DNA and protein tetragrams: Biological sequences as tetrahedral movements

Clifford A. Pickover

IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

A graphical approach is introduced for representing information-containing sequences in biology. In particular, the procedure takes DNA or protein sequences containing n *bases or amino acids, respectively, and computes* n *three-dimensional real vectors. When displayed on connected tetrahedra these characteristic patterns appear as DNA or amino acid tetragrams* T(n). *Experiments indicate that these tetragrams are sensitive to certain important patterns in the sequence of bases and allow the human observer to visually detect various properties of biological sequences. The system presented is special in its focus on the fast characterization of the progression of sequence data using a graphics supercomputer with several controlling parameters.*

Keywords: DNA sequences, protein sequences, tetragrams, patterns and sequences

INTRODUCTION

Among the methods available for characterizing information-containing sequences in biology, computer graphics is emerging as an important tool.[1,2] Fairly detailed nongraphical comparisons between sequences like DNA and proteins are useful and can be achieved by a variety of brute-force statistical computations,[2–4] but sometimes at a cost of the loss of an intuitive feeling for the structures. Differences between sequences may obscure the similarities when standard approaches alone are used. The graphical approach described here involves the mapping of sequence data to a three-dimensional pattern on connected tetrahedra to visualize similarities between, and biochemical properties of, DNA and amino acid sequences. The idea of mapping genetic sequence information to graphic patterns traced by a computer has precedent in H curves,[5,6,8] DNA vectorgrams and faces,[9] and standard and cumulative line-extension formats.[7] Color DNA and amino acid *tetragrams*, in this paper, can complement these past approaches especially when computed and displayed on powerful graphics workstations that allow real-time rotation, magnification, lighting, and hidden-surface removal for these representations.

Color Plates for this article are on page 17.

Address reprint requests to Dr. Pickover at IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA.
Received 27 March 1991; accepted 5 August 1991

METHOD

DNA is symbolically represented by a long string of characters (i.e., G, C, A, and T) representing the four DNA bases. Using this representation, the human observer may find difficulty in distinguishing among different sequences, assessing base composition, and finding various patterns. A technique that has proved useful in overcoming this drawback is the DNA tetragram. A computer inspects the DNA sequence one base at a time and assigns a tetrahedral direction of movement corresponding to the base. Therefore, each letter causes a vector to be drawn from a point in the center of a tetrahedron to one of four adjacent vertices. This procedure is repeated, and therefore a pattern characteristic of the DNA sequence is drawn in three-space. The fact that all four directions are spatially equivalent avoids some of the arbitrariness inherent in other possible assignment schemes. Similarly, the tetragram can be used to represent protein sequences by assigning the 20 amino acids to four categories: polar, nonpolar, positively charged, and negatively charged.

Color Plate 1 shows DNA tetragrams computed for two viruses—one for the viral harvey murine sarcoma DNA (lower right) and the other for the Kirsten sarcoma virus (upper right). Each small sphere represents a single base, and colors are assigned as follows: G, red; C, yellow; A, green; and T, blue. The Kirsten sarcoma virus clearly shows a direction trend that is not the same as the viral harvey murine DNA and has visually distinct (AT) regions. Other obvious periodicities are evident. When viewed from different angles, the viral harvey murine sarcoma tetragram clearly shows a general red–yellow (G–C) trend. Three different regions are made noticeable by the tetragram: The two terminal domains are separated by an intermediate segment with several base repeats of (GC)-rich sequences. The rightmost domain appears to be more random.

The two transparent spheres in Color Plate 1 indicate where the termini of the tetragrams would be if the DNA sequences were each composed of a random string of bases, the outer sphere corresponding to the Kirsten virus. The radius r for each sphere is computed from $r = L\sqrt{N}$, where

