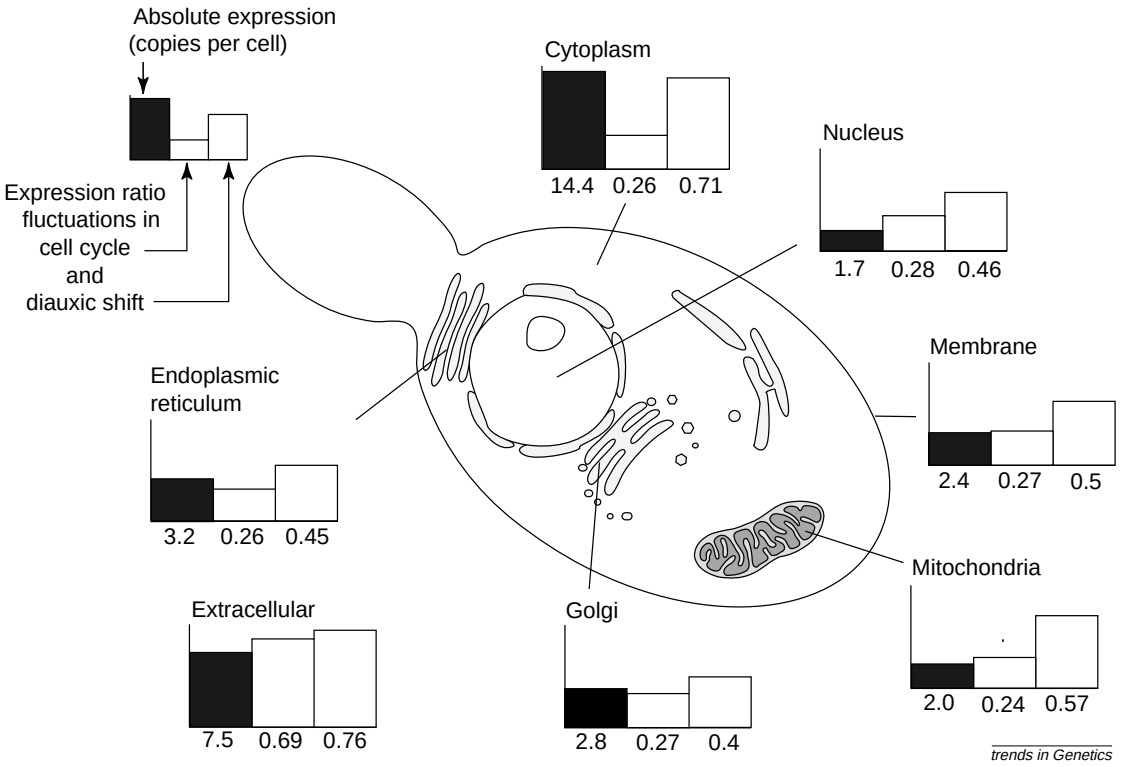
Genome-wide analysis relating expression level with protein subcellular localization

We investigate the relationship between protein subcellular localization and gene expression for a variety of whole-genome expression datasets. We find high expression levels for cytoplasmic proteins and low ones for nuclear and membrane proteins. Excreted proteins have large fluctuations in expression level over various time courses. Our results can be interpreted in terms of protein structure and function. Detailed statistics are at http://bioinfo.mbb.yale.edu/genome/expression.

FIGURE 1. Expression patterns in major compartments of the yeast cell



Expression patterns of major compartments of the yeast cell. The graph for each compartment consists of three parts (left, middle and right) described below. All scales are logarithmic. Left (solid), average absolute expression level of all transcripts in the compartment. Values refer to the dataset by Holstege *et al.*[4], which we used as a reference. We chose the Holstege data because we believe that it is the most accurate absolute gene expression level data. However, one would get similar results picking other datasets as a reference. In fact, on our website we show how similar results can be derived from a 'combined' dataset derived from averaging many different experiments. Note that further detail on this figure is shown in Figure 2, which gives the numbers of different transcripts for each compartment. Middle, fluctuation of expression levels during the time course of the yeast cell cycle with alpha-factor arrest[6]. To calculate the fluctuation, we start with the logarithm of the expression ratio for gene i at time t (Eqn 1):

