Reading of DNA Sequence Logos: Prediction of Major Groove Binding by Information Theory

Thomas D. Schneider*

version = 1.50 of oxyr.tex 1999 October 12

Methods in Enzymology: RNA Polymerase and Associated Factors, Part B, Volume 274, pages 445-455, 1996
Edited by Sankar Adhya

Supplemental information:
http://www.lecb.ncifcrf.gov/~toms/how.to.read.sequence.logos/index.html

running title: Reading Sequence Logos

*National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Experimental and Computational Biology, National Cancer Institute, P. O. Box B, Building 144, Room 469, Frederick, MD 21702. email address: toms@ncifcrf.gov. http://www.lecb.ncifcrf.gov/~toms/

Information theory was introduced in the late 1940s by Claude Shannon for the study of communications systems[1, 2]. With this mathematical tool, information passing through a telephone or computer data line can be measured and compared to the theoretical limit of the line, called the channel capacity. The information measure is given in units of bits per second, where one bit is the choice between two equally likely possibilities. Surprisingly, one of Shannon's theorems states that as long as the channel capacity is not exceeded, the communications may have as few errors as desired[3]. An example of the practical use of this result is the clear music of compact discs (CDs) which are specially coded to protect against noise. Cleaning instructions for CDs say that they should be wiped in a radial direction. If any scratches are introduced, they will have less effect on the coding, which runs in concentric circles and is capable of correcting up to 4000 data bit errors[4].

Evolutionary conservation indicates the functional importance of biological structures, so a robust and precise measure for conservation is necessary for empirical studies. The measurement of conservation in bits is uniquely suited to this task because, unlike any other measure, bits are additive and consistent from one system to the next. Bits provide a universal scale for measuring biological conservation not only in DNA and RNA but also for proteins and other macromolecules[5, 6].

OxyR is a tetrameric protein that binds to the DNA of several promoters in *Escherichia coli* and activates transcription of genes encoding antioxidant enzymes[7, 8]. The initial investigation of the DNA-binding sites by Tartaglia *et al.* showed that OxyR binds to a large region of DNA, but the consensus sequences obtained were weak and sparse, making the sites difficult to characterize.

In conjunction with the information analysis described here, recent experimental work on this protein[9] clarified the way in which OxyR binds to DNA sequences. This would not have been possible using a consensus sequence.

Materials and Methods

Sequences

When possible, wild-type OxyR DNA binding site sequences[7, 9] were obtained from GenBank 83. The following list includes the site name, accession number, and coordinate of the central (zero) base: oxyR, J04553, 163; katG, M21516, 68; ahpC, D13187, 116; dps, X69337, 202; gorA, U00039, 60491; Mu mom, V01463, 59; *S. typhimurium* orf, (not in GenBank) 29.

Both wild-type and the random sequences that bound to OxyR are given elsewhere[9].

Programs and Data

Information analysis and the sequence logo technique were performed as described previously [5, 10, 11]. A primer on information theory is available on the internet at ftp://ftp.ncifcrf.gov/pub/delila/primer.ps. Programs (written in Pascal[12]) and data are available by anonymous ftp from ftp://ftp.ncifcrf.gov/pub/delila/ or via the World Wide Web site http://www-lmmb.ncifcrf.gov/~toms/.

The programs for constructing sequence logos and performing other tasks were used in approximately this order:

- dbbk.p: convert GenBank to Delila (*DE*oxyribonucleic *LI*brary *LA*nguage) format[13].
- makebk.p: convert raw sequences to Delila format.
- catal.p: create a Delila database.
- delila.p: extract binding site sequence fragments from a Delila database.
- alist.p: display aligned sequence fragments (Fig. 1).
- encode.p: convert aligned sequence fragments to binary vectors[14].
- rseq.p: use binary vectors to perform information analysis of a binding site[5]. At each position l in the binding site, the frequency of each base b is determined. The frequency, $f(b,l)$, is used to compute the information content at each position according to:

  $$R_{sequence}(l) = 2 - \left(e(n(l)) - \sum_{b \in \{A,C,G,T\}} f(b,l) \log_2 f(b,l) \right) \quad \text{(bits per base)} \tag{1}$$

  where $n(l)$ is the number of bases at l and $e(n(l))$ is a correction for small sample size[5]. The total information content is

  $$R_{sequence} = \sum_{l} R_{sequence}(l) \quad \text{(bits per site).} \tag{2}$$

- dalvec.p: convert information table to symbolic vector format.
- makelogo.p: convert symbolic vector to a sequence logo[10] (Fig. 2).
- malign.p: information theory based multiple alignment.
  A paper describing this program is at
  ftp://ftp.ncifcrf.gov/pub/delila/malign.ps
- rsim.p: compute the standard deviation corresponding to the mean $R_{sequence}$[15].
- ttest.p: perform Student's t test[16].
- dnag.p and dops.p: draw DNA base pairs (just the atoms in Fig. 3).

