with larger receptive fields could prevent this confounding of stimulus properties.

We cannot yet prove that either spike count or temporal modulation are actually used in visual processing. However, there are many advantages in regarding temporal modulation as the intrinsic neuronal code underlying visual perception. For example, the simultaneous encoding of several stimulus features by temporal modulation, a general mechanism found throughout the visual system, could be used to avoid confounding information. Furthermore, the overlap of the temporal codes in all areas would allow concurrent, rather than sequential, visual processing.

REFERENCES AND NOTES

1. D. C. Van Essen and J. H. R. Maunsell, *Trends Neurosci.* **6**, 370 (1983); M. Mishkin, L. G. Ungerleider, K. A. Macko, *ibid.*, p. 414; E. A. DeYoe and D. C. Van Essen, *ibid.* **11**, 219 (1988); S. Zeki and S. Shipp, *Nature* **335**, 311 (1988).
2. H. B. Barlow, *Perception* **1**, 371 (1972).
3. B. J. Richmond and L. M. Optican, *J. Neurophysiol.* **57**, 147 (1987).
4. L. M. Optican and B. J. Richmond, *ibid.*, p. 162; J. W. McClurkin, T. J. Gawne, L. M. Optican, B. J. Richmond, *ibid.*, in press; B. J. Richmond, J. W. McClurkin, T. J. Gawne, L. M. Optican, *Soc. Neurosci. Abstr.* **14**, 308 (1988).
5. B. J. Richmond and L. M. Optican, *J. Neurophysiol.* **64**, 370 (1990).
6. All animal protocols were approved by the National Eye Institute and the National Institute of Mental Health animal care and use committees and conformed with the U.S. Public Health Service guidelines for animal care and use.
7. N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing* (Springer-Verlag, Berlin, 1975).
8. K. L. Coburn, J. W. Wesson Ashford, J. M. Fuster, *Behav. Neurosci.* **104**, 62 (1990).
9. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).
10. W. R. Uttal, *The Psychobiology of Sensory Coding* (Harper & Row, New York, 1973).

4 December 1990; accepted 24 May 1991

Technical Comments

Counting and Discounting the Universe of Exons

R. L. Dorit *et al.* (*1*), beginning with the premise that all proteins are constructed of modules encoded by discrete exons, undertake to count the "underlying universe" of exon types. They calculate that a reasonable estimate is between 1000 and 7000. I question whether there is a suitable method for counting the members of such a universe, and whether the sequence-alignment methods used by Dorit *et al.* actually identified homologous pairs in most cases. Finally, I challenge whether there is an underlying universe of exons to count.

It is impossible to prove that two sequences have not evolved from a common ancestor; they may just have changed so much by amino acid replacement that the relationship is obscured. All we can do is make statistical judgments about the likelihood that similarities are not due to chance (*2*). Let us assume that the number of exon types has been constant from the time of the first encoded proteins and that all proteins are indeed constructed from that prototypic set. Can the method used by Dorit *et al.* establish the primordial number? Their strategy depends on determining two fundamental numbers: (i) the number of nonhomologous protein types in their database collection of exons, and (ii) the number of pairs of the exons themselves that are homologous but embedded in nonhomologous proteins. Their judgments about which sequences belong in which class depend exclusively on sequence compairson.

Suppose that when Dorit *et al.* compared their exon types they did not find *any* homologous pairs. Should they conclude, as their formula would demand, that the number of exotypes is infinite? The exact opposite could also be correct: there could have been a single exotype that started the entire expansion, but amino acid replacements over the past 3 to 4 billion years have eroded all possible recognition at the pairwise level.

This is not to say that one cannot use amino acid sequence comparison to show that portions of proteins, whether or not they are encoded by exons, have been shuffled about during evolution. There are many cases where parts of proteins are more similar to portions of other proteins than are the parental proteins in which they are embeded. It is also possible that, under carefully specified conditions, sampling may allow an upper bound to be placed on the number of entities being shuffled. No lower bound short of unity can be established, however.