N is the number of bases in a sequence, and L is the length of the step taken from each base. Other experiments, not illustrated in this paper, include tetragrams for a human bladder oncogene. The bladder oncogene gives an extremely long, linear tetragram that travels roughly ten times further then expected for random DNA. For the human somatostatin I gene, the regions of repeating T–A are clearly seen. For *X. laevis* oocyte 5S DNA, the (A T)-rich spacers are easy to see. Tetragrams were computed also for the simian (African green monkey) immunodeficiency virus, the human immunodeficiency virus type 1 (HIV-1), and the human T-cell leukemia virus type 1 (HTLV1). The immunodeficiency viruses are quite similar, particularly in the first 80% of the sequence where they show a long, fairly parallel, linear trend. The leukemia virus has a very different trend, suggesting a very different base composition. One also can construct plots of the distance D traveled by the tetragrams as a function of base number N. When there is a predominance of a particular base, the tetragram grows faster, and this is seen as a continually changing D-versus-N curve. These plots also can show where periodicities are. When there is roughly an equal mix of bases, the tetragram remains stationary, and, for example, this can be seen as a plateau after base 200 in the D-versus-N curve for the viral Harvey murine sarcoma virus. A subsequent rapid rise in the D-versus-N curve corresponds to the sudden insertion of a (GC)-rich region. For the bladder oncogene, the D-versus-N curve can be roughly divided into two regions having two different slopes. Interestingly the change in slope at about $N = 1350$ occurs very close to the place that separates control signals and "enhancer regions" from the coding groups to follow. Various repeating patterns are obvious in the D-versus-N plots for the *X. laevis* sequence.

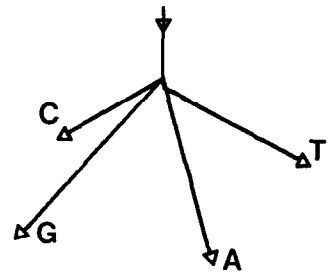
Color Plate 2 shows amino acid tetragrams computed as discussed, with the following coloration: polar, red; nonpolar, yellow; positivcly charged, blue; and negatively charged, green. The origin of the amino acid tetragram is at the N-terminus. Color Plates 2a, b, and c compare the amino acid sequences of hen egg white lysozyme, human lysozyme, and bovine α-lactalbumin (a milk protein). The lysozymes and α-lactalbumin are evolutionarily related.[10] All three tetragrams exhibit a Z-like pattern corresponding to a predominately nonpolar stretch of amino acids, followed by a polar stretch, followed by a nonpolar stretch. In fact, the two forms of lysozyme give rise to very similar tetragrams despite about 40% difference in amino acid composition. Interestingly, the rightmost stretch of amino acids (corresponding to the first 40 residues) forms an obvious subregion in the tetragrams for the lysozyme sequences, and this corresponds precisely to the buried helical core, known from crystal data. Another helical core, from residues 101 to 129, clearly corresponds to the leftmost run of spheres in the tetragram. Lysozyme, like the globins, conforms to the principle of "hydrophobic-in–hydrophilic-out," and the walk of the tetragram in these two core areas indicates a predominance of nonpolar residues. It has been hypothesized that rapid evolution of α-lactalbumin from lysozyme caused divergence of its sequence composition.[10]

Amino acid tetragrams are useful for showing functionally important regions in a protein. I have computed amino acid tetragrams for a range of protein sequences, such as human erythrocyte glycophorin, a transmembrane protein. A large stretch of nonpolar residues from positions 73 through 95 is seen clearly as a linear string of yellow balls, and this is the region that passes through the membrane. For calf histone 2a, the tetragram clearly shows that the first 36 residues contain 12 positive charges and no negative charges. It has been hypothesized that this region interacts with the negative phosphates of a DNA double helix. Collagen displays a long linear trend, suggesting a fairly constant ratio of polar to nonpolar residues.

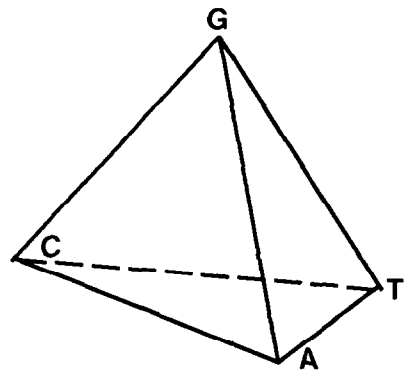
Finally, tetragrams can be used to show evolutionary relationships between proteins. Color Plate 3 shows an amino acid tetragram for a cytochrome C sequence from a human (top), spider monkey, penguin, rattlesnake, honeybee, and spinach (bottom). As expected, more closely related organisms give similar tetragrams. The spinach and human molecule have about 50% difference in amino acid composition. When rotating the tetragrams using a computer, it is easy to identify changing and invariant regions through an evolutionary sequence of proteins.

