Wavelet analysis of DNA sequences

A. A. Tsonis,[1] P. Kumar,[2] J. B. Elsner,[3] and P. A. Tsonis[4]
[1]*Department of Geosciences, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53210-0413*
[2]*Department of Civil Engineering, University of Illinois–Urbana, Urbana, Illinois 61801*
[3]*Department of Meteorology, Florida State University, Tallahassee, Florida 32306-3034*
[4]*Department of Biology, The University of Dayton, Dayton, Ohio 45469*
(Received 11 May 1995)

In this paper we use wavelet analysis in order to probe the localized structure of DNA sequences. We demonstrate that, unlike other conventional approaches, wavelets are able to decompose seemingly homogeneous regions in noncoding sequences into smaller distinct regions that obey their own repetition and construction rules. The significance of this result to gene evolution is discussed.

PACS number(s): 87.15.Mi

The evolution of genetic information and the generation of genes is one of the most challenging problems facing evolutionary and molecular biologists. The principles by which nature produced the genetic information and subsequent generation of genes are still not well understood. DNA sequences are strings of the bases (nucleotides) *A*, *T*, *C*, and *G*. Bases *C* and *T* are pyrimidines, and bases *A* and *G* are purines. DNA sequences are characterized as coding (intron-less) sequences and as noncoding (intron-containing) sequences.

Since the early 1970s, scientists have attempted to discover some kind of order or hidden structure(s) in DNA sequences, to discriminate coding from noncoding regions, to find translation initiation sites, to explore and understand function in genes, etc. [1–4]. With the advent of DNA sequencing techniques in the late 1970's, scientists had the opportunity to probe the DNA for such order. It soon became apparent that a periodicity of 3 reflects the use of codons, but there was no hypothesis at that time how they might be acquired. In a series of papers, however, Ohno [3,4] showed that modern DNA sequences have evolved from primordial blocks of maybe seven nucleotides. According to this, such blocks duplicated many times, and as they did so they also mutated. This led to a variety of sequences found in today's DNA. From his studies he was also able to state that the construction rule was *TG-CA-CT* excess and *TA-CG* deficiency. True enough, *CT* define the first two nucleotides for the codon for leucine, the most abundant amino acid in proteins, and *TA-CG* are found only in promoted regions which constitute a small part of the genome. In an attempt to further elucidate the presence of these periodicities in DNA sequences, Tsonis, Elsner, and Tsonis [5] applied Fourier analysis on coding and noncoding sequences. They found that while noncoding sequences showed spectra similar to those of random sequences, coding sequences revealed specific periodicities of variable length and a common periodicity of three. Furthermore, they were able to reconstruct the spectra of a given *m*RNA from an artificial periodic sequence mutated in such a way as to represent the actual content of amino acids found in proteins. Spectral analysis in DNA sequences was also used by Voss [6], who confirmed the existence of the above mentioned periodicity of 3 and suggested that DNA sequences exhibit spectra appropriate to $1/f^{a}$ noises (see also [7,15]).

Conventional Fourier analysis, however, can only reveal "global" periodicities. Hidden localized periodicities that might provide hints about underlying construction rules cannot be extracted. Such a difficulty can be overcome with the use of wavelets, a mathematical approach that can transform a signal into a sum of smaller waves that can break up more complex signals. Just recently, wavelets were applied to random walks generated from DNA sequences [8] in order to investigate the proposed in [7] long-range correlations in such walks. The purpose of our work was quite different, in that our interest was to search for the construction rules in the actual DNA sequences.

The wavelet transform [9–11] is a localized transform in both space (time) and frequency. In mathematical terms wavelet transforms are integral transforms using integration kernels called wavelets. The wavelet transform of a function $f(t)$ with finite energy is defined as the integral transform with a family of functions $\Psi_{\lambda,t}(u)\equiv(1/\sqrt{\lambda})\Psi((u-t)/\lambda)$, and is given by

$$Wf(\lambda_1 t)=\int_{-\infty}^{\infty} f(u)\Psi_{\lambda,t}(u)du \ , \quad \lambda>0$$

$$=\int_{-\infty}^{\infty} f(u)\frac{1}{\sqrt{\lambda}}\Psi\left[\frac{u-t}{\lambda}\right]du \ . \tag{1}$$

Here λ is a scale parameter, t a location parameter, and the functions $\Psi_{\lambda,t}(u)$ are the wavelets. In case $\Psi_{\lambda,t}(u)$ is complex, we use the complex conjugate $\overline{\Psi}_{\lambda,t}(u)$ in the above integration. Changing the value of λ has the effect of dilating ($\lambda>1$) or contracting ($\lambda<1$) the function $\Psi(t)$, and changing t has the effect of analyzing the function $f(t)$ around the point t. The normalizing constant $1/\sqrt{\lambda}$ is chosen so that

$$\|\Psi_{\lambda,t}\|^2 \equiv \int |\Psi_{\lambda,t}(u)|^2 du = \int |\Psi(t)|^2 dt$$

for all scales λ [notice the identity $\Psi(t) \equiv \Psi_{1,0}(t)$]. We also choose the normalization $\int |\Psi(t)|^2 dt = 1$.

An important property of wavelets called time-frequency localization. The advantage of analyzing a signal with wavelets as the analyzing kernels is that it enables one to study features of the signal locally with a detail matched to their scale, i.e., broad features on a large scale and fine features on small scales. This property is especially useful for signals that are either nonstationary, or have short lived transient components, or have features at different scales, or have singularities. One might see wavelets as the elementary building blocks in a decomposition or series expansion akin to the familiar Fourier series. Thus a representation of the process using wavelets is provided by an infinite series expansion of dilated and translated versions of a *mother wavelet,* each multiplied by an appropriate coefficient. For processes with finite energy this wavelet series expansion is optimal; i.e., it offers an optimal approximation to the original signal in the least squares sense.

