4408–4415 Nucleic Acids Research, 1997, Vol. 25, No. 21

Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences

Thomas D. Schneider

National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Mathematical Biology, PO Box B, Frederick, MD 21702-1201, USA

Received April 8, 1997; Revised and Accepted July 7, 1997

ABSTRACT

A graphical method is presented for displaying how binding proteins and other macromolecules interact with individual bases of nucleotide sequences. Characters representing the sequence are either oriented normally and placed above a line indicating favorable contact, or upside-down and placed below the line indicating unfavorable contact. The positive or negative height of each letter shows the contribution of that base to the average sequence conservation of the binding site, as represented by a sequence logo. These sequence 'walkers' can be stepped along raw sequence data to visually search for binding sites. Many walkers, for the same or different proteins, can be simultaneously placed next to a sequence to create a quantitative map of a complex genetic region. One can alter the sequence to quantitatively engineer binding sites. Database anomalies can be visualized by placing a walker at the recorded positions of a binding molecule and by comparing this to locations found by scanning the nearby sequences. The sequence can also be altered to predict whether a change is a polymorphism or a mutation for the recognizer being modeled.

INTRODUCTION

Sequence logos are a graphical method that use letters to quantitatively depict the average sequence conservation and base frequencies in a set of aligned sequences (1). Logos have been used to help understand DNA/protein interactions (2–4), RNA/protein interactions (5), protein structure (6–9) and English word structure (10). However, as useful as they are for characterizing an entire set of sequences, logos only convey a vague idea of how a protein would interact with a specific DNA sequence.

The walker method described here solves this problem by combining letter graphics with a unique weight matrix (11–14) defined by information theory (15–17). Although a walker looks like a logo, it is not the same. In a logo a stack of letters depicts the relative frequencies of bases or amino acids at each position in an aligned set of sequences. The height of the stack is the sequence conservation, measured in bits of information. In contrast, walkers apply to a single sequence, so in a walker only a single letter is drawn for each position on the sequence (Fig. 1). The height of the letter is in bits, and represents that base's contribution to the sequence conservation of the entire set of sequences. A walker represents the individuals that make up the logo, with the logo representing the average sequence conservation (18). As a walker is moved along a DNA, one can immediately see how the matrix 'responds' to particular sequences. This new method is complementary to, and a natural extension of, sequence logos. Walkers allow one to visualize complex genetic regions, to interpret their structure, to understand the effects of sequence changes and to simultaneously engineer overlapping binding sites.

MATERIALS AND METHODS

The first step to using walkers is to create a model of the binding sites. To do this the Delila programs **dbbk**, **catal**, **delila** and **alist** were used to extract and align DNA sequences from GenBank flat files (4,19,20). Rapid multiple alignment based on maximizing the information content of the binding sites was performed with the **malign** program to check the alignment (21). An information curve was made using **encode** (20) and **rseq** (22) and the average shown as a sequence logo using **dalvec** and **makelogo** (1).

Once a sequence logo was established, the **ri** program was used with the same sequences to create an information theory based model of the binding sites which is called an 'individual information weight matrix' (18). After making the matrix, **ri** also determines the information content of every binding site, thereby establishing a distribution of information values that can be graphed with the **genhis** and **genpic** programs. Important properties of this matrix and distribution are discussed in the following section.

Portions of sequences to be analyzed were also extracted by **delila**. For Figures 2 and 3, sequences were extracted and renumbered by using a new feature in **delila** instructions (19), 'default coordinate 0' that redefines the coordinates of the sequence. For example, the **delila** instructions for Figure 2 were: [title 'Delila instructions for tgt/sec'; default coordinate 0; organism E.coli; chromosome E.coli; piece M37702; name 'tgt/sec promoter'; get from 1897 –83 to same –42;]. (The 'same' in this get statement refers back to the 'from' coordinate 1897.)

The **scan** program applies an individual information weight matrix to every possible position in a set of sequences. The evaluations can be plotted against position using **dnaplot** and **xyplo** (23). These graphs can show vast regions, but they lack details.

In contrast, a single interactive walker shows the sequence in great detail (Figs 1, 4 and 6). The user may interactively move the walker around, change parameters of the display and modify the sequence. Only one walker can be seen at a time by this method, which was