Sequence comparison was the only basis on which Dorit *et al.* determined the two samples sets needed for the combinatorial estimation: in both instances the determinations are vulnerable to errors of judgment. It might seem that compiling the starting list of exons from nonhomologous proteins, the numerator n in the sampling statistic, is straightforward. Although species redundancies are readily removed, homologous entries that result from past gene duplications present a much greater challenge. Nonetheless, let us assume that the culled list generated by Dorit *et al.* represents 1255 exons from nonhomologous proteins.

The objective is to identify any homologous exons among them. The criterion is that a pair of exons be significantly more similar than are the proteins from which they are drawn. To this end, Dorit *et al.* compared exon sequences using a program that did not allow gaps and scored only identities. In line with their premise that exon size was a principal determinant, comparisons were restricted to exons of similar lengths. The cutoff for deciding whether a match was significant was determined by using a global measure based on the overall amino acid composition of the database.

How effective was this search? I subjected each internal measure of the 14 pairs they found to a conventional alignment and scrambling test (*3*) to see how the results fared statistically by an internal measure (Table 1). Of the 14 pairs, four had reasonably significant scores (>4 standard deviations above the random mean). All of these sequences had been reported previously by others as examples of "exon shuffling" and are likely valid. They included the epidermal growth factor (EGF) domain being moved into the blood clotting factors IX and XII (*4*), collagen-like segments being translocated into complement component Clq (*5*) and a mannose-binding protein (*6*), and a shuffle of a segment between thyroglobulin and the Ii-antigen (7). Other than the four, only one other pair reached the generally applied minimum threshold of 3 standard deviations above the random mean, this being an unlikely match between a hydrophobic signal peptide from a chloroplast gene product and a membrane-spanning sequence from the mouse red cell band 3 protein.

None of the remaining cases bears up to scrutiny. Indeed, there is good evidence to reject several of the matches a priori. In the case of the proposed relationship of a collagen exon and an elastin exon, for example, the repeat structures of the two proteins are fundamentally different, even though both are rich in glycine and proline (Fig. 1). The elastin repeat leads to a spiral of β turns (*8*), whereas collagen is a three-stranded cable.

There are other inconsistencies regarding the alleged homologous pairs. In the matchup of a mouse collagen exon 5 (MAC5) with a similar sized complement Clq exon, the complement exon contains the signal peptide region over its first half and must be

structurally dissimilar to the $(Gly\text{-}X\text{-}Pro)_x$ skein with which it is matched in MAC5. Finally, the characteristic Gly-X-Pro rhythm of the various collagen sequences shown is independent of exon length.

As for the alleged relationship between an albumin and a keratin exon, the only thing these two have in common is that both are parts of helix-rich proteins. In the case of keratin, the helices are wound around each other to form coiled-coils; in albumin the helices are components of a single-chained multidomained structure. I searched each exon against the Protein Identification Resource database (*9*). The keratin exon retrieved keratins, desmins, laminins, and neurofilament proteins, but *no* albumins. The albumin exon, in contrast, retrieved all known albumins, α fetoproteins, and vitamin D–binding proteins, but *no* keratins. Clearly the "no gap-identity only" program, its emphasis on length notwithstanding, retrieved a chance match well outside the bounds of anything reflecting common ancestry.

All the other posited homologous exons had sequences enriched with particular amino acids, including a number of pairs involving signal peptides of the sort Dorit *et al.* sought to remove during the purging process because of their biased amino acid compositions. Compositional bias was evident in the pair that had a chorion protein, which is similar to keratin in being glycine-rich. Similarly, the neurofilament exon specified is rich in glutamic acid, as are the heat shock protein and *fps* oncogene exons with which it was paired. None of these relationships was statistically significant as measured by a randomization test that took account of such biases in composition. Only five at most of the 14 pairs listed by Dorit *et al.* meet the minimum criteria for pairwise homology. As such, the estimated number of exons in the posited "universe of exons" could have been given as 157,000.