COMPARISON WITH OTHER METHODS

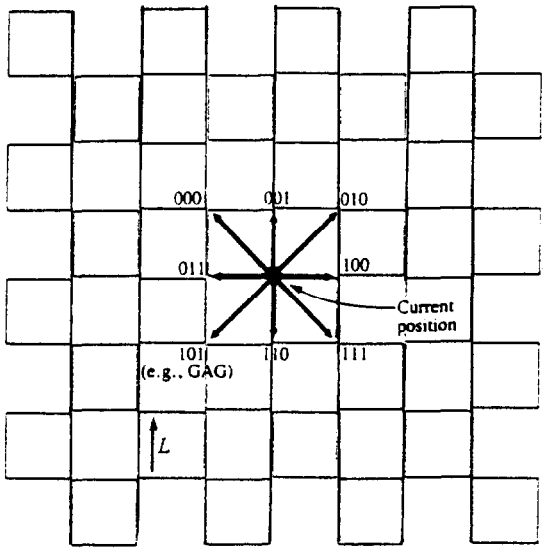
Readers may be interested in how the approach in this paper differs from certain previous methods. As background, *H-curves* use five directions to trace out a pattern in three dimensions. Four directions are arranged at the corners of a square, and a fifth direction is used to show the order of the nucleotides in the DNA sequence:[8]



With the DNA tetragrams, the connected tetrahedra can be diagramed as:



DNA vectorgrams, on the other hand, operate on a binary representation of DNA. The Gs and Cs are assigned the value of 1, and the As and Ts are assigned a value of 0. The vectorgram groups the bases three at a time to direct a trace in eight possible directions on a cellular lattice of length L:[9]



Cumulative line-extension formats are similar to H-curves in that one direction represents position on a sequence, while Gs and Cs produces an upward movment, and As and Ts produce a downward movement.[7] See the respective papers for details.

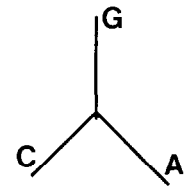
Color *DNA tetragrams,* in this paper, can complement these past approaches especially when computed and displayed on a graphics supercomputer. The relative advantages and disadvantages of these various methods are discussed here. In contrast with many past methods, the tetragram's use of color, hidden surfaces, transparency, and depth cueing on a graphics supercomputer make it of significant interest. I suggested that a transparent or opaque sphere be placed at the origin of the tetragram to indicate how far the sequence would be expected to travel by chance alone. This gives a rough idea of how nonrandom the sequence is. A color tetrahedral axis is placed at the origin to help orient the viewer. Color is extremely useful because it helps the viewer, at a glance, determine sequence composition. Hidden surfaces and depth cueing are used to produced easy-to-interpret three-dimensional (3D) tetragrams. This is particularly useful for following complicated spatial trajectories. Spheres are used because hidden surfaces and shading are easy to see and 3D relationships easy to understand. Lines could be used as well. A graphics supercomputer is used to rotate, magnify, translate, and light the structures in real time. As many as 20,000 nucleotides can be manipulated easily in real time using a Stellar GS 1000 computer. Hidden-surface and shading changes are also manipulated in real time. The C program that computes the tetragram is quite portable and requires less than a second to create a graphics metafile, even for several thousand bases. The metafile contains all the information required to produce the plot and can be edited with a text editor to change color, sizes, and other parameters. Another few seconds are required by a display program to read and display several thousand spheres specified in the metafile. Once displayed, all interactions are in real time.

Perhaps the most important geometrical property of the tetragram is that, as opposed to H-vectors and cumulative line-extension formats, *all directions are equivalent and spatial.* This means that a long linear trend directly and visually implies a significant DNA feature. With H-curves it is sometimes difficult to tell at a glance if a spatial persistence in one direction is due to the predominance of a particular base or due to the fifth spatial dimension being used to represent sequence position—which necessarily produces long drawn-out structures. Tetragrams have neither of these problems. We therefore obtain the greatest spatial resolution with the tetragram.

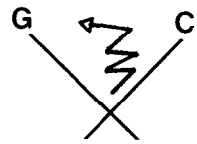
Another possible drawback of H-vectors is that there is an arbitrariness as to how the bases are assigned. For instance there are 4! = 24 ways that 4 base vectors can be assigned to the 4 nucleotides. With tetragrams, the symmetry properties avoid such arbitrariness.

