FIG. 1 Chemical ionization (isobutane) mass spectra of (*E*)-4,8-dimethyl-1,3,7-nonatriene released from beet armyworm-damaged cotton plants immediately before the introduction of $^{13}CO_2$ in synthetic air (top), during a 3-h collection in which a 2-h pulse label with $^{13}CO_2$ in synthetic air was administered (middle) and a 3-h collection starting 1 h after the pulse label ended (bottom). Laboratory-reared[19], third instar beet armyworm larvae, starved for 7 h, were placed on six-week-old, greenhouse-grown cotton plants (variety Deltapine Acala 90) at 1500 h, and allowed to feed continuously. Filtered air (1 l min^{-1}) passed down over the plants contained within a 34 × 6-cm diameter glass sleeve and out of a hole in a split plate at the base that closed loosely around the stem of the plant[20]. Synthetic premixed air (360 p.p.m. (≥ 99% ^{13}C) CO_2, 20.7% O_2 and a balance of N_2) was introduced at 0900 h on day 2 by flushing the chamber at 10 l min^{-1} for 2 min and then reducing the flow to 1 l min^{-1} for 2 h. The same procedure was used to switch back to atmospheric air. Plant volatiles were collected between 0600 and 2100 h in 3-h intervals by pulling half of the air passing over the plant through Super Q adsorbent (25 mg) traps around the base of the volatile-collection chamber; the rest of the air vented out from the bottom of the system. Volatiles were eluted from the absorbent traps with 150 μl CH_2Cl_2; 400 ng each of *n*-octane and nonyl acetate were added as internal standards and 1-μl aliquots analysed by gas chromatography–mass spectroscopy on a Finnigan MAT ITS40 (ion trap) mass spectrometer.