$$R(i,t) = \log_2\left(\frac{r(i,t) - r_{Background}(i,t)}{g(i,t) - g_{Background}(i,t)}\right) \quad \text{Eqn 1}$$

where $r(i,t)$ and $g(i,t)$ represent the red and green fluorescent signals at a particular time point. It is usual to analyse the logarithm of the expression ratio, rather than the expression ratio itself, because the logarithm is generally distributed symmetrically around zero[6]. Note that because of the structure of microarray experiments, all gene-to-gene comparisons have to be done using expression ratios rather than absolute measurements. Then we calculate the standard deviation $\sigma(i)$ in this quantity over all time points t for each gene i (Eqn 2):

$$\sigma^2(i) = \left\langle R^2(i,t) - \left\langle R(i,t)\right\rangle^2 \right\rangle \quad \text{Eqn 2}$$

Finally, to get the numbers in the graph, we average $\sigma(i)$ over the genes in a given compartment (Eqn 3):

$$\sigma_{Compartment} = \left\langle \sigma(i) \right\rangle_{i \in Compartment} \quad \text{Eqn 3}$$

The results from the other cell cycle experiments with CDC15 (Ref. 6) and CDC28 (Ref. 3) show similar trends (data not shown). Right, fluctuation of expression levels during the timecourse of the diauxic shift experiment[2]. Values are calculated in the same way as above.

Amar Drawid *‡
Amar.Drawid@yale.edu

Ronald Jansen *‡
Ronald.Jansen@yale.edu

Mark Gerstein * †§
Mark.Gerstein@yale.edu

Department of *Molecular Biophysics & Biochemistry and †Computer Science, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA.

‡These authors contributed equally to this work.

§ Corresponding author

0168-9525/00/$ – see front matter PII: S0168-9525(00)02108-9

The recent advent of experiments that measure gene expression levels (mRNA transcript abundance levels) on a genome-wide scale allows a comprehensive view of gene activity patterns in cells. For instance, these experiments have demonstrated that the expression patterns of many functionally related genes are similar[1–8]. Here, we show that, for yeast, expression levels in these experiments are clearly correlated with the subcellular localization of the corresponding protein. Furthermore, this correlation can be interpreted in terms of broad classes of protein structures and functions.

We scaled the expression levels generated by a range of techniques (gene chip, SAGE, cDNA microarray) for yeast in a variety of conditions into a common framework and cross-referenced them with the known localizations of approximately 2000 yeast proteins found in the MIPS[9] and YPD[10] databases. (Further details are given in the caption to Figure 1 and on the associated website, http://bioinfo.mbb.yale.edu/genome/expression.

Absolute expression levels

As shown in Figure 1, high expression levels can be observed for cytoplasmic proteins, low levels for nuclear and membrane proteins, and middling levels for secretory pathway proteins, i.e. those secreted or in the endoplasmic reticulum (ER) and golgi apparatus.

Figure 2 shows a more detailed representation of the absolute expression levels in Figure 1. We chose the dataset of Holstege *et al.*[4] as a reference because it results from the careful averaging over many experiments. Figure 2 shows simplified box-plot representations of the underlying distributions of expression levels for each of the different subcellular compartments for this dataset. It is evident that each compartment shows an appreciable spread of expression levels and that the distributions for the compartments with the highest expression levels are more spread-out than for those with lower expression levels. Full representations of each expression level distribution are shown on our website (http://bioinfo.mbb.yale.edu/genome/expression). These show that the distributions are roughly exponential, although the distribution tails are somewhat longer. The exponential shape may reflect the fact that many genes are expressed at a basal level whereas a smaller number are highly active in the particular state of the cell.

Table 1 shows key statistics based on the box-plots in Figure 2 for the Holstege *et al.*[4] dataset. For comparison, we show the same statistics for a variety of other expression experiments that used different techniques (gene chips and SAGE). The figure shows that our results are largely consistent over the variety of experiments we analysed. By this we mean that the overall trend of high expression in the cytoplasm and low expression in the nucleus can be observed consistently in the different datasets. However, the datasets do differ in the exact value of each statistic, with these differences probably resulting from the slightly different protocols and growth conditions employed in each experiment.