Note that H-curves do have certain advantages over tetragrams: For example, H-curves cannot travel back on themselves. Related to this is the fact that the height of an H-curve will indicate the number of nucleotides in a sequence, whereas tetragrams do not necessarily show this. Comparing the tetragram with the vectorgram: The DNA vectorgram groups bases three at a time, which is potentially useful, but since it operates on a binary representation of DNA, the vectorgram cannot distinguish sequences like ATATA and AAAA that have different folding properties.

Notice that the tetragram can be oriented easily so that one of the axes is pointed to the viewer, thereby isolating three of the four bases.



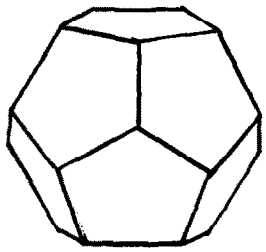
For the tetragram, note that a sequence that alternates between two bases executes a zig-zag in a plane, such as:



As would be expected, various common short-sequence control elements in the DNA can give signature patterns.

The H-curve, tetragram, vectorgram, and cumulative line-extension format can show both short-range detail and global nucleotide distribution trends. The sphere size in the tetragram is arbitrary. I chose a radius such that the spheres just touch. However, by increasing the sphere size used in the tetragram, while keeping the tetrahedron distances the same, a graphic smoothing takes place and global features can be emphasized.

Other avenues for future research include the mapping of the information using more direction vectors. For example, it would be interesting to use other uniform polyhedra to represent more than 4 possible choices. In particular, the dodecahedron's 20 vertices can be used to direct vectors for 20 different amino acids:



I have tested this for the sequences coding of human lysozyme, hen lysozyme, and lactalbumin. Even though these kinds of vectorgrams are not as straightfoward to interpret, I mention them here because they are sensitive to changes that the condensed tetragram representation (described above) cannot show. In particular, if one were to compare structurally or evolutionary related proteins, which often differ only in amino acids within the same tetragram category (e.g., nonpolar-to-nonpolar substitution), the tetragrams would be identical. When computing the dodecagram for the lysozymes, I used the same four color codes as were used with the tetragram—even though there were now 20 different position possibilities. I also clustered the 4 categories of amino acids in spatially separated positions in the dodecahedron. As with the DNA tetragram, even though the sequence of amino acids does not seem to form any conspicuous pattern by a casual inspection of the amino acid sequence, the *dodecagram* has a long directed walk with various twists and turns corresponding to changes in amino acid composition. Like the tetragram, the two lysozymes appeared similar in the dodecagram representation. The α-lactalbumin sequence, however, was quite different, since it has about a 62% different amino acid composition from the lysozyme molecules. The dodecagrams emphasize this difference more than the tetragrams. Perhaps dodecagrams and tetragrams can be used together to represent amino acid sequences, the dodecagram being sensitive to amino acid composition, and the tetragram being sensitive to the charge characteristics of the amino acids. In the future it would be interesting to cluster or color code the amino acids according to their relative frequency of occurrence in proteins. For example, ALA and GLX occur relatively frequently in proteins, whereas TRP and HIS are relatively infrequent. Also, in some fibrous proteins, certain amino acids appear in periodic sequences. Collagen has a preponderance of glycine, proline, and hydroxyproline resides which occur in the periodic sequence -GLY-*X*-*Y* where *Y* is frequently proline or hydroxyproline. This periodicity would show up in the dodecagram.

CONCLUSION AND SUGGESTIONS FOR FUTURE EXPERIMENTS

As a result of the proliferation of biological sequence information,[11] which has been far greater than ever anticipated, it is useful to develop tools to help characterize both small- and very large-scale sequence information. An alternative method used to capture sequence periodicities is the power-spectrum approach.[12,4] Although this technique can be very illuminating, in many applications it does have certain significant drawbacks. For example, power spectra are phase insensitive. One nontrivial consequence of this is that both very orderly and random data can, in theory, give rise to similar spectra.[13] On the other hand, tetragrams $T(n)$ employ a computationally fast and conceptually simple algorithm that directs the observer's eye to the precise sequence area of interest. The DNA and amino acid tetragrams complement one another. Amino acid tetragrams may be used when researchers are looking for patterns and similarities in the structure and function of proteins, whereas DNA tetragrams can be used to examine patterns and differences in the statistical properties of both coding and noncoding regions. Obviously, colors can be used in other ways, e.g., for differentiation interons and exons, and for showing active sites of proteins.