Allowing that their initial attempt to identify homologous exon pairs in nonhomologous settings may have been less than perfect, Dorit *et al.* reexamined the results of the pairwise comparisons of the 1255 exons in a slightly different fashion referred to as a "wedge calculation." They concentrated on the top 5% of all pairwise scores within a set of similarly sized exons in an effort to show that comparisons of authentic exon sequences yielded an excess of high scores relative to comparisons of random sequences, even if the individual scores themselves did not achieve statistical significance. They determined the excess to be 830 matches. Unfortunately, this procedure did not allow the identification of just which pairs were included in those 830 matches, thereby precluding a possibility of determining whether the excess similarities were attributable to homologous exons in nonhomologous proteins, or whether the excess was wholly attributable to proteins that shared overall common ancestry but had not been excluded in their initial selection process. The presumption that the excess was solely attributable to pairs of homologous exons in nonhomologous settings is unjustified. The 830 comparison pairs represent only a 1% excess of the matches in the upper 5% range. This small percentage could easily be the result of matches between distantly related proteins that had not been completely culled from the original list. Furthermore, the upper 5% of scores, as the authors note, includes matches as low as 17% identity. For the 40 to 49 residue size range, for example, eight identities in a string of 47 residues would be included. This is an unacceptably low rate of similarity for assessing pairwise homology.

One could inquire what factors other than common ancestry might contribute to such an excess. As Dorit *et al.* note, commonly occurring secondary structure rhythms—α helices, β structure, membrane spanners, and the like—could easily be a factor. Furthermore, certain constellations of amino acids, presumably for valid evolutionary reasons, are not found as frequently as random expectation would specify. The tripeptide Glu-Pro-Asp, for example, occurs less than

Table 1. Significance of pairwise matches reported by others. The pairs of exon sequences studied are taken from table 1 of Dorit *et al.* (*1*); minor discrepancies have been adjusted. Thus their exon "X" for human α-1 (II) collagen would appear to be exon 9, and their exon 2 for human lymphotoxin would appear to be exon 1. Also, their exon 8 for human elastin is denoted "exon 10" in the paper cited by Dorit *et al.* (*14*). The differences in identity percentage between their data set and our is mostly due to the fact that our alignment allows gaps and theirs did not.

Protein	Residues (No.)	Dorit *et al.* ID (%)	This study	
			ID (%)	SD*
Human αI (II) collagen exon Rat mannose-binding protein A exon 2	36 38	50	50	+8.5†
Human apolipoprotein exon 1 Human EGF receptor exon 1	24 29	46	41	+0.8
Human factor XII exon 7 Human factor IX exon 4	34 37	41	41	+7.1†
Human pro-α1 type I collagen exon 47 Human elastin exon 10	34 41	38	38	+1.9
Mouse major urinary protein exon 1 Rabbit collagenase exon 1	32 34	38	38	+2.6
Chicken steroid inducible hsp exon 7 Human neurofilament subunit NF-L exon 4	40 47	38	43	+2.2
Human lymphotoxin (TNF-β) exon 1 Rat asialoglycoprotein receptor exon 3	32 38	36	44	+3.3
Schizophyllum (IG2 gene (fruiting exon 1 Human fibronectin exon 1	40 49	33	33	+2.6
Chicken c-fes proto-oncogene exon 8 Human neurofilament subunit NF-L exon 4	40 47	37	33	+2.3
Mouse α2 type IV collagen exon 5 Human complement C1q B-chain exon 1	60 64	32	38	+4.3†
Murine Ii gene, Ia antigen-associated exon 6b Bovine thyroglobulin exon 18	63 64	30	33	+10.1†
Silkmoth chorion exon 2 Mouse keratin, intermediate filament exon 7	108 112	24	29	−0.5
C. reinhardtii chloroplast *psbA* gene exon 4 Mouse band 3 exon 17	77 84	23	20	+1.6
Human serum albumin exon 4 Human K6b epidermal keratin exon 7	70 73	23	27	−0.1

*The alignment method we used is based on the Needleman-Wunsch algorithm (*15*) and employs the Minimum Mutation Matrix of Dayhoff *et al.* (*16*). We used 64 randomizations (8 × 8) to assess significance in the form of standard deviations above (or below) the random mean. Scores with standard deviations greater than 4 are denoted with daggers.