Useful information can also be extracted by interpreting the wavelet transform (1) as a time-scale transform. This was well illustrated by Rioul and Vetterli [12], and is sketched below. In the wavelet transform (1) when the scale λ increases, the wavelet becomes more spread out

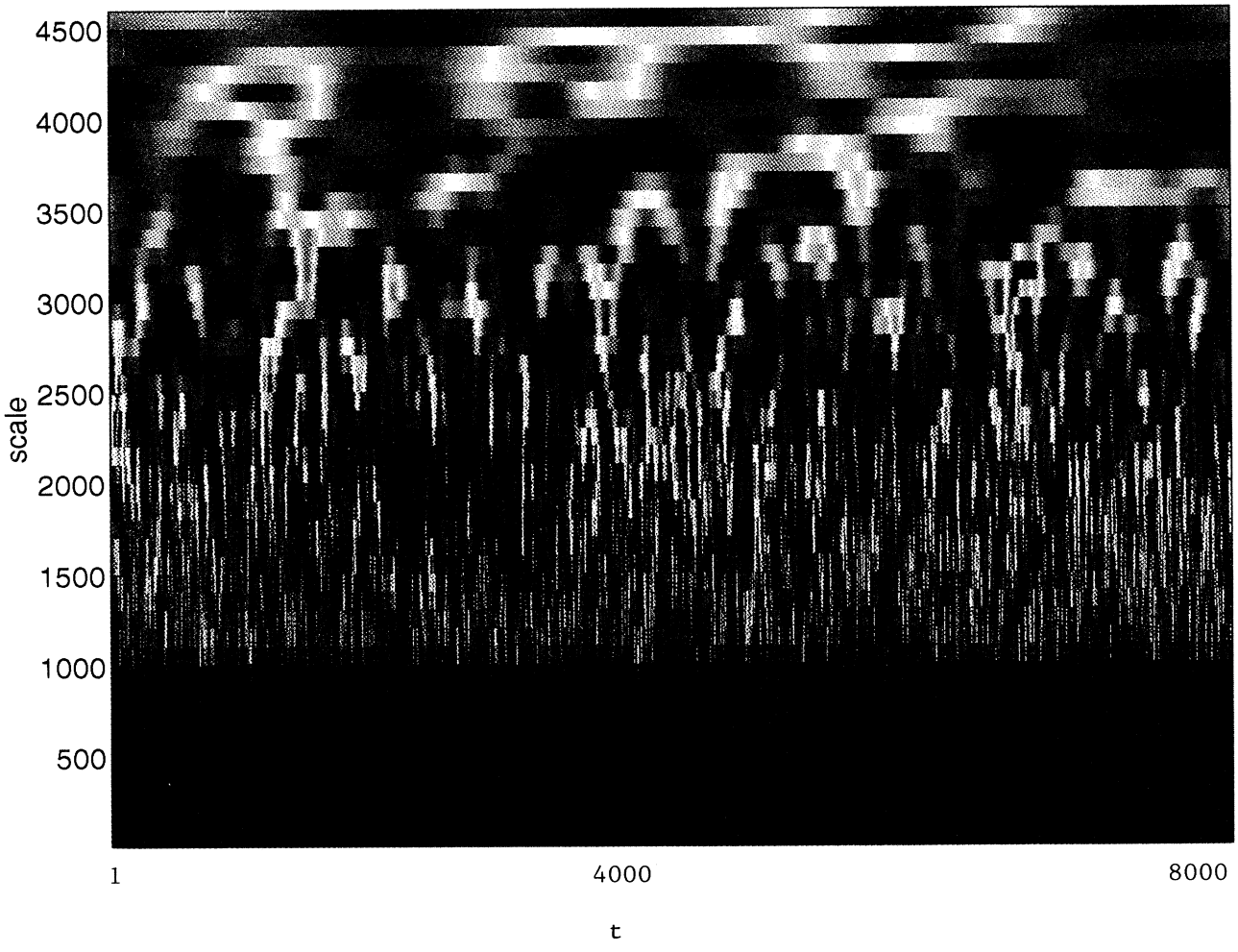

FIG. 1. Wavelet analysis of the genomic sequence of the chicken *c*-myb oncogene, which is 8200 bases long. The wavelet used here is the Morlet wavelet. The scale λ is related to the frequency ω via the relation $\omega = \omega_0/\lambda$, where $\omega_0 = 5$. Time (t) runs from 1 to 8200. The colors can be interpreted similarly to the way peaks are interpreted in a conventional Fourier analysis. Blue corresponds to low power, and red to high power. Such bright colors will indicate strong localized periodicities or construction rules. We observe three distinct "hot" spots centered at scales of 4200, 4400, and 4500 and times 1200, 3400, and 6000 respectively. As is explained in the text, these spots may represent local construction rules in the DNA sequence.

and takes only long time behavior into account, as seen above. However, by a change of variables, Eq. (1) can also be written as

$$Wf(\lambda,t)=\int_{-\infty}^{\infty}\sqrt{\lambda}f(\lambda u)\Psi\left[u-\frac{t}{\lambda}\right]du\ . \tag{2}$$

Since the mapping $f(t)\rightarrow f(\lambda t)$ has the effect of contracting $f(t)$ when $\lambda>1$ and magnifying it when $\lambda<1$, the above equation indicates that, as the scale grows, a contracted version of the function is seen through a fixed size filter, and vice versa. Thus the scale factor λ has the interpretation of the scale in maps.

We have applied wavelet analysis to genomic (more than 90% intron containing) and to coding (intronless) DNA sequences. For mathematical purposes any DNA sequence may be transformed to a sequence of integers in the interval [1,4] (for example $A=1$, $T=2$, $C=3$, and $G=4$). This allows us to consider the sequence as an "observable," $f(t)$, and to apply statistical and mathematical techniques. In our analysis we employed the Morlet wavelet. This wavelet is given by

$$\Psi(t)=\pi^{-1/4}\exp(-i\omega_0 t)\exp(-t^2/2)\ ,\quad \omega_0\geq 5\ . \tag{3}$$

This wavelet is complex, allowing us to extract information about the amplitude and phase of the process being analyzed. The Fourier transform of Eq. (3) is given by

$$\hat{\Psi}(\omega)=\pi^{-1/4}\exp[-(\omega-\omega_0)^2/2]\ ,$$

where ω is the frequency. The Fourier transform of the scaled wavelet $\Psi_{\lambda,0}(t)$ is given by