In conclusion, the different DNA sequences tested produce different looking tetragrams; some travel in one direction along the connected tetrahedra, while others travel in an opposite direction. As might be expected for DNA, the tetragram is in general not random. Interspersed repeats are manifest by repeating motifs on the tetragram. This is evident, for example, in *X. laevis* oocyte 5S DNA. It is possible to compare similar sequences and see where differences occur. For example, the human bladder cancer gene was graphically compared with its normal homolog by superimposing the sequences at their 5′ ends. After the point in which the mutation occurs, the graphic pattern bifurcates because even a single base difference can cause a difference in the tetragrams.

Tetragrams are useful in two ways: They provide a qualitative and comparative measure of patterns and periodicities in information-containing molecules, even for segments having significantly different base or amino acid content; and they can be used to search for many pronounced structural features within one sequence with a single calculation. Sometimes it is useful to obtain a visual representation of periodic patterns that is *independent* of the labeling of bases, and the tetragram should yield the same kinds of shapes whether the base sequence has many simple repeats of the type *ggcggcggc* or *aagaagaag*. With such a representation, one could make visual comparisons between different segments of DNA with different base content. This is not the case for H-curves.

Since $T(n)$ presents both nucleic acid and amino acid sequence data in a way that can be visually interpreted by the researcher, sequence characterizations are facilitated. The interactive nature of the research station allows for the rapid generation of these functions using several parameters and magnifications. A report such as this can be viewed only as introductory due to the large variety of DNA and amino acid parameters that potentially can be visualized by this method. For example, for DNA the sphere size chosen can be changed to emphasize global features, or triply bonded bases can be color differentiated from singly bonded ones, several tetragrams may be superimposed, etc. The exploration of this large parameter space provides a provocative area for future research. It may be possible to discover interesting properties and periodicities in the DNA sequence by having the program produce many tetragrams by automatically iterating through a larger number of input parameters and mappings. In this way, the program may suggest to the human analyst important features and parameters that would not be considered otherwise. The correlation of resultant features with biological relevance would be the next necessary area of study.

As a final experiment, I created synthetic DNA using a stationary Markov process, and the simulated DNA tetragrams visually resemble the real ones. Readers may wish to experiment by correlating two of the bases, B (B can have two values, 0 or 1), $\{B_i, i = 1,2,3, \ldots N\}$ where the 0/1 sequence is not "completely random" but can be described as a stationary Markov process with the transition matrix P:

$$P = \begin{bmatrix} P_0 & 1 - P_0 \\ 1 - P_1 & P_1 \end{bmatrix} \quad [1]$$

Here, P_0 and P_1 are the probabilities that B_i is equal to 0 or 1, respectively, if B_{i-1} is equal to 0. Similarly, P_1 and $1 - P_1$ are the probabilities that B_i is equal to 1 or 0, respectively, if B_{i-1} is equal to 1. The values of B_i depend on the values at B_{i-1}, and deviations from randomness (P_0, $P_1 \neq 0.5$) produce DNA-like tetragrams. This may mean that DNA can be modeled as a stationary Markov process.

It would be interesting also to find out if most of the known DNA sequences are constrained to walk in certain preferred overall directions, or if the end points of a trace are equally likely to be in any direction in three-space. It is hoped that the tetrahedron application preliminarily presented here will provide a useful tool for future representations of nucleic acid sequences. Recent proposals to sequence the entire 3-billion base sequence of the human genome,[11] advances in understanding viral sequences that induce cancer[14] and in understanding the mechanisms by which oncogenes are activated in tumors,[15] and improved, commercial gene mapping techniques (with accompanying proliferation of sequence data[16]) motivate further assessment of the DNA tetragram.†

Readers wishing a directed reading list for various past papers concerning the visual display of biological sequence information, should see References 18–23. These references are taken from a larger list appearing in reference,[17] which also lists various papers on the musical interpretation and representation of genetic and amino acid sequences, and I look forward to receiving additions to this list from readers.