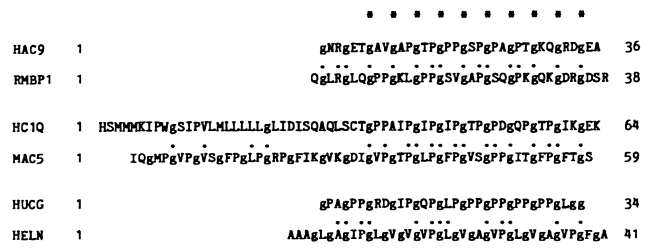

Fig. 1. Alignment of five collagen exon sequences and one elastin exon sequence; all are represented in Table 1. HAC9, human α-1 (II) collagen exon 9; RMBP, rat mannose-binding protein A exon 2; HClQ, human complement ClQ B chain exon 1; MAC5, mouse α-2, type IV collagen exon 5; HUCG, human pro-α-1 type I collagen exon 47; HELN, human elastin exon 10 [reference 8 in (*1*)]. Dots denote identities between members of a pair. All glycine residues (g) are shown in lower case for emphasis; asterisks denote essential glycine positions in the five collagen-type sequences; Numbers at left and right denote first and last residues of expressed amino acids.

half as often as the occurrence of those three amino acids in data banks would predict (*10*). These and other unrecognized biases attributable to natural selection could easily account for the wedge effect.

Dorit *et al.*, anticipating such criticism, performed another operation to see if protein structural constraints alone could account for the excess. For one set of their exons (the 40- to 49-residue size class), they translocated amino terminal segments (blocks of 15 to 25 residues) to the carboxyl termini, leaving the segmental sequences intact. When these rearranged exon sequences were compared, they apparently behaved as if they had been totally randomized (*11*). This is a puzzling result in that protein sequences are *not* random, and one would have expected a slight excess of matches for the block-transposed compared with the truly randomized sequences. The issue is moot, however, because no distinction can be made as to whether the excess high scores are reflections of low-level similarity in homologous proteins or are homologous exons within nonhomologous proteins.

Even if the most sophisticated searching and alignment scheme available had been used, the strategy was doomed unless all sequences root back to a single starter type. With any result short of that extreme, there would always be a question of whether common ancestry was hidden by the relentless rain of amino acid replacement.

On another front, all of the pairs in their initial set of 14 that exhibited a credible level of confidence are sequence types found only in animals. Until similar sequences are identified in plants, fungi, protists, or bacteria, there is no basis for presuming they existed even a billion years ago, never mind at the time primitive proteins were first being assembled. Unless the number of exon types has been immutable from the beginning, each line of descent will have a unique and restricted population of types.

If the number of exon types has been increasing, or even if a steady state has existed whereby the gain of new types is offset by the loss of old, the result is that each line of descent must have a unique population of types. Barring horizontal transfers, shuffling can only occur within an individual genome. The time of first appearance for a particular type of exon is therefore a major factor, because lineages that diverged before that time would be deprived of combinatorial advantages. Prokaryotes and eukaryotes, for example, could only share those exon types that existed at the time of their last common ancestor, 2 billion years ago. Similarly, some types should be unique to plants, others to animals, and so forth. As a result, any calculation based on a simple sampling of a comprehensive database containing entries from all types of eukaryotes would be erroneous.