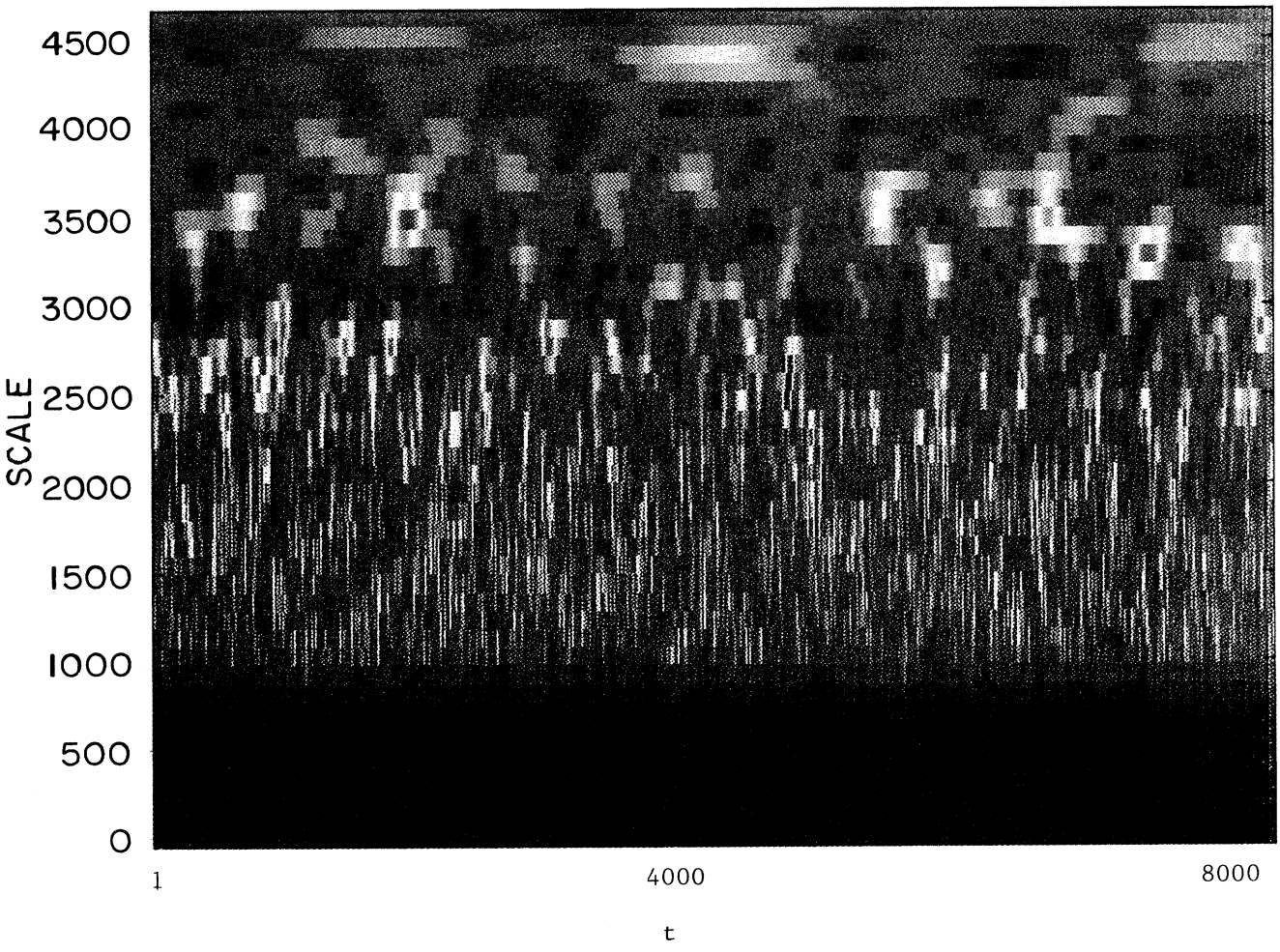

FIG. 2. Same as Fig. 1, but for a surrogate sequence obtained by shuffling the *c*-myb sequence. Such surrogate data have similar statistical properties to the DNA sequence, but are random.

$$\hat{\Psi}_{\lambda,0}(\omega)=\lambda\pi^{-1/4}\exp[-(\omega-\lambda\omega_0)^2/2]\ .$$

The wavelet $\Psi_{\lambda,0}(t)$ has the property that it is centered at t with a spread λ, while when its Fourier transform is supported almost entirely on $\omega>0$ it is centered at ω_0/λ and has a spread $1/\lambda$.

Figure 1 shows the results for the chicken *c*-myb oncogene. This is a genomic sequence 8200 bases long. We observe three distinct and pronounced "hot" spots (indicated by red color) at high scale (low frequency) values. At higher frequencies such features are hard to distinguish. The question now arises: Are those hot spots statistically significant (and thus represent nonrandom features of the sequence), and if they are what do they correspond to? In order to answer the first part of the question we generated surrogate sequences by shuffling the above DNA sequence. This way we destroy whatever dynamics or features in the sequence, thereby producing random sequences that exhibit the same statistical properties (such as mean, variance, distribution, etc.). Figure 2 is a typical result from the surrogate sequences. Here we observe that at $4200<(\text{scale})<4500$ there are no pronounced hot spots. At other higher frequencies (smaller scales) the features of the two figures are not significantly different. Thus the hot spots in Fig. 1 represent significant features in the DNA sequence. From the analysis of a number of such sequences (another example is shown in Fig. 3) and their surrogates, we find that those hot spots at low frequencies are characteristic features of noncoding sequences which are not likely the product of a random string of bases. Thus they indicate certain "construction" features in the sequences. A similar analysis of coding sequences (see, for example, Fig. 4) does not reveal such pronounced features, as they often appear indistinguishable from random strings of similar statistical character. This is due to the fact that coding regions are much smaller than the noncoding regions, and thus possible repetitions or construction rules do not provide a large enough sample size for statistically significant differences between coding sequences and their