Expression fluctuations

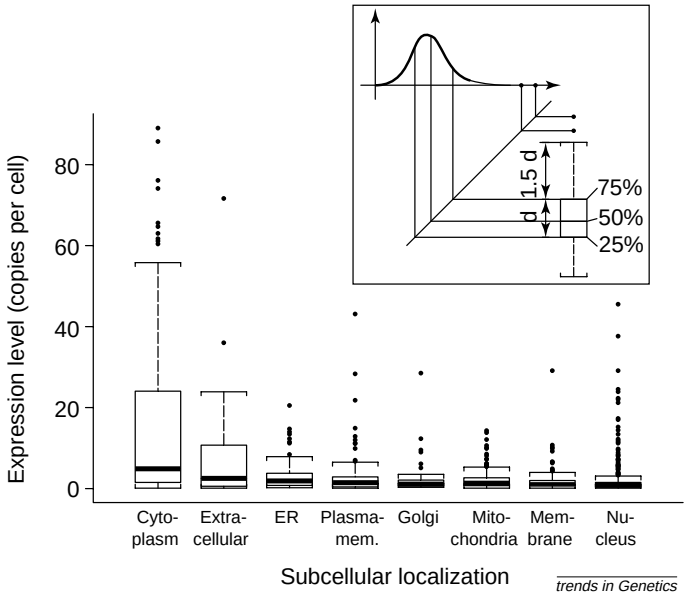
Expression measurements over time also enable us to relate fluctuations in expression to localization. As described in the caption to Figure 1, and shown in the body of the figure, we measure the magnitude of fluctuation in terms of the standard deviation in expression ratio over a time course. For the yeast cell cycle time course[3,6] we find that secreted proteins have, perhaps predictably, high fluctuations. (However, proteins in the secretory pathway with final destinations in the endoplasmic reticulum (ER) or golgi show fluctuations slightly below average.) Plasma membrane proteins, needed for transporting molecules out of the cell, have the second highest fluctuations. Biologically, these results are quite reasonable as the export of proteins from the cell is quite variable and depends on the exact state of the cell, whereas the amount of intracellular protein has to be maintained at a more constant level.

For the diauxic shift time course[2], which is also shown in Figure 1, we again observe high fluctuations for secreted proteins. We also observe particularly high fluctuations for cytoplasmic and mitochondrial proteins, as expected when the cell shifts from fermentation to respiration and alters the activity of many metabolic enzymes.

Structure–function analysis

Analysis of the functions of the associated proteins further elucidates the relationship between expression and localization. Using the MIPS classification[9], we can subdivide the yeast genome into various functional categories, e.g. 'cell structure' and 'transcription'. Figure 3a shows the average absolute expression levels for each functional category for a variety of experiments, doing for function what Figure 1 does for localization. Likewise, Figure 3b shows box-plot representations of the expression distributions for select functional categories in a similar fashion to Figure 2.

FIGURE 2. Standard box-plot representation of expression distributions in different yeast compartments



Standard box-plot representation of the distribution for the dataset by Holstege *et al.*[4] This complements with Figure 1, which shows only the mean expression levels, and gives some sense of the distribution of levels. The box-plot is explained in the inset. The thick line in the middle of the central vertical box shows the median. The central box contains the values for 50% of the total number of transcripts around the mean. Thus, 25% transcripts have values above the upper boundary of the central box, i.e. it is the 75% or top quartile line. Similarly, 25% transcripts have values below the lower boundary of the central box, i.e. it is the 25% or bottom quartile line. The brackets extend up to 1.5 times the interquartile distance (the distance between the 25% and 75% lines) above and below the central box. Points beyond the brackets are defined as outliers. All measurements are in copies per cell.

TABLE 1. Key statistics of the box-plot representations[a]