The main reason to think the number of exon types has increased during evolution is because so many of them are phylogenetically restricted. As an example, the prototypic collagen exon encoding a $(Gly\text{-}X\text{-}Pro)_x$ unit is found only in animals (*12*). This exotype could not have been available as a starter-type unless protist, fungal, and plant lineages had independently lost it.

The formula used by Dorit *et al.* presumes that all members of the underlying set are randomly mixed and follow some uniform distribution. Therefore the pairwise estimate might have been verified by examining the numbers of triple and quadruple matches found. In this regard, collagen exons appeared on the list of 14 three times, and a particular neurofilament exon twice, implying quadruple and triple matches, repectively (*13*). These repeat occurrences are substantially greater than expected for only 14 pairwise matches uncovered from a sample set of 1255. The chances of finding any triple and quadruple matches when the pairwise count is only 14 should be vanishingly small (0.1 and 0.0005, respectively).

Put the other way, if a redundancy (tenfold or more) for some exon type were found, as could have been the case for the widely distributed EGF domain, the formula would yield a small number for the "underlying universe" (N) and certainly less than 500. The reason for so many occurrences of this domain certainly has to do with its being naturally selected in relatively recent times. Using its frequency of occurrence as a sample of the general mixing of all protein modules throughout all time seems misdirected. Even allowing for the fact that ten times more animal sequences have been reported than for plants and fungi, EGF domains should by now have been found in the latter if these domains were equitably distributed.

RUSSELL F. DOOLITTLE
Departments of Chemistry and Biology, Center for Molecular Genetics, University of California, San Diego, La Jolla, CA 92093–0634

REFERENCES AND NOTES

1. R. L. Dorit, L. Schoenbach, W. Gilbert, *Science* **250**, 1377 (1990).
2. R. F. Doolittle, *Methods Enzymol.* **183**, 99 (1990).
3. ———, *Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequences* (University Science Books, Mill Valley, CA, 1987).
4. D. E. Cool and R. T. A. MacGillivray, *J. Biol. Chem.* **262**, 13662 (1987).
5. K. B. M. Reid, *Biochem. J.* **231**, 729 (1985).
6. M. E. Taylor, P. M. Brickell, R. K. Craig, J. A. Summerfield, *ibid.* **262**, 763 (1989).
7. N. Koch, W. Lauer, J. Habicht, B. Dobberstein, *EMBO J.* **6**, 1667 (1987).
8. D. K. Chang, C. M. Venkatachalan, K. V. Prasad, D. W. Urry, *J. Biomol. Struct. Dynam.* **6**, 851 (1989).
9. W. C. Barker, D. G. George, L. T. Hunt, *Methods Enzymol.* **183**, 31 (1990).
10. R. F. Doolittle, in *Prediction of Protein Structure and the Principles of Protein Conformation*, G. D. Fasman, Ed. (Plenum, New York, 1989), pp. 599–623.
11. Although the data themselves were not provided, it was reported that "there is the same significant excess of matches of real exonsd over both the 'scrambled exon' simulations and the 'block-transposed exon' simulation" (*1*).
12. Recently a short (18-residue) collagen-like sequence has been found in a bacteriophage capsid protein [J. H. Bamford and D. H. Bamford, *Virology* **177**, 445 (1990)]. Although intriguing, its origin remains mysterious.
13. This assumes that the collagen exons are derived from a single type.
14. Z. Indik *et al.*, *Proc. Natl. Acad. U.S.A.* **84**, 5680 (1987) (HUMEL).
15. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
16. M. O. Dayhoff, R. M. Schwartz, B. C. Oraitt, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, suppl. 8.

6 June 1991; accepted 12 June 1991

Response: Doolittle has provided a sharp criticism of our paper. For completeness, we call the reader's attention to another critical essay by Laszlo Patthy (*1*).