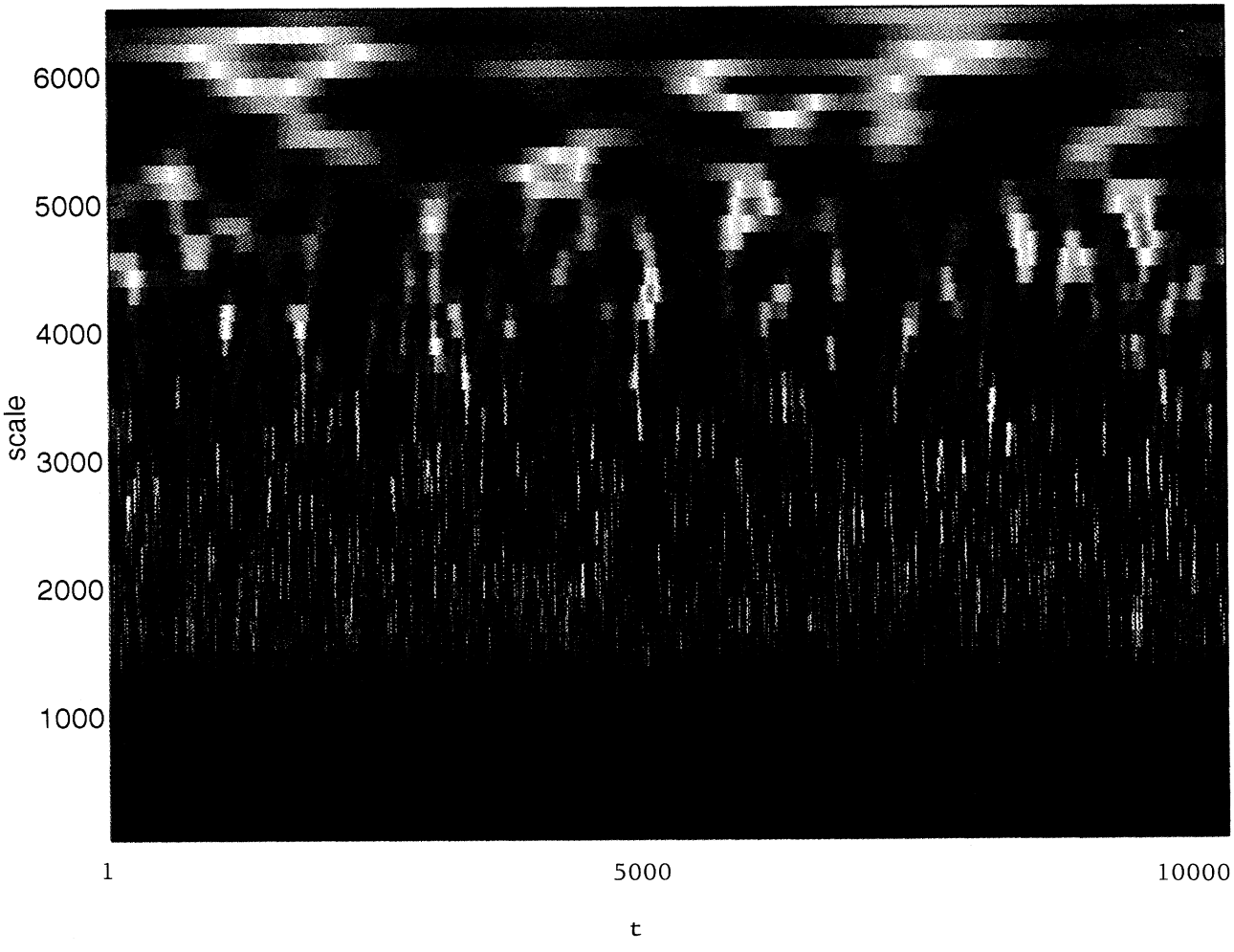

FIG. 3. Same as Fig. 1, but for the genomic gotglobe gene (noncoding, 10 200 bases long).

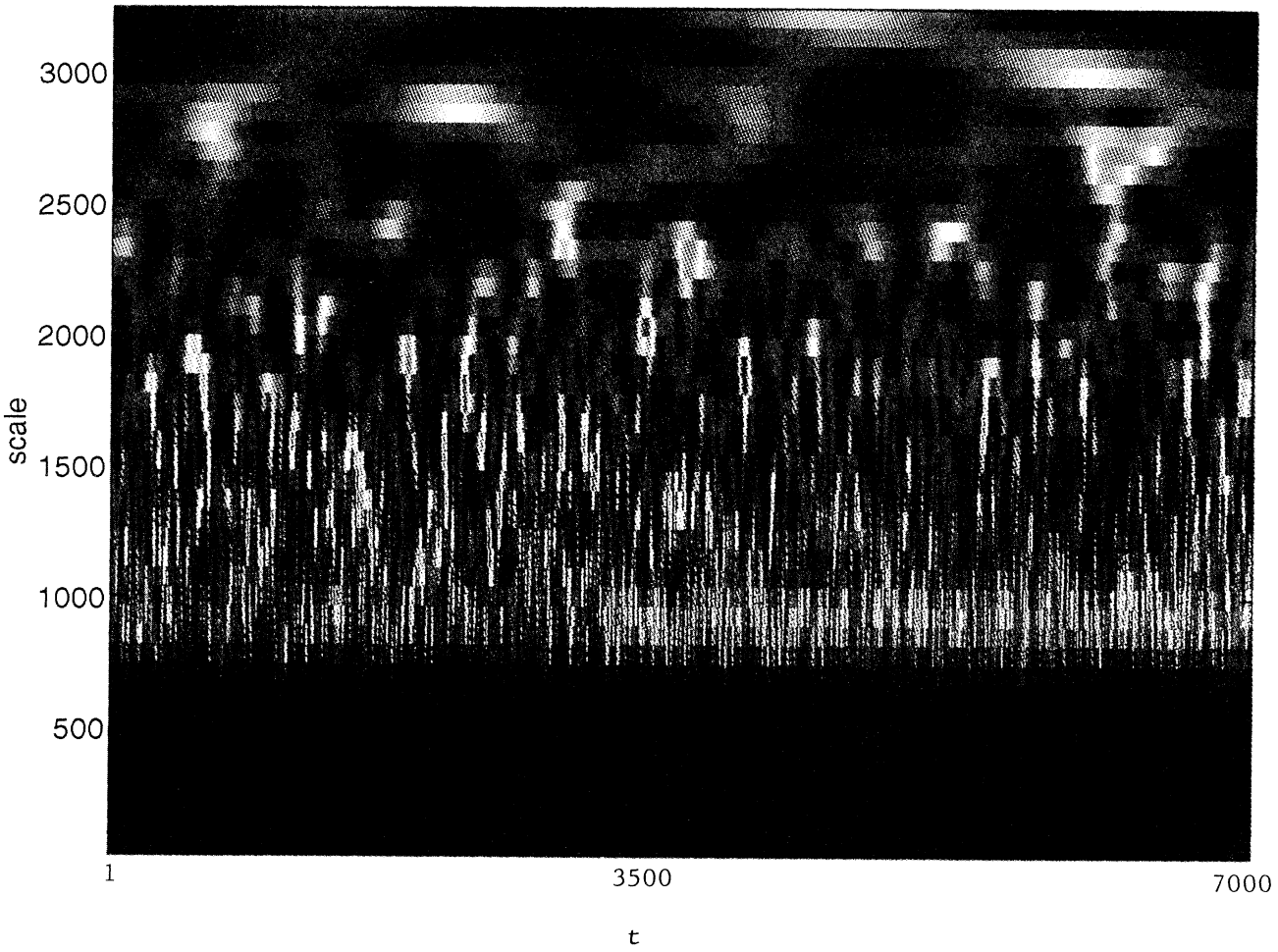

FIG. 4. Same as Fig. 1, but for the coding sequence of human *b*-cardiac MHC (6000 bases long).

surrogate data. Having established the above, we proceed with investigating exactly what those features in the noncoding sequences indicate, and what function do they serve.

In order to obtain insights into gene evolution and structure, we compare the above results with those obtained from conventional analyses of repeated elements (or homologies) in DNA sequences. The most common approach is the dot matrix analysis [13]. This technique is explained in Fig. 5, which shows results for the *c*-myb DNA sequence analyzed in Fig. 1. Two distinct areas (units) are revealed one from 750 to 3500 nt (nucleotides) and the other from 3500 to 8000 nt. Thus, Fig. 5 suggests that this genomic sequence has been constructed by two different sets of repeating units or construction rules. The first prevails in the region 500–3500 nt, and the second in the rest of the gene. The first unit coincides with the two first hot spots seen by wavelet analysis, and the second with the third. If we set the parameters to be more stringent (20 matches out of 30 bases long segments) the two regions remain the same (Fig. 6). In fact in this case the noise is cleared, and the second region becomes more defined in the region from 5000 to 7500 nt which corresponds to the third hot spot of Fig. 1. Thus a comparison of Figs. 1 and 5 or 6 indicates that the wavelet analysis was able to decompose the first region into two. Another example is the gotglobe gene. In the dot matrix analysis (Fig. 7) we see a uniform distribution of "dots," indicating widespread periodicities throughout the gene. This suggests that one general construction rule is responsible for the whole gene. With wavelets, however (Fig. 3), we discover that the gene is broken up into two different regions that suggests the possibility of two general construction rules. Similar results are obtained from other genomic sequences.