Detailed documentation on each program is given at the beginning of the source code. Binaries for Sun4 Sparc computers are in ftp://ftp.ncifcrf.gov/pub/delila/bin/sun4/. A World Wide Web sequence logo server has been set up by Steven E. Brenner (University of Cambridge School of Biological Sciences and MRC Laboratory of Molecular Biology) at http://www.bio.cam.ac.uk/seqlogo/.

The aligned listing and sequence logos were printed on a Tektronix Phaser 140 inkjet printer.

Sequence Logos

The sequence logos[10, 11] in Fig. 2 summarize the data in a set of aligned sequences such as Fig. 1. The height of a stack of letters is the sequence conservation measured in bits of information according to equation (1). The height of each letter within a stack is proportional to its frequency at that position in the binding site. The letters are sorted, with the most frequent on top. The cosine wave represents the twist of B-form DNA. Wave peaks are all on one face of the DNA and represent the major groove facing the protein. Error bars indicate the variability of a comparable number of random sequences[5].

Results and Discussion

Analysis of OxyR Sequence Logos for DNA Binding Face

Wild-type OxyR binding site sequences were aligned (Fig. 1) and information analysis was used to generate the sequence logo in Fig. 2a. The logo shows a correlation between the strongest sequence conservation, as given by ⇐Fig 1 ⇐Fig 2 the heights of the stacks of letters, and the face of B-form DNA, as given by the cosine wave. The same correlation is seen in other proteins (λ cI/cro, λ O, 434 cI/cro, ArgR, CRP, TrpR, FNR and LexA, see figure 6 of Papp *et al.*[11]). The extent of DNase I footprinting[7] and this correlation suggests that OxyR binds to one face of B-form DNA in 4 successive major grooves.

A second line of evidence that can be read from the sequence logo also supports this model. When a protein is in contact with a major groove, the two base pairs and their two orientations can be distinguished, as recognized by Seeman *et al*[17, 11], so the protein is capable of "choosing" one of the four possibilities: A=T, T=A, C≡G, or G≡C. This can be explained with the help of Fig. 3, which depicts the two base pairs. The possible chemical contacts for T=A in the major groove are (reading from left to right): methyl group, hydrogen acceptor, hydrogen donor and hydrogen acceptor, or T:Me-a-d-a:A for short. This can easily be distinguished from the complementary pattern of A=T, which is A:a-d-a-Me:T. Likewise C:(blank)-d-a-a:G is distinguishable from G:a-a-d-(blank):C. Finally, GC/CG can be distinguished from AT/TA. ⇐Fig 3

This choice of 1 possibility in 4 can be made with 2 bits of information. (This is calculated as $\log_2 4/1 = 2$ bits. For further explanation, see Pierce[2] or the primer on information theory whose location is given in Materials and Methods.) Completely conserved positions in the major groove are described by 2 bits and this is the highest point on the vertical scale of the sequence logos. It is easiest to think of a bit as a knife slice that dissects the bases in Fig. 3. A horizontal slice is the first bit and a vertical slice is the second one. The first bit determines whether the base is above or below the slice and the second bit determines whether it is to the left or the right. Because they are at right angles to one another, the slices provide independent choices and no more than 2 bits are needed to specify a single base. For example, "top, left" selects the T. The average number of bits needed to describe the observed frequency of bases is the information content or sequence conservation. Because it is an average it does not need to be an integer and so the heights of the letter stacks in the sequence logo are real numbers. Sequence conservation in the major groove can range anywhere between 0 and 2 bits depending on the strength of the contacts involved, as seen in figure 6 of Papp *et al.*[11]. Just because it is *possible* for sequence conservation to be as high as 2 bits from the major groove does not mean that the protein *will* evolve to that high a value. The important factors are the total sequence conservation[5] and the coding of the binding site that distinguishes it from other sequences.