REFERENCES AND NOTES

1 See, for example, the *J. Mol. Graphics* (Butterworth and Company). Also see: Pickover, C. Spectrographic representations of globular protein breathing motions. *Science* 1984, **223**, 181

2 Cantor, C. and P. Schimmel. *Biophysical Chemistry* W.H. Freeman and Company, San Francisco, 1980

3 Friedland, P. and Kedes, L. Discovering the secrets of DNA. *Commun. ACM* 1985, **28**, 1164–1186

4 Silverman, B.D. and Linsker, R. A measure of DNA periodicity. *J. Theor. Biol.* 1986, **118**, 295–300

5 Hamori, E. and Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.* 1983, **258**(2) 1318–1327

6 Hamori, E. Novel DNA sequence representations. *Nature* 1985, **314**, 585–586

7 Lathe, R, Findlay, R. Novel DNA sequence representations. 1985, **314**, 585–586

8 Hamori, E., Varga, G., and LaGuardia, J. HYLAS: program for generating H curves (abstract three-dimensional representations of long DNA sequences). *Comp. Appl. Biol. Sci.* 1989, **5**(4) 263–269

9 Pickover, C. *Computers, Pattern, Chaos and Beauty* St. Martin's Press; New York, 1980; Pickover, C. *Computers and the Imagination* St. Martin's Press; New York, 1991; Pickover, C. DNA Vectorgrams: representation of cancer gene sequences as movements along a 2-D cellular lattice. *IBM J. Res. Dev.* 1987, **31**: 111–119; Pickover, C. Computer-drawn faces characterizing nucleic acid sequences. *J. Mol. Graphics* 1984, **2**, 107–110

10 Dickerson, R. and Geis, I. *The Structure and Action of Proteins.* Benjamin/Cummings; Massachusetts, 1969

11 Lewin, R. Proposal to sequence the human genome stirs debate. *Science* 1986, **232**, 1598–1599

12 Pickover, C. Frequency representations of DNA sequences: Application to a bladder cancer gene. *J. Molec. Graphics* 1984, **2**, 50

13 For example, graphs of both a smooth, regular curve and a very noisy curve having the same power spectra are illustrated and discussed in Pickover, C. and Khorasani, A. Fractal characterization of speech waveform graphs. *Comp. Graphics* 1986, **10**, 51–61

14 Bishop, J. Oncogenes. *Scien. Amer.* March 1982, 81–92

15 Wong, A., Ruppert, J., Eggleston, J., Hamilton, S., Baylin, S., and Vogelstein, B. Gene amplification of *c-myc* and *N-myc* in small cell carcinoma of the lung. *Science* 1986, **233**, 461–464

16 Schneider, K. Gene mapping is improved. *New York Times.* 26 June 1986

17 Pickover, C. *The Visual Display of Biological Information,* in press.

18 Gates, M. A simple way to look at DNA. *J. Theor. Biol.* 1986, **119**, 319–328

19 Jeffrey, H. Chaos game representation of gene structure. *Nucl. Acids Res.* 1990, **18**(8) 2163–2170

20 Mizraju, E. and Ninio, J. Graphical coding of nucleic acid sequences. *Biochimie* 1985, **67**, 445–448

21 Swanson, R. A vector representation for amino acid sequences. *Bul. Math. Biol.* 1984, **46**(4) 623–639

22 Cowin, J., Jellis, C., and Rickwood, D. A new method of representing DNA sequences which combines ease of visual analysis with machine readability. *Nucl. Acids Res.* 1986, **14**(1) 509–520

23 Melcher, U. A readable and space-efficient DNA sequence representation: application to caulimoviral DNAs. *Comput. Appl. Biosci.* 1985, **4**, 93–96

†To compute the directions of movements in the C-language program, the following commands were repeated for each base in the sequence c, variables x, y, and z are initially set to 0.

s2 = $\sqrt{2}$

```
if (c[i] = = 'G') {x = x+ 0.5; y = y+.5*s2; strcpy(st , "G");}
if (c[i] = = 'A') {x = x+0.5; y = y-.6*s2; strcpy(st , "A");}
if (c[i] = = 'T') {x = x-0.5; z = z+.5*s2; strcpy(st , "T");}
if (c[i] = = 'C') {x = x-0.5; z = z-.5*s2; strcpy(st , "C");}
```

The pictures in the article are oriented, by hand, using a computer mouse to show features of interest