The above results point out limitations in the dot matrix analysis, and suggest wavelets as a complimentary approach. More importantly, the results here bear

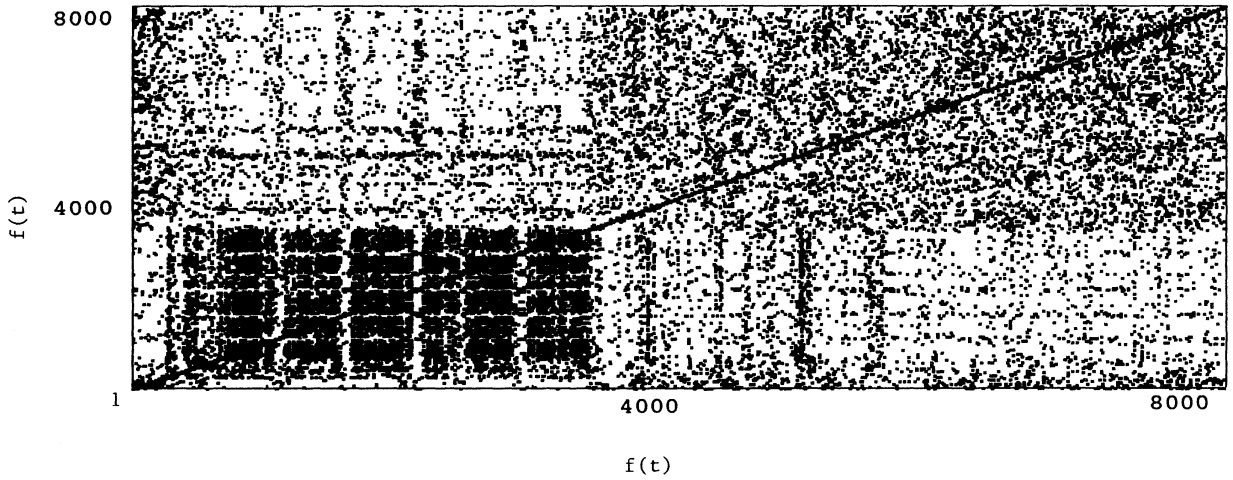

FIG. 5. The dot matrix analysis compares a given DNA sequence $f(t)$ with itself [i.e., with $f(t)$], with the help of a sliding "window" whose length may vary. For example, if we compare a given sequence 1000 bases long to itself, and the window is assumed to be 1000 bases long in a two-dimensional plot with coordinates $[f(t),f(t)]$, the result will be the diagonal line. This will indicate 100% homology. If the comparison is attempted with shorter windows (say 20 or 50 bases long), and repetitions do exist, diagonal lines are drawn in the regions where the sequences are repeated. Such an analysis can thus scan a given DNA sequence for repetitions within it. In this figure we compare the *c*-myb DNA sequence analyzed in Fig. 1. The length of the window is 20, and we assume a homology limit of 10. Accordingly, we consider the segment $1<t<20$ and we slide the window across the sequence. If at least ten bases in the window and in the segment $1<t<20$ match, then the diagonal of that region in the $[f(t),f(t)]$ plane is drawn. Then we consider the segment $2<t<21$ and repeat the sliding procedure, and so on, until we scan the whole DNA sequence. As we can see, two distinct areas (units) are revealed, one from 750 to 3500 nt (nucleotides) and the other from 3500 to 8000 nt (note that due to the relative short window length, the diagonals are very short and appear like dots). Such a result indicates that the repetitive units must be different in the two regions (otherwise the dots will fill the plane uniformally).

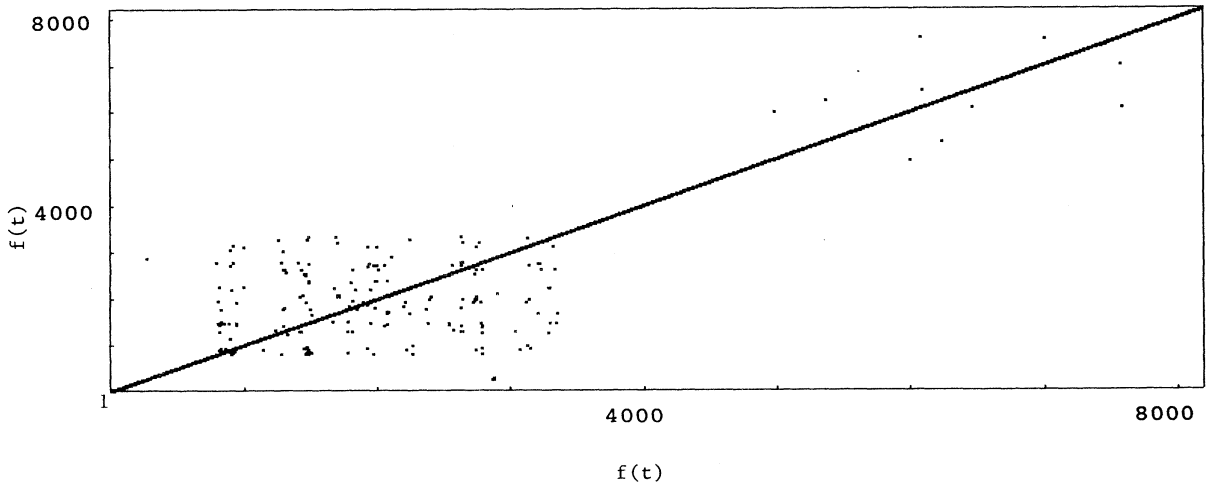

FIG. 6. Same as Fig. 5, but for a window of 30 and a homology limit of 20 (66%).