Our paper is an effort to estimate how many kinds of exons might be used in the construction of modern protein diversity. Although our basic hypothesis is that introns are most commonly lost, we did not seek to identify "fused" exons or motifs in arbitrary proteins to trace such loss. We compared amino-acid sequences of known exons with other exons of the *same* size in

order to get a measure of the "simple" reuse of exons and thus obtain estimates of how many ancestral exons one might infer from a specified set of data. Thus, for example, we chose to treat collagen exons of significantly different lengths (34 versus 60 amino acids) as representing different exons, rather than homologous exons that have diverged in length as a result of fusion or intron loss.

There are several difficulties involved in this procedure. First, we relied on sequence comparisons. As with all measures of similarity, sequence comparisons cannot distinguish between convergence (analogy) and evolutionary relatedness (homology). Our study does assume, however, that pairwise similarity between two exons embedded in otherwise dissimilar proteins reflects common exon ancestry. We explicitly stated that, in the first part of our calculation, we considered shared compositional bias as an indication of homology. Second, we realize the confounding effect of time and sequence change on our ability to purge homologous proteins from the database. Despite our best efforts, the final database of exons may still contain unrecognized related proteins, contributing to the excess of statistically significant pairwise matches in our wedge calculation. We think this effect is a component of our low second estimate. Time and sequence change, however, also hinder our ability to identify homologous exons involved in shuffling events, leading us to overestimate the size of the exon universe. We cannot at this point evaluate the importance of these two counterbalancing biases.

There is no reason to assume either that all exon sequences are monophyletic, or that all possible amino acid chains were present at one time and only a limited set have survived. Our calculation is an effort to trace back, as far as one can, ancestral relationships embedded in the current structure of genes. Any such calculation tries to detect indications of evolutionary homology that have survived the "rain of amino acid replacements." We did not purport to predict the ultimate origins; we only made an effort to see back as far as possible.

The calculation is based on comparing the amino acid sequences of *exons*, drawn from a limited database based on genomic structures, in order to identify similarities that are unusual given that *limited* set of comparisons. We would not have done the calculation the way we did if we thought that the test used by Doolittle to detect significant *single* sequence matches in the total protein database was appropriate for the problem at hand. However, from the nature of our statistical comparison, we would expect a few of our candidate exon shuffles to be in error, as we attempted to work at a 95% confidence limit; we would also expect to miss other exon matches which truly are examples of shuffling but which our methods would fail to identify. One might expect a Poisson error in our number of 14, that is, a standard deviation of ±4. It should be obvious from the two ways of doing the calculation that the true uncertainties in the estimate of the size of the exon universe involve much larger variations than this.

That the exon pairs we identified were drawn primarily from vertebrate proteins may simply reflect the strong phylogenetic biases of the current sequence databases. Although Doolittle asserts that there must be a unique population of exon types for each line of descent, such a conjecture is premature. At best one can only say that certain motifs have not *yet* been found in all kingdoms. Our hypothesis predicts that most exon types will turn out *not* to be phylogenetically restricted.

Our model does assume that all exons have an equal probability of being involved in a shuffling event. We are aware that chance and natural selection must yield a far more complex probability distribution, as certain exon motifs (such as EGF) may be more likely than others to reappear in a variety of unrelated proteins. At present, however, we cannot glimpse the shape of the overall probability distribution for all exon types, and therefore choose the simplest model. As additional sequence information becomes available, we expect to be in a position to refine our assumptions.

Patthy argues that many introns must have been inserted and hence that much of the intron-exon breakup is recent. He adheres strongly to the hypothesis that introns can never slide to a different phase in the reading frame. We think this is too restrictive an assumption and consider introns that are in similar position but out of phase with each other to be the same.

We agree that both Doolittle and Patthy point to specific examples of "matches" that probably are not examples of shuffling of three-dimensional structural elements but rather represent other coincidences. However, we do not feel that those observations change the general force of our argument.

ROBERT L. DORIT
LLOYD SCHOENBACH
WALTER GILBERT
Department of Cellular and Molecular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138

REFERENCES

1. L. Patthy, *Bioessays* **13**, 187 (1991).

26 June 1991; accepted 27 June 1991