Distribution parameters[b]	Subcellular localization								
	Cytoplasm	Extracellular	ER	Plasma-membrane	Golgi	Mitochondria	Membrane	Nucleus	Data set
75%	23.5	10.5	3.6	2.7	2.3	2.6	1.8	1.3	Ref. 4
50% (median)	4.5	2.5	1.7	1.2	1.3	1.1	0.9	0.7	
25%	1.3	0.6	0.8	0.5	0.9	0.6	0.4	0.4	
No. of ORFs	479.0	16.0	113.0	106.0	48.0	266.0	148.0	698.0	
75%	26.8	5.5	3.0	1.9	2.0	2.3	1.5	1.0	Ref. 7
50% (median)	4.2	2.5	1.1	0.5	0.9	0.8	0.6	0.4	
25%	0.9	0.4	0.5	0.2	0.5	0.4	0.2	0.2	
No. of ORFs	480.0	17.0	116.0	115.0	48.0	262.0	149.0	668.0	
75%	11.9	6.7	4.7	3.3	3.3	3.4	3.3	3.3	Ref. 5
50% (median)	3.4	3.3	3.3	1.7	1.9	2.7	1.7	1.4	Mating type a
25%	2.0	1.5	1.4	1.2	1.3	1.4	1.2	1.2	
No. of ORFs	494.0	18.0	117.0	128.0	49.0	273.0	166.0	739.0	
75%	15.0	5.2	5.5	2.0	3.0	3.0	2.0	2.0	Ref. 1
50% (median)	4.0	2.0	2.0	1.5	2.0	2.0	1.0	1.0	SAGE
25%	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	log
No. of ORFs	242.0	6.0	55.0	48.0	23.0	97.0	50.0	215.0	phase

[a]Key statistics derived from Fig. 2. For comparative purposes we show these statistics for a variety of different experiments as well as for that of Holstege *et al.*[4]

[b]Number of different transcripts with expression levels that are associated with the cellular compartment in each experiment. These were obtained by cross-referencing with the localization databases; 75% expression value of the top quartile line, i.e. the 75% line. For instance, 75% of transcripts belonging to the cytoplasm in the Holstege data have a value of less than 23.5 copies per cell. Similarly, the row '50%' contains the median value of expression for transcripts in each compartment and the row '25%' contains the expression value of the bottom quartile line.

Although the experiments vary in methodology and growth conditions, the same trend can be observed. The differentiation of expression according to subcellular compartmentalization is evident for all cases. More detailed statistics can be found on the website. All the data has been scaled into a common reference system based on the Holstege data. Furthermore, on the website, we give results similar to those in this table for a 'combined' dataset derived from averaging many different experiments. Abbreviation: ORFs, open reading frames.

We observe that proteins in the 'transcription' and 'transport' categories have lower average expression levels (1.8 and 2.7 versus 3.2 for all classified genes; all values are in copies per cell for the data set of Holstege *et al.*[4]; see Figure 3a). On the other hand, expression levels in the categories 'protein synthesis' and 'energy' are above average (16.0 and 5.4). This is all in accord with the localization results (as noted above, nuclear and membrane proteins have low expression levels and cytoplasmic proteins have high expression levels). Proteins involved in transcription include DNA-binding and regulatory proteins with clear nuclear localization; proteins involved in extracellular transport are often located in the cell membrane. Furthermore, the high level of expression associated with protein synthesis is due largely to the ribosomal proteins (average expression level of 23.2), which are in the cytoplasm. In contrast, the amino-acyl-tRNA synthetases, which are also part of the broad category of protein synthesis, have considerably lower levels of expression. Proteins involved in energy production include the cytoplasmic proteins involved in glycolysis, which have high levels of expression (20.5), and the more lowly expressed mitochondrial proteins involved in the tricarboxylic acid (TCA) cycle (2.0).

The relationship between expression and localization is also linked to protein structure, although to a lesser extent. Using membrane protein prediction plus classifications of the known soluble protein folds[11,12], we can subdivide the proteins in the yeast genome into helical membrane proteins, soluble proteins and, among the latter, proteins with a structural architecture that is all α, all β and mixed αβ. Figure 3c shows average expression levels for these structural classes, in analogy to the presentation in Figures 1 and 3a for localization and function. We find a low average expression level for the transmembrane proteins (1.7). On the other hand, proteins with mixed αβ architecture, which are typically found in the cytosol[13], are the most highly expressed among the soluble proteins (3.5 versus 2.5 for other soluble proteins).

Conclusions

To conclude, we find a clear statistical relationship between a gene's expression level and its subcellular localization, with cytoplasmic proteins tending to be highly expressed and nuclear or membrane proteins more lowly expressed. This relationship may be useful in predicting subcellular localization, given expression information[15]. The correlation between expression and localization may be related to the volumes of the various subcellular compartments. The cytoplasm, for instance, has much more space for proteins than the other compartments. To achieve the same effective concentration, the expression level for freely diffusing proteins destined for larger compartments may need to be higher than for smaller ones. In other words, genes associated with cytoplasmic proteins may be regulated differently, in a different dynamic range, than those associated with membrane and mitochondrial proteins.