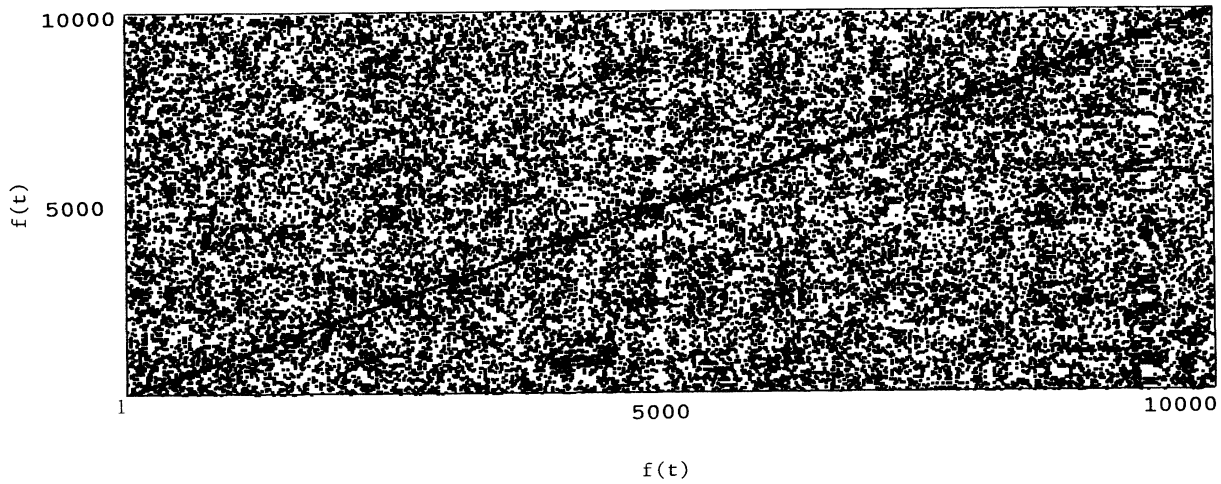

FIG. 7. Same as Fig. 5, but for the gotglobe gene used in Fig. 3. The window is 20 and the homology limit is 10 (50%). The distribution of the dots is rather uniform, suggesting one construction rule (repetition unit) throughout the gene.

significance if we consider the aspect of gene evolution and structure. In both cases wavelets were able to decompose seemingly "homogeneous" region in DNA in smaller distinct areas. Given the type of analysis we presented here, these smaller areas have probably been evolved from different primordial blocks. In this respect our data suggest that genes have been constructed not only by one primordial block, but by more. This could indicate that these different constructions were combined later, possibly by a similar mechanism to exon shuffling. Such "supershuffling" could lead to generation of genes by combining long genomic sequences that contain introns and exons. This in turn would suggest that exons and introns may have been made by the construction principle, as proposed by Ohno [4]. Accordingly, our results offer support to the intron-early theories [14], which postulate that ancient genes were made from both introns and exons rather than the intron-late theories which postulate that modern genes have arised from initially uninterrupted genes by later insertion of introns.

[1] J. C. W. Shepherd, CSH Symp. Quantum Biol. **47**, 1099 (1982).
[2] A. D. Nazarea, D. P. Bloch, and A. C. Semrau, Proc. Natl. Acad. Sci. U.S.A. **82**, 5337 (1984).
[3] S. Ohno, Proc. Natl. Acad. Sci. U.S.A. **85**, 4378 (1988).
[4] T. Yomo and S. Ohno, Proc. Natl. Acad. Sci. U.S.A. **86**, 8452 (1989).
[5] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, J. Theor. Biol. **151**, 323 (1991).
[6] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[7] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sclortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[8] A. Arneodo, E. Barcy, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).
[9] Y. Meyer, *Wavelets: Algorithms and Applications* (SIAM, Philadelphia, 1993).
[10] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, Geophysics **47**, 203 (1982).
[11] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, Geophysics **47**, 222 (1982).
[12] O. Rioul and M. Vetterli, IEEE Signal Proc. Mag. **33(2)**, 14 (1991).
[13] G. von Heinge, *Sequence Analysis in Molecular Biology* (Academic, San Diego, 1987).
[14] P. Senapathy, Science **268**, 1366 (1995).
[15] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, Biochem. Biophys. Res. Commun. **197**, 1288 (1993).