In contrast to the major groove, contacts in the minor groove of B-form DNA allow both orientations of each kind of base pair so that rotations about the dyad axis cannot easily be distinguished. This is because from the minor groove C≡G appears nearly identical to G≡C and A=T appears identical to T=A. Fig. 3 shows that C≡G in the minor groove has the chemical moiety pattern a-d-a, which is, to a good first approximation, identical to the pattern of the complementary orientation. The hydrogen donor N2 of G is almost exactly on the dyad axis (the dashed line) and so its position does not change much in the complement. Hydrogen atoms held in a hydrogen bond vibrate vigorously and probably make such a fine positional distinction difficult because they vibrate almost independently of the donor and acceptor[18]. The base pair A=T has a-(blank)-a, which is identical in the other orientation. So A=T can be distinguished from C≡G in the minor groove only by a donor contact to the N2 or by a physical probe which blocks the N2 (in the blank at the black dot). Because both of these are close to the dyad axis only the horizontal knife slice works for the minor groove.

Because only 2 of the 4 possibilities can be distinguished, when a B-form minor groove is probed by a protein no more than 1 bit of information ($\log_2 4/2 = 1$ bit) can be obtained. That is, positions with more than 1 bit of information are likely to represent major grooves facing the protein or, if they do represent minor groove contacts, then the DNA is probably not B-form[11]. In the OxyR sequence logo, positions ±4, ±5 and ±7 are conserved by more than 1 bit of information, so these positions probably represent major grooves facing the protein. This is consistent with the correlation between sequence conservation and the face of the DNA discussed earlier.

After this prediction was made, it was confirmed by hydroxyl radical footprinting and missing base experiments[9].

Prediction of Specific OxyR DNA Contacts by "Reading" the Sequence Logo

Because position 0 in the wild-type sequence logo shows equally likely A and T (Fig. 2a), that position may represent a contact from OxyR which collides with the minor groove N2 of G (ref.[17]), allowing only A=T and T=A, and disallowing C≡G entirely. This prediction was supported by methylation interference data which indicate that methylation at N3 of A blocks binding (Fig. 2a).

The logo also shows that positions +4, +7 and +13 are predominantly either A or G (T or C at the negative coordinates), which suggests a contact by OxyR to the major groove N7 group because only the N7 acceptor is conserved in an A↔G transition (T:Me-a-d-$\boxed{\text{a}}$:A matches C:(blank)-d-a-$\boxed{\text{a}}$:G only in the last moiety, see Fig. 3). This is also confirmed by methylation interference data, but the disproportionate frequencies of these bases suggests that other contacts or effects are also involved.

Position −15 is mostly T or G (A or C at +15), suggesting a weak contact to the major groove A-N6 or C-N4 and/or T-O4 and G-O6 groups. These contacts could be conserved in T↔G transversions because they shift by only ∼1Å. (T:Me-$\boxed{\text{a}}$-$\boxed{\text{d}}$-a:A matches G:a-$\boxed{\text{a}}$-$\boxed{\text{d}}$-(blank):C in the middle two moieties, see Fig. 3). This kind of contact is likely to be at position −6 (and +7) of CRP sites[11], where the crystal structure shows that Arg^{180} donates hydrogen bonds to O6 and N7 of guanine[19]. The preferred binding order $G > T > A \approx C$ is directly visible in the CRP sequence logo[11], and is confirmed by mutations[20, 21, 19]. Apparently when G is replaced by T in a CRP site, the T-O4 contact is used instead of G-O6, but the G-N7 contact is lost. This accounts for the binding order except for the A. Substitution by A would break the G-O6 contact but would maintain the N7 contact. The binding order $T > A$ suggests that the O4 contact is stronger than the N7 contact[20], even though when T→G, the O4/O6 contact shifts by ∼1Å whereas the N7 contact does not move at all.

Other positions in the OxyR binding sites show methylation interference that do not have apparent correlation to the observed sequence conservation. This could be because the sequence conservation was derived from only a few sequences and so is noisy, as indicated by the large error bars. Alternatively, OxyR may pass close to the DNA at some points but not make actual contact unless the DNA is abnormally methylated. Other effects, such as DNA bending or twisting, also might account for these discrepancies.