Acknowledgements

We thank the NIH and the Keck foundation for support.

References

1 Velculescu, V.E. *et al.* (1997) Characterization of the yeast transcriptome. *Cell* 88, 243–251

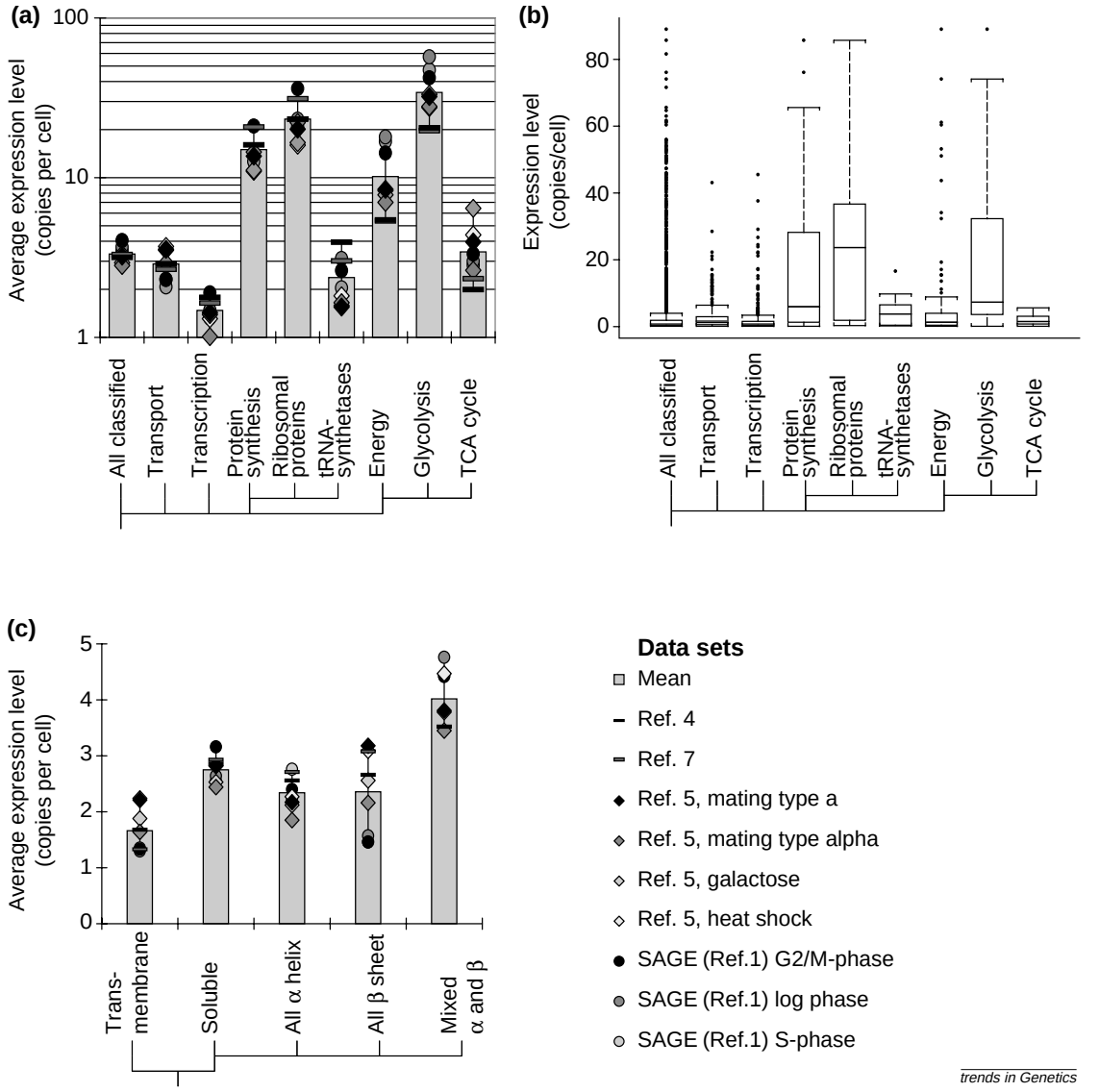
2 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686