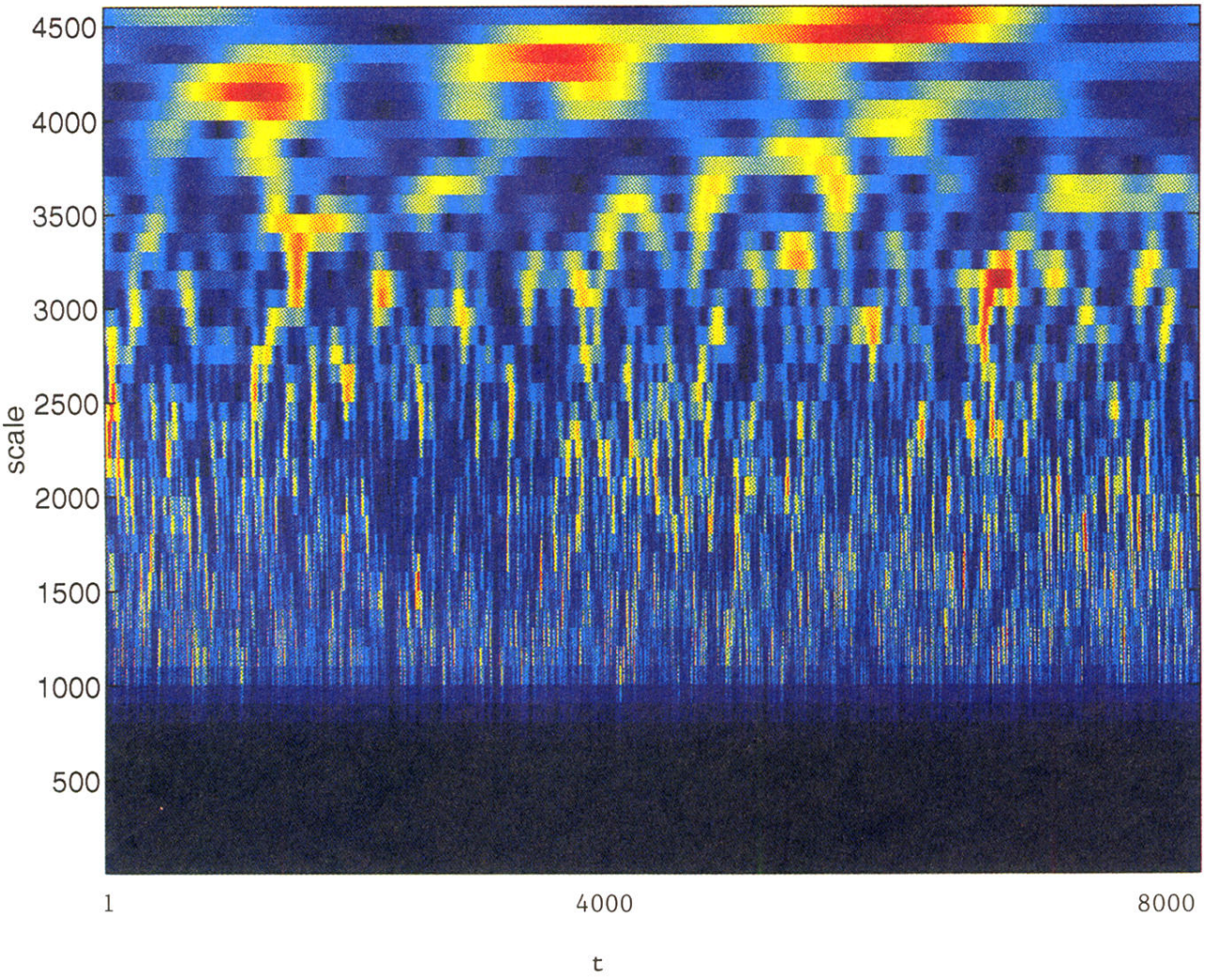

FIG. 1. Wavelet analysis of the genomic sequence of the chicken *c*-myb oncogene, which is 8200 bases long. The wavelet used here is the Morlet wavelet. The scale λ is related to the frequency ω via the relation $\omega=\omega_0/\lambda$, where $\omega_0=5$. Time (t) runs from 1 to 8200. The colors can be interpreted similarly to the way peaks are interpreted in a conventional Fourier analysis. Blue corresponds to low power, and red to high power. Such bright colors will indicate strong localized periodicities or construction rules. We observe three distinct "hot" spots centered at scales of 4200, 4400, and 4500 and times 1200, 3400, and 6000 respectively. As is explained in the text, these spots may represent local construction rules in the DNA sequence.

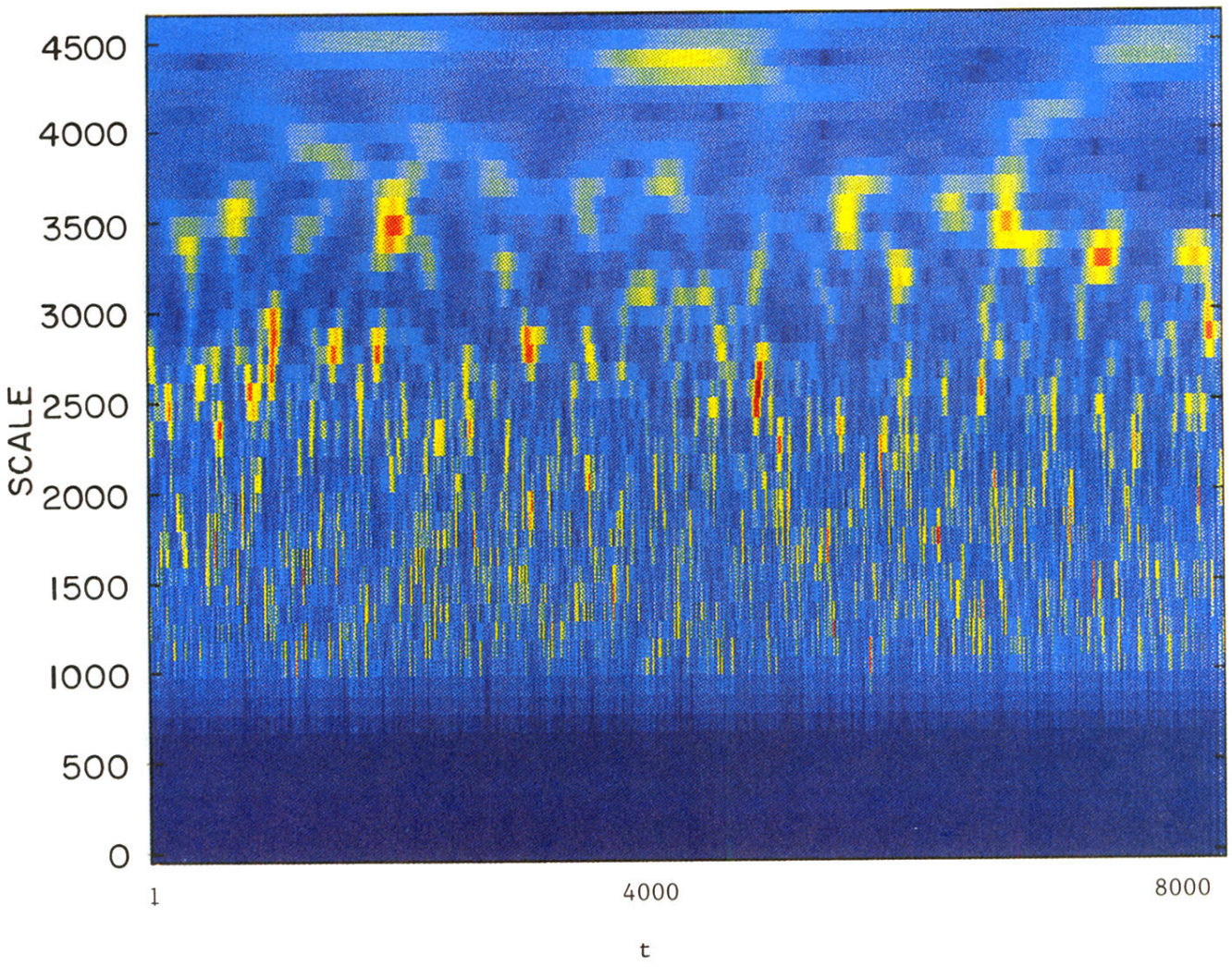

FIG. 2. Same as Fig. 1, but for a surrogate sequence obtained by shuffling the *c*-myb sequence. Such surrogate data have similar statistical properties to the DNA sequence, but are random.