A more subtle piece of evidence can be found in the overall shape of the sequence logo. Notice how the stack heights at positions −4, −5, and −7 follow along under the cosine wave (Fig. 2a). This effect can be observed in other sequence logos, in particular λ cI/cro, λ O, 434 cI/cro, CRP, FNR and LexA[11]. It can be explained by the geometry of a globular protein approaching the cylindrical DNA. During the process of finding the binding sites the protein moves toward and away from the DNA[22]. Contacts at the center of the DNA cylinder are closest to the protein and so should be the easiest to evolve. Contacts become progressively more difficult to make as the approach is made further off axis (Fig. 4). If one were to rotate a ⇐Fig 4
DNA molecule on its long axis, a point on its surface would cycle between being visible and not visible. This naturally results in a cosine function of

accessibility along a linear DNA molecule. If we define accessibility as a cosine function that runs from 0 to 1 bit in the minor groove and another cosine function that runs from 0 to 2 bits in the major groove, then the sum of these two functions is the cosine wave drawn on the logo. The correlation between sequence conservation and accessibility is consistent with the proposal that positions ± 4, ± 5, and ± 7 are read from the major groove.

To summarize, there are at least four interrelated techniques that can be used to read a sequence logo:

1. correlation of the strongest sequence conservation with major grooves.

2. stack heights above 1.0 bit suggest major grooves.

3. stack composition suggesting major or minor groove contacts.

4. stack heights following the cosine accessibility wave.

The various observations that we have made for positions 0, ± 4, ± 5, ± 7, ± 13 and ± 15 all support a model in which OxyR binds to 4 major grooves in the orientation shown in Fig. 2a.

Analysis of Synthetic OxyR Binding Sites

A "randomization" experiment[23, 11, 24] was performed in which OxyR protein was used to gel shift 30 base pair equi-probable random sequences[9]. Unfortunately this gave a dismal logo (Fig. 2b), possibly because the protein was prevented from binding properly by the flanking constant sequence of the vector. The high conservation at the ends (± 14 and ± 15) comes more from one side of the cloning sites and so may represent an artifact.[1] Still, some correlation with Fig. 2a is visible in the logo, in particular the A preference at ± 4, the T preference at ± 5, and the G preference at ± 7. But other positions are just as conserved and do not reflect the wild-type sequences.

To clarify this situation, the randomization experiment was repeated with 45 base pair equi-probable random sequences which were then aligned by an information theory technique using the malign.p program (Fig. 2c)[9]. Only some of the patterns evident in Fig. 2a were confirmed by this experiment, whereas others became more predominant. The T or C at position ± 2 closely reflects the wild-type sequences there (7 Cs, 5 Ts, 1 G, 1 A). The conservation at ± 4 and ± 5 increased but positions 0 and ± 7 barely increased.

[1] This is not visible in the sequence logo because both the sequences and their complements were used. It is visible when only one orientation is displayed, not shown.

The wild-type cluster at ± 13, ± 15 and perhaps ± 17 did materialize but not strongly. The almost insignificantly weak preferences for A at ± 6, T at ± 9, and T at ± 18 of wild-type appear amplified. Additional conservation not seen previously appeared at ± 8 (?), ± 12, ± 16, ± 19 and ± 22. The reason for these quantitative discrepancies between the wild-type sequence logo and the logo from experimentally selected sites is unknown, but might be accounted for by the small sample sizes.

Another difficulty with this kind of experiment is that it always contains at least one unknown parameter, the stringency of selection. If the concentration of OxyR protein were large, then its non-specific binding should cause more DNA sequences to shift in the gel. This would lead to a sequence logo with a low information content relative to the natural sequences. However, a low concentration of OxyR protein should lead to a much higher measured information content, perhaps higher than is naturally found. This is also a danger in amplification protocols such as SELEX[25].

To counter this, the protein concentration could be adjusted so that the total information content from the randomization experiment matches that of the natural sequences. Presumably the two logos would then look the same. Were they the same for the experiment that was done? The area under the logo ($R_{sequence}$) was 15.4 ± 1.9 bits for the 7 wild-type sequences and their complements (Fig. 2A) but was 17.5 ± 1.2 bits for the 16 sequences and their complements in the 45mer experiment (Fig. 2C). These can be compared by a two-tailed Student's t test[16]. Because both the sequences and their complements were used, the two halves of the sequence logo are not independent and the test must be done with half-sites. This reduces each mean by a factor of 2 and reduces the variance by a factor of 2 so the standard deviations reduce by $\sqrt{2}$. The number of samples in the half-site set remains at 14 and 32 respectively. The t test on the half-sites shows that the natural sites are significantly different in information content from the experimental set ($t = 2.7$ with 44 degrees of freedom, $p < 0.02$). Unfortunately this criterion for matching the wild-type binding sites was not met.