3 Cho, R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73

4 Holstege, F.C. *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95, 717–728

5 Roth, F.P. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome

FIGURE 3. Expression related to structure and function



Statistics on gene expression related to functional and structural categories in analogy to the presentation relating expression to localization in Figures 1 and 2. (a) Average expression levels of genes in selected MIPS (Ref. 9) categories. The tree below the *x*-axis indicates a subset of the hierarchy of the different classifications. For example, it shows how all yeast proteins have as a subset 'energy production' proteins and how these, in turn, have subsets associated with glycolysis and the TCA cycle. The *y*-axis displays the average expression levels for transcripts associated with each respective category. We analysed nine datasets[1,4,5,7], each represented by different symbols. The columns represent the averages of all datasets. The total number of ORFs in the dataset by Holstege *et al.*[4] classified by MIPS is 2471 ('all classified'). The number of transcripts for the other categories is 712 (transport), 721 (transcription), 318 (protein synthesis), 190 (ribosomal proteins), 37 (tRNA synthetases), 210 (energy), 25 (glycolysis) and 19 (TCA cycle). (b) Box-plots of the expression level distributions for selected MIPS categories. The box-plots are again generated for the dataset by Holstege *et al.*[4] and show the same distribution parameters as in Figure 2a. (c) Average expression levels of proteins in broad structural classes. For structural classification we predicted, as described previously[11], transmembrane segments in the yeast genome using a hydropathy scale and used PSI-blast[14] to find matches of the yeast genome with PDB structures. For the dataset by Holstege *et al.*[4] the number of transcripts in each category are 600 (transmembrane), 4743 (soluble), 290 (all α helix), 202 (all β sheet), and 781 (mixed α and β). The number of transmembrane and soluble ORFs does not add up to the complete number of ORFs in the yeast genome because we are considering only those ORFs that can be assigned to either category with high confidence. The number of transcripts for the other structural classes reflects the amount of structure matches that could be found with the PDB. More detailed statistics can be found at our website http://bioinfo.mbb.yale.edu/genome/expression. Abbreviations: ORFs, open reading frames; PDB, Protein Data Bank.

mRNA quantitation. *Nat. Biotechnol.* 16, 939–945
6 Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297
7 Jelinsky, S.A. and Samson, L.D. (1999) Global response of Saccharomyces cerevisiae to an alkylating agent. *Proc. Natl. Acad. Sci. U. S. A.* 96, 1486–1491
8 Niehrs, C. and Pollet, N. (1999) Synexpression groups in eukaryotes. *Nature* 402, 483–487
9 Mewes, H.W. *et al.* (1999) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 27, 44–48
10 Hodges, P.E. *et al.* (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* 27, 69–73

11 Gerstein, M. (1998) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* 33, 518–534
12 Jansen, R. and Gerstein, M. (2000) Analysis of the yeast transcriptome with broad structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res.* 28, 1481–1488
13 Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* 288, 147–164
14 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
15 Drawid, A. and Gerstein, M. (2000) A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* 360, 1077–1093

Intron sliding in conserved gene families

Intron sliding (also called intron slippage or migration) can be defined as the relocation of intron–exon boundaries over short distances (1–15 bases) in the course of evolution. The reality of intron sliding has been debated for a long time. Here we present a Monte Carlo statistical analysis of broadly sampled data on intron positions; it indicates that intron sliding by one base is a real phenomenon.

Compared to coding regions, little is known of the evolutionary processes that occur within introns. The evolutionary origin of introns has been a source of debate since their discovery[1]. The 'introns early' hypothesis suggests that introns existed prior to the divergence of prokaryotes and eukaryotes, and have been subsequently completely lost in archaeal and bacterial genomes[2,3]. The 'introns late' hypothesis counters that introns have been inserted into eukaryotic genomes after the split from the other two domains of life[4,5]. Several mechanisms of intron evolution have been discussed by proponents of both hypotheses, such as intron deletion, intron insertion and intron sliding[6–9]. Intron sliding has been postulated by advocates of the 'introns early' hypothesis to explain the surprising finding that the position of apparently orthologous introns can vary among lineages[9]. However, the 'introns late' camp have maintained that intron sliding, if it occurs at all, has contributed little to the diversity of intron positions[7,10]. Indeed, to date, no unambiguous cases of intron sliding have been reported. A recent analysis of the distribution of intron insertion positions among a number of genes and lineages attributed most differences in intron position to alignment artefacts[10].