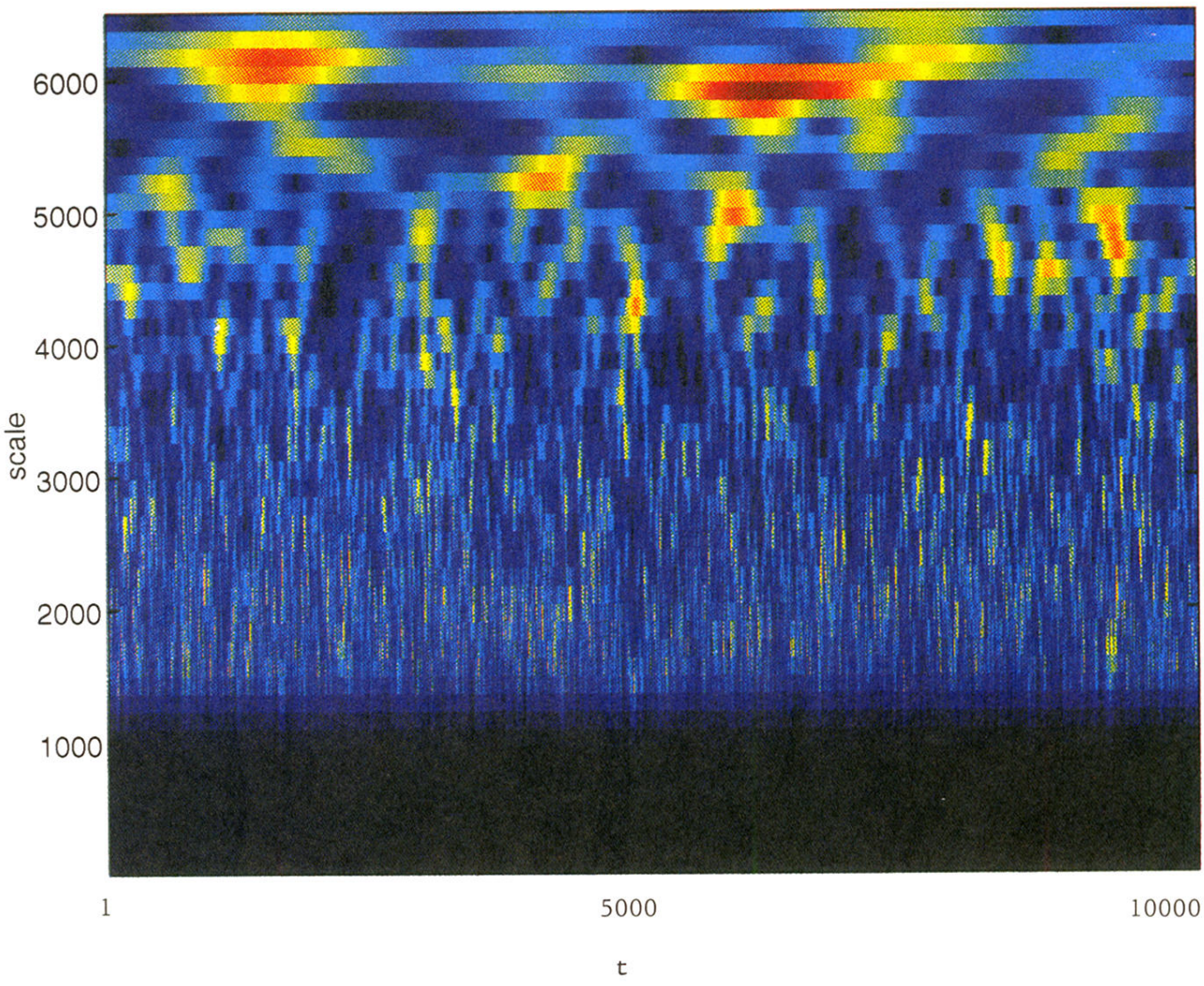

FIG. 3. Same as Fig. 1, but for the genomic gotglobe gene (noncoding, 10 200 bases long).

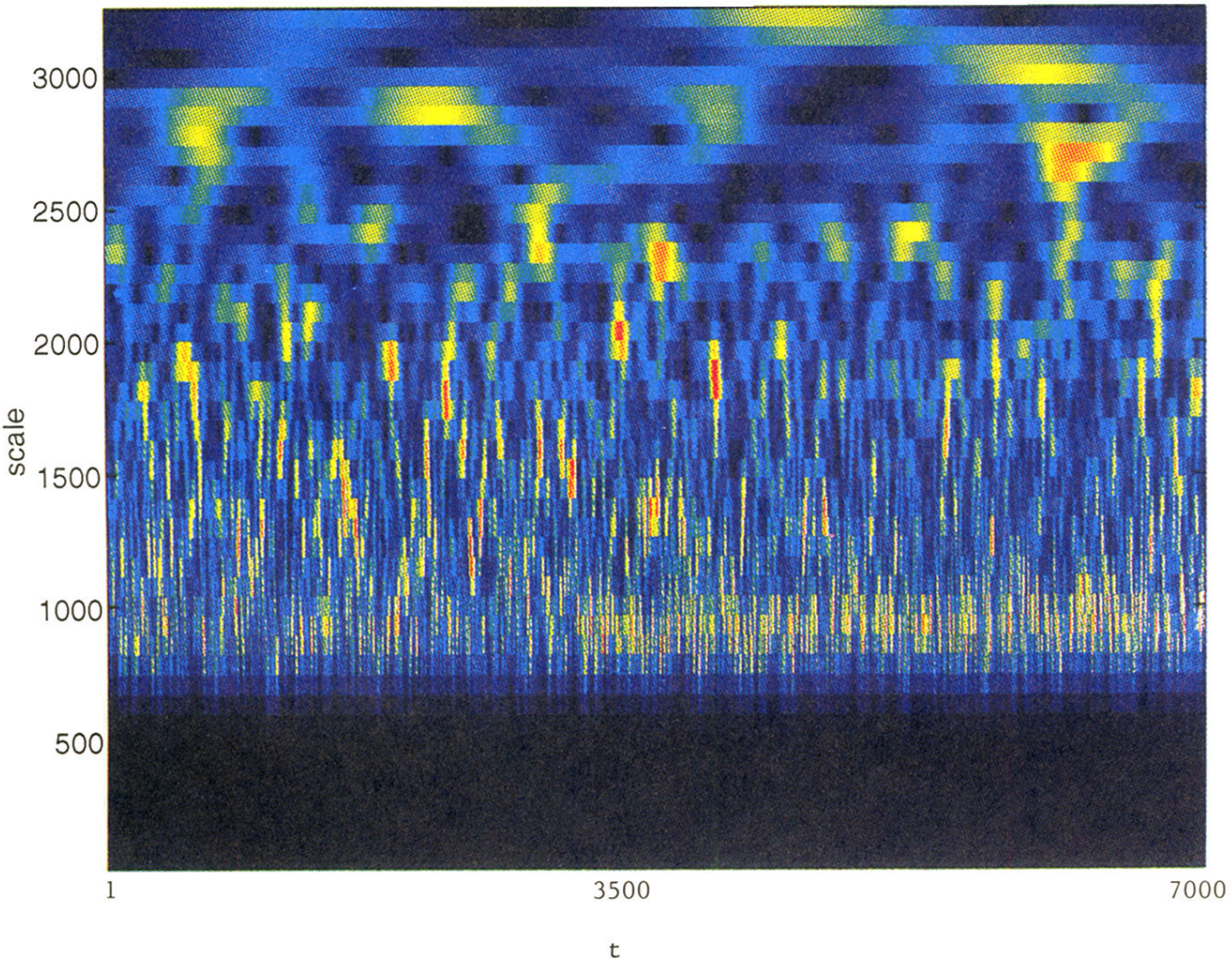

FIG. 4. Same as Fig. 1, but for the coding sequence of human *b*-cardiac MHC (6000 bases long).