Even so, it is clear that the sequence conservation is not uniformly proportional across the two logos. One possibility that might account for the observed conservation at the edges of the experimentally derived sites is that OxyR is still affected by the constant flanking sequences of the cloning vector and binds to one or the other side in some cases. The discrepancy may also reflect different conditions between *in vivo* evolutionary factors and the *in vitro* gel shift experiment. For example, no spermidine was used in the gel shift experiment[9], yet it is well known that spermidine is important for precise recognition by other DNA binding proteins[26, 27]. Comparison of the wild-type sequence logo to a series of random gel shift sequence logos might be used to determine precisely what the *in vivo* binding conditions are.

Usefulness of Sequence Logos Versus Consensus Sequences

The case of OxyR demonstrates the usefulness of sequence logos as a replacement for consensus sequences. The pattern bound by the protein is difficult to detect by eye (Fig. 1), and no agreement could be found for a consensus sequence. In contrast, the sequence logos are created automatically and without any ambiguity. They show clear and easily interpretable patterns. Because information theory is quantitative, statistical tests can be applied to collections of binding site sequences.

A consensus replaces the natural frequencies of bases with arbitrarily chosen ones [28]. For example, in our data set[11], CRP position -6 has 2 As, 2 Cs, 44 Gs and 10 Ts. Taking the consensus alters this to 100% G and 0% of the other bases. When the consensus sequence G was chosen by this method, a subtle pattern of sequence conservation was lost[29, 30, 31, 32, 20, 33, 21, 19]. That the T contact occurs naturally apparently went unrecognized until the present work, although the mechanism of the contact was already understood from mutations[20].

The art of predicting specific base contacts is well known[17] but the pervasive use of consensus sequences in the modern molecular biology literature has prevented full use of the available sequence data. In the case of CRP discussed earlier, the subtle G to T switch, which probably destroys one hydrogen bond while keeping the other, was missed because only the G was retained in the consensus model. In contrast, because it visually displays the relevant information in a compact, quantitative form, the sequence logo allows direct interpretation of the data and leads to specific predictions that can guide experimentation. Further, when anomalies appear, the logo displays them so blatantly that new phenomena are revealed[11]. Even correlations between positions in a binding site[15] could be presented in a three-dimensional sequence logo, but software to generate this display has not been written yet.

As seen by the sequence logo, OxyR does not have a special kind of binding site as has been suggested[7]. OxyR merely happens to have a long binding site with low overall information content, so it tends to have low sequence conservation per position. As a consequence of the Second Law of

Thermodynamics, DNA-protein contacts tend to spread out over the available surface on an evolutionary time scale. With at least 4 major grooves and three minor grooves to make contacts in, OxyR can "afford" to have many small contacts. Paradoxically, a mathematically rigorous theorem shows that having many small contacts like those used by OxyR can *improve* sequence discrimination[34]. Furthermore, many binding sites are like OxyR in that they have variations in their information content at different positions. This is immediately apparent upon inspection of splice junction sequence logos[15] and the "gallery" of DNA recognition sequence logos[11].

The arbitrary and artificial distinctions between strong and weak binding sites, between the "core" and the periphery of a site and between the inside and outside of binding site "boxes" that have been fostered by the use of consensus sequences are eliminated when one adopts the concept that sequence conservation is a real number that can be measured precisely in bits of information.

Summary

DNA sequences to which the OxyR protein binds under oxidizing conditions were analyzed by the sequence logo method, a quantitative graphic technique based on information theory. A sequence logo shows both the sequence conservation and the frequencies of bases at each position in a site. Unlike the consensus sequence, the sequence logo analysis revealed that OxyR should bind to four major grooves of DNA. This was later confirmed by experiments. Detailed interpretation of the sequence logo also allowed the prediction of likely major and minor groove OxyR-DNA base contacts, consistent with available experimental results. Because the sequence logo shows the original base frequencies in a clear, easily interpreted graphic that does not distort the data, highly refined analysis of binding site contacts becomes easy. Not only can these methods be applied to any DNA sequence binding site, they can also be applied to sites on RNA and proteins.