Tel: +1 301 846 5581; Fax: +1 301 846 5598; Email: toms@ncifcrf.gov

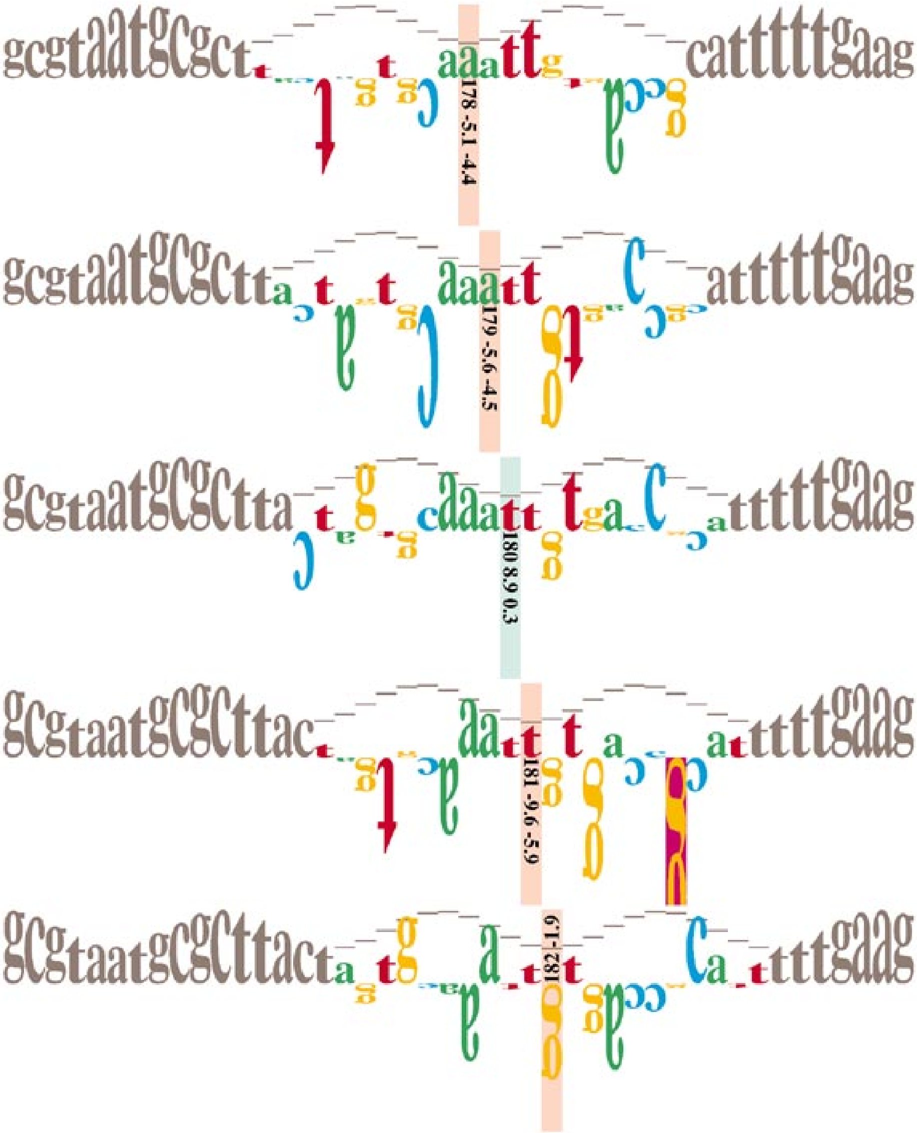


Figure 1. Walkers displayed around the *S.typhimurium hin* distal Fis site at GenBank accession V01370 positions 156–199 using **makewalker**. Five frames of a 'movie', in five successive rows, show the walker progressing along the sequence. The walker is the colored letters. The height of each letter is the information content in bits. The green or red rectangle provides the scale, with –4 bits on the bottom, 0 bits at the middle and +2 bits at the top.

implemented using the **makewalker** and **GhostScript** (http://www.cs.wisc.edu/~ghost/index.html) programs. The **makewalker** program creates a PostScript program (the 'walk') which is then run by the interactive PostScript interpreter **GhostScript**. Commands to change the display are implemented as PostScript procedures.

The **lister** program (19) provides an intermediate scale between **dnaplot**/**xyplo** graphs and the detailed but interactive **makewalker** display, by showing many walkers simultaneously (Figs 2, 3 and 5). To use **lister** one collects 'features' from various programs. **Scan** supplies a weight matrix and the locations for one or more walker or ASCII representations of sites, **palinf** supplies the locations of palindromes, **live** supplies the orientation of DNA as a colored strip, **exon** supplies coding regions based on GenBank features, and **search** supplies the locations of restriction sites and other simple patterns. All of these features, along with user defined marks, are combined into a single 'map' by **lister**. This display is not interactive, but it can be automatically regenerated by using the **atchange** program (http://www-lmmb.ncifcrf.gov/~toms/atchange.html) to detect when sequences or parameters are altered.

Programs for individual information analysis are written in Pascal (24) and can be automatically translated to C by **p2c**, http://www.synaptics.com/people/daveg/) (Table 1). All graphics are in PostScript (25,26). Web-linked manual pages describe each program's function input and output files, and present limitations.

The programs described here are available as Sun Sparc binaries. A confidential disclosure agreement with the National Institutes of Health (NIH) is required for academic and government researchers, and a licensing agreement can be negotiated with the Office of Technology Transfer at NIH for commercial applications (see http://www-lmmb.ncifcrf.gov/~toms/walker/contacts.html). Further information and examples are at http://www-lmmb.ncifcrf.gov/~toms/. Other **delila** programs

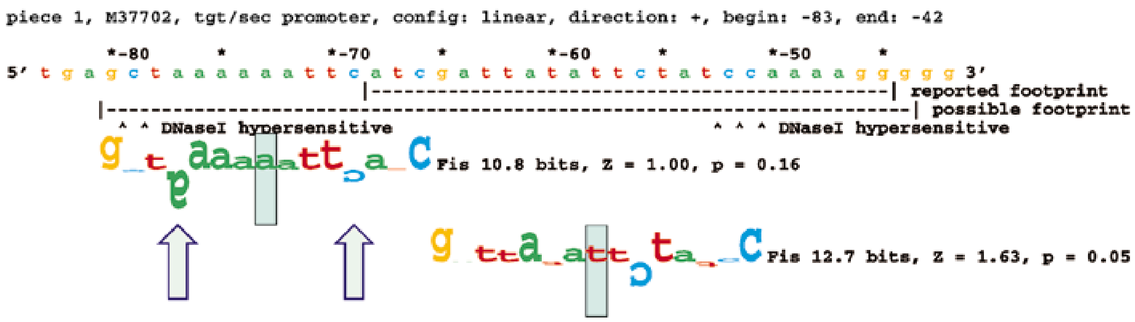

piece 1, M37702, tgt/sec promoter, config: linear, direction: +, begin: -83, end: -42
*-80 * *-70 * *-60 * *-50 *
5' t g a g c t a a a a a a t t c a t c g a t t a t a t t c t a t c c a a a a g g g g g 3'
|---| reported footprint
|--| possible footprint
^ ^ DNaseI hypersensitive
^ ^ ^ DNaseI hypersensitive
Fis 10.8 bits, Z = 1.00, p = 0.16
Fis 12.7 bits, Z = 1.63, p = 0.05

Figure 2. Prediction of an additional binding site using **scan** and **lister**. The *tgt/sec* promoter region was scanned for Fis sites in the region of coordinate –58 where a Fis site had been reported (31). The top line of the figure identifies the DNA piece number, the GenBank entry, the name of the DNA piece, and the topology and coordinates of the DNA sequence (19). Coordinates are numbered every 10 bases and marked by an asterisk (*) every 5 bases. The start of transcription at 1897 of GenBank entry M37702 was renumbered 0 by **delila** (not shown). Horizontal dashes below the sequence represent regions protected by Fis from DNase I. According to the original report, the 5′ limit of the DNase I footprint was unclear so both the reported footprint region and the footprint as far as it might extend are marked, along with DNase I hypersensitive sites (&) that are affected by Fis. The arrows indicate bases that do not fit the consensus (G/T)N_2**Y**RN_2(A/T)N_2Y**R**N_2(C/A) used in previous work (31,44).

(including **dbmutate**, **xyplo**, **live** and **mergemarks** are freely available from the web site.

RESULTS AND DISCUSSION

Mathematical basis of walkers: individual information

The mathematics used to construct walkers is a natural extension of Shannon's information theory (15,17,27). Information is a useful measure for sequence conservation because it is additive when the positions are independent; no other measure (e.g., counting mismatches to a consensus) has this property (16). To get the total sequence conservation, represented by the area of a sequence logo and called R_{sequence}, the information content is summed across all positions in a site (22). Information measured in bits is the average number of binary choices needed to specify the observed degree of sequence conservation.

The idea behind walkers is that a sequence conservation can be assigned to *individual* sequences, so that the average of these individuals is also R_{sequence}. This is done in three steps. First, a matrix is created from the frequency $f(b,l)$ of each base b at position l in the aligned sequences, according to $R_{iw}(b,l) = 2 + \log_2 f(b,l) - e[n(l)]$, where $e[n(l)]$ is a small sample correction (22). Second, this matrix is used to evaluate the individual information content of each site. That is, after aligning the matrix with a sequence, each base in the sequence selects one of the four weights in $R_{iw}(b,l)$, and all weights for the site are summed for all positions l to produce the 'individual information', R_i. Third, if the individual information values of all the aligned sequences are averaged the result is R_{sequence}. There is only one way to define the matrix to obtain this result. The details of this method and other properties of the weight matrix are described elsewhere (18), but some of the important properties of the method that affect interpretation of walkers are discussed briefly below.

(i) Only functional sites are needed to create an individual information matrix, so the intensive training procedures required by neural networks (12) are avoided. Gathering large sets of negative data (showing where a protein does not bind) is also unnecessary. As a result, clean models can be created by restricting the input to biochemically proven binding sites. A model for a binding site can be created with as few as six examples, but this gives unreliable results. A set of 20 or more examples generally gives a reasonable sequence logo and weight matrix.

Table 1. Delila programs (1,4,19,20) used for individual information analysis

Program	Version	Function
dbmutate	1.30	mutate GenBank database entries
dnaplot	3.40	graph individual infromation across large DNA sequences
exon	1.86	convert exons and CDSs to features for a **lister** map
lister	8.63	list sequences with translation, features, walkers and hand-defined marks
live	1.14	add a color bar to a **lister** map to show DNA periodicity
makewalker	3.47	walk an information weight matrix across a sequence
mergemarks	1.04	merge **live** marks with hand-defined marks
ri	2.37	compute individual information weight matrix and individual information for every site
scan	2.88	scan sequences with an individual information matrix to find sites
xyplo	8.63	general x,y data plotter

See http://www-lmmb.ncifcrf.gov/~toms/delila.html for information about other Delila programs.