FIGURE 1. Examples of probable intron sliding

(a)

```
    CDS          Intron                 Intron         CDS
GAC AAG GAG G-/gtgagtcgag............tcccttgtag/GC ATT CCG CCG  1
GAC AAG GAG GG/gtgcgtgcct............catgctatag/-C ATT CCT CCC  2
```

(b)

```
    CDS          Intron
GGC CTG GAG T/gt--gagtaggc-ctgctggg.....gcta--gaagcaaaacct-gct  1
AGC CAG GAG -/gtgtgagttggcgctcctggg.....gccactgaag-aaaacctggct  2

                                  Intron              CDS
gtgctgggcccagtgacagccagg.....tggcctctcccacccag/-AC CCC ATG CTG  1
gtgctggccccaagga-agccagg.....tggc-tctcc-acccag/TAC CCC ATT GTG  2
```

(c)

```
    CDS          Intron
TTC TCG TCC AG/gtaagatgatggatggcttcatgattatgatgacgtagtagttt...  1
ATC TCT ACC G-/gtaagatgatgaacggcttcatgattatg-----ctagtagttt...  2

                                  Intron              CDS
...ttttttggatgcattaagttgatggg-ttttgctgttcttgcag/-G TTT GGT GCG  1
...ttcttttgatgcatttagctgatgggatttttctgttcttacag/GT CTT GGT GCT  2
```

trends in Genetics

(a) Algal and fungal polyubiquitin genes (1, *Volvox carteri*, X74214; 2, *Schizophyllum commune*, AF031628). Given the extraordinary evolutionary conservation of ubiquitin, the amino acid sequence-based alignment is unambiguous. Sliding seems to have occurred via insertion of a G in the donor site (highlighted in black in sequence 2) accompanied by deletion of a G in the acceptor site (highlighted in gray in sequence 1), or alternatively via a deletion in the donor site and an insertion in the acceptor site. (b) Mammalian cholesterol esterase genes (1, human, M94579; 2, rat, M69157). The exon–intron structure of the human gene is identical to that of the gorilla ortholog (AF206618) and is consistent with several complementary DNA (cDNA) sequences (NM_001807, AF081673, M54994); the exon–intron structure of the rat gene is consistent with several cDNA sequences (X16054, M15893, NM_012732). (c) Plant alcohol dehydrogenase II gene (1, *Oryza sativa*, M36469; 2, *Oryza officinalis*, AF148613). The exon–intron structure of the *Oryza sativa* gene is confirmed by the cDNA sequence (X16297); the exon–intron structure of the *Oryza officinalis* gene is confirmed by several homologous sequences (X02915, X12733, X12734).

Nevertheless, several very likely cases of intron sliding are apparent. A straightforward example includes polyubiquitin genes, whose alignment does not contain any insertions/deletions and seems to be unambiguous. In the alga *Volvox carteri*, an intron interrupts a GGC codon (amino acid no. 35) between the first and second positions, whereas in the fungus *Schizophyllum commune*, the apparently orthologous intron interrupts the second and third positions of the same codon (Fig. 1a). Typically, orthologous introns from distant species do not show any appreciable sequence conservation due to their evolution under the neutral (or nearly neutral) model[11]. Nevertheless, probable intron sliding in mammalian cholesterol esterase genes (Fig. 1b), which was noticed originally by Stoltzfus and co-workers[10], is supported by significant sequence similarity between the introns (49% identity, probability of occurring by chance, $P = 0.0008$ as estimated using the ALIGN and RSS programs of the FASTA package). Another example of potential intron sliding, with an even greater conservation of the intron sequences themselves, has been observed in alcohol dehydrogenase II genes from two species of rice (Fig. 1c). These observations, anecdotal as they are, suggest that intron sliding could occur in evolution, and the problem merits further analysis.

We analyzed the intron positions in 40 conserved gene families (Table 1). For each of these families, nucleotide sequence alignments were derived from protein alignments and were essentially unambiguous. Therefore reliable information on changes in the location of introns in the course of evolution could be obtained. All intron locations in the genes from each family were mapped onto an operational 'scaffold gene' (Fig. 2)[10,12]. Under this procedure, a series of introns that occupy exactly the same position in all aligned genes produce just one point on the scaffold. Altogether, we found 665 variable intron locations, with