Acknowledgments

I thank Paul N. Hengen, Denise Rubens, Paul A. Smith, R. Michael Stephens, and R. E. Wolf for useful comments on the manuscript.

```
                 -----------------                   +++++++++++++++++
                 22222221111111111--------- +++++++++11111111112222222
                 65432109876543210987654321012345678901234567890123456
                 .....................................................
J04553    163  1 tagggataatcgttcattgctattctacctatcgccatgaactatcgtggcga
J04553    163  2 tcgccacgatagttcatggcgataggtagaatagcaatgaacgattatcccta
M21516     68  3 acaatatgtaagatctcaactatcgcatccgtggattaattcaattataactt
M21516     68  4 aagttataattgaattaatccacggatgcgatagttgagatcttacatattgt
D13187    116  5 gaaggttgtaaggtaaaacttatcgatttgataatggaaacgcattaccggaa
D13187    116  6 ttccggtaatgcgtttccattatcaaatcgataagttttaccttacaaccttc
X69337    202  7 tttttcacgcttgttaccactattagtgtgataggaacagccagaatagcgga
X69337    202  8 tccgctattctggctgttcctatcacactaatagtggtaacaagcgtgaaaaa
U00039  60491  9 aagctggatcgtgccggagtaattgcagccattgctggcacctattacgtctc
U00039  60491 10 gagacgtaataggtgccagcaatggctgcaattactccggcacgatccagctt
V01463     68 11 tagaaaacgacgatcgaatcaattaaatcgatcggtaatacagatcgattatg
V01463     68 12 cataatcgatctgtattaccgatcgatttaattgattcgatcgtcgttttcta
sal.orf    29 13 ctggcacgccagctcttacctatgtctgtgataggcatcatcattaatactct
sal.orf    29 14 agagtattaatgatgatgcctatcacagacataggtaagagctggcgtgccag
```

Figure 1: Aligned list of OxyR binding sequences.

7 OxyR wild-type binding sites (odd numbers) and their complementary sequences (even numbers) were listed by the alist program. The numbers in the bar on the top are read vertically and give the position in the binding site, running from -26 to $+26$. The GenBank accession number, the coordinate of the zero base, and the number of each sequence are given on the left-hand side of the figure.

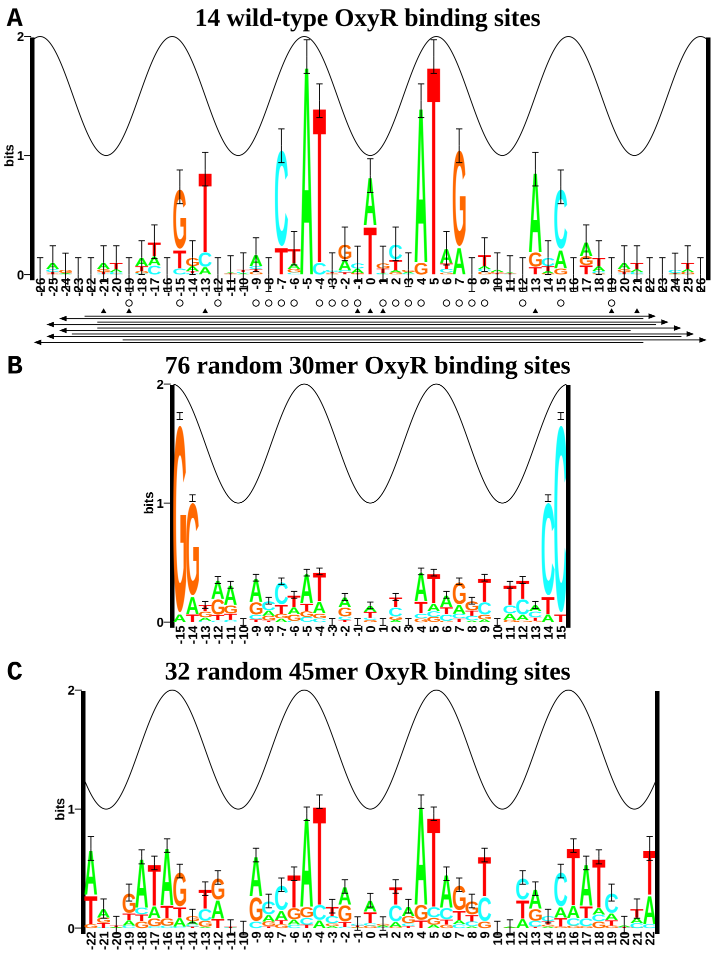

Figure 2: Sequence logos for OxyR binding sequences.

(A) 7 OxyR wild-type binding sites and their complementary sequences. The total sequence conservation, obtained by adding together the stack heights to determine the "area" under the logo (equation (2)), is 15.4 ± 1.9 bits per site for the range -22 to $+22$ with error calculated by program rsim according to [15]. Methylation of guanines at N7 which interfere with OxyR binding are indicated by open circles ($\circ$) and methylation of adenines at N3 which interfere with OxyR binding are indicated by filled triangles (▲) [35, 7]. DNase I protected regions for the sites in Tartaglia et al.[7] are shown by arrows drawn $5' \rightarrow 3'$. (B) 38 randomly synthesized sequences selected by OxyR protein, 30 bases wide and their complementary sequences. $R_{sequence} = 11.6 \pm 0.6$ bits per site. in the range -15 to $+15$. (C) 16 randomly synthesized sequences selected by OxyR protein, 45 bases wide and their complements. $R_{sequence} = 17.5 \pm 1.2$ bits per site in the range -22 to $+22$.

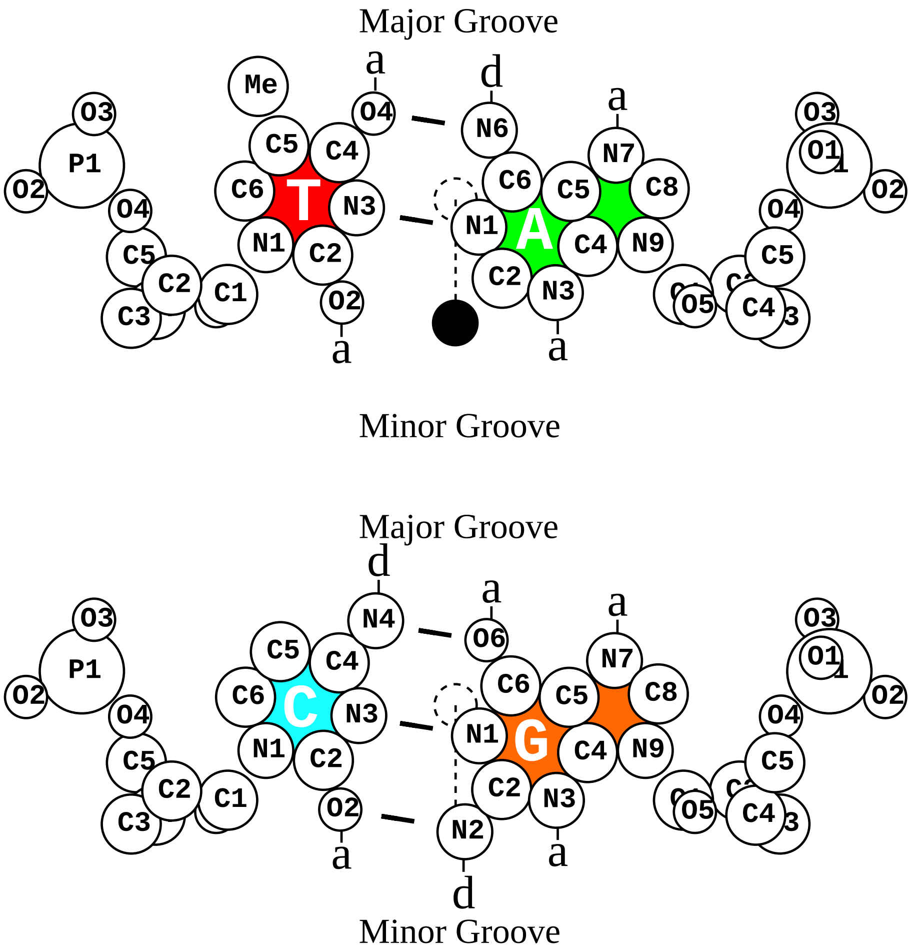

Figure 3: DNA base pairs

DNA base pairs drawn by the dnag program using coordinates for B-DNA [36] with atomic radii [37]. Short line segments indicate hydrogen bonds between the bases. a: acceptor of hydrogen bond; d: donor of hydrogen bond. The scale is shown by a 1Å diameter dashed circle placed on the helical axis. The two-fold dyad axis is indicated by a dashed line. Rotation by 180° on this axis brings the backbone sugars and phosphates into register again. The next base pair is 3.38Å above the page and is rotated 36° counterclockwise.

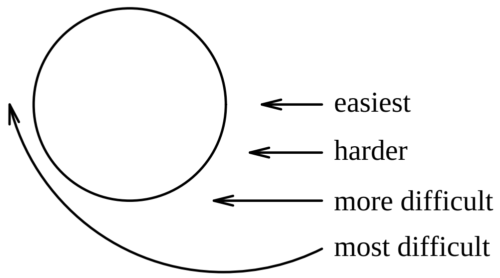

Figure 4: DNA accessibility.

References

[1] N. J. A. Sloane and A. D. Wyner, "Claude Elwood Shannon: Collected Papers." IEEE Press, Piscataway, NJ, 1993.

[2] J. R. Pierce, "An Introduction to Information Theory: Symbols, Signals and Noise." Dover Publications, Inc., New York, 1980.

[3] C. E. Shannon, Proc. IRE 37, 10 (1949).

[4] K. A. S. Immink, "Coding Techniques for Digital Recorders." Prentice-Hall, Inc., N. Y., 1991.

[5] T. D. Schneider, G. D. Stormo, L. Gold and A. Ehrenfeucht, J. Mol. Biol. 188, 415 (1986).

[6] G. D. Stormo, Meth. Enzym. 208, 458 (1991).

[7] L. A. Tartaglia, C. J. Gimeno, G. Storz and B. N. Ames, J. Biol. Chem. 267, 2038 (1992).

[8] G. Storz and L. A. Tartaglia, J. Nutr. 122, 627 (1992).

[9] M. B. Toledano, I. Kullik, F. Trinh, P. T. Baird, T. D. Schneider and G. Storz, Cell 78, 897 (1994).

[10] T. D. Schneider and R. M. Stephens, Nucl. Acids Res. 18, 6097 (1990).

[11] P. P. Papp, D. K. Chattoraj and T. D. Schneider, J. Mol. Biol. 233, 219 (1993).

[12] K. Jensen and N. Wirth, "Pascal User Manual and Report." Springer-Verlag, New York, 1975.

[13] T. D. Schneider, G. D. Stormo, J. S. Haemer and L. Gold, Nucl. Acids Res. 10, 3013 (1982).

[14] T. D. Schneider, G. D. Stormo, M. A. Yarus and L. Gold, Nucl. Acids Res. 12, 129 (1984).

[15] R. M. Stephens and T. D. Schneider, J. Mol. Biol. 228, 1124 (1992).

[16] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, "Numerical Recipes in Pascal. The Art of Scientific Computing." Cambridge University Press, Cambridge, 1989.

[17] N. C. Seeman, J. M. Rosenberg and A. Rich, Proc. Natl. Acad. Sci. USA 73, 804 (1976).

[18] G. J. Kearley, F. Fillaux, M.-H. Baron, S. Bennington and J. Tomkinson, Science 264, 1285 (1994).

[19] S. C. Schultz, G. C. Shields and T. A. Steitz, Science 253, 1001 (1991).

[20] C. Jansen, A. M. Gronenborn and G. M. Clore, Biochem. J. 246, 227 (1987).

[21] X. Zhang and R. H. Ebright, Proc. Natl. Acad. Sci. USA 87, 4717 (1990).

[22] I. T. Weber and T. A. Steitz, Nucl. Acids Res. 12, 8475 (1984).

[23] T. D. Schneider and G. D. Stormo, Nucl. Acids Res. 17, 659 (1989).

[24] D. Barrick, K. Villanueba, J. Childs, R. Kalil, T. D. Schneider, C. E. Lawrence, L. Gold and G. D. Stormo, Nucl. Acids Res. 22, 1287 (1994).

[25] C. Tuerk and L. Gold, Science 249, 505 (1990).

[26] A. Pingoud, Eur. J. Biochem 147, 105 (1985).

[27] C.-Y. Wan and T. A. Wilkins, PCR Methods and Applications 3, 208 (1993).

[28] G. D. Stormo, Meth. Enzym. 183, 211 (1990).

[29] B. de Crombrugghe, S. Busby and H. Buc, Science 224, 831 (1984).

[30] R. H. Ebright, P. Cossart, B. Gicquel-Sanzey and J. Beckwith, Nature 311, 232 (1984).

[31] I. T. Weber and T. A. Steitz, Proc. Natl. Acad. Sci. USA 81, 3973 (1984).

[32] R. H. Ebright, P. Cossart, B. Gicquel-Sanzey and J. Beckwith, Proc. Natl. Acad. Sci. USA 81, 7274 (1984).

[33] M. E. Gent, A. M. Gronenborn, R. W. Davies and G. M. Clore, Biochem J. 242, 645 (1987).

[34] T. D. Schneider, Nanotechnology 5, 1 (1994).

[35] U. Siebenlist and W. Gilbert, Proc. Natl. Acad. Sci. USA 77, 122 (1980).

[36] S. Arnott and D. W. L. Hukins, Biochem. Biophys. Res. Commun. 47, 1504 (1972).

[37] M. Karplus and R. N. Porter, "Atoms & Molecules." Benjamin/Cummings Publishing Co., Menlo Park, CA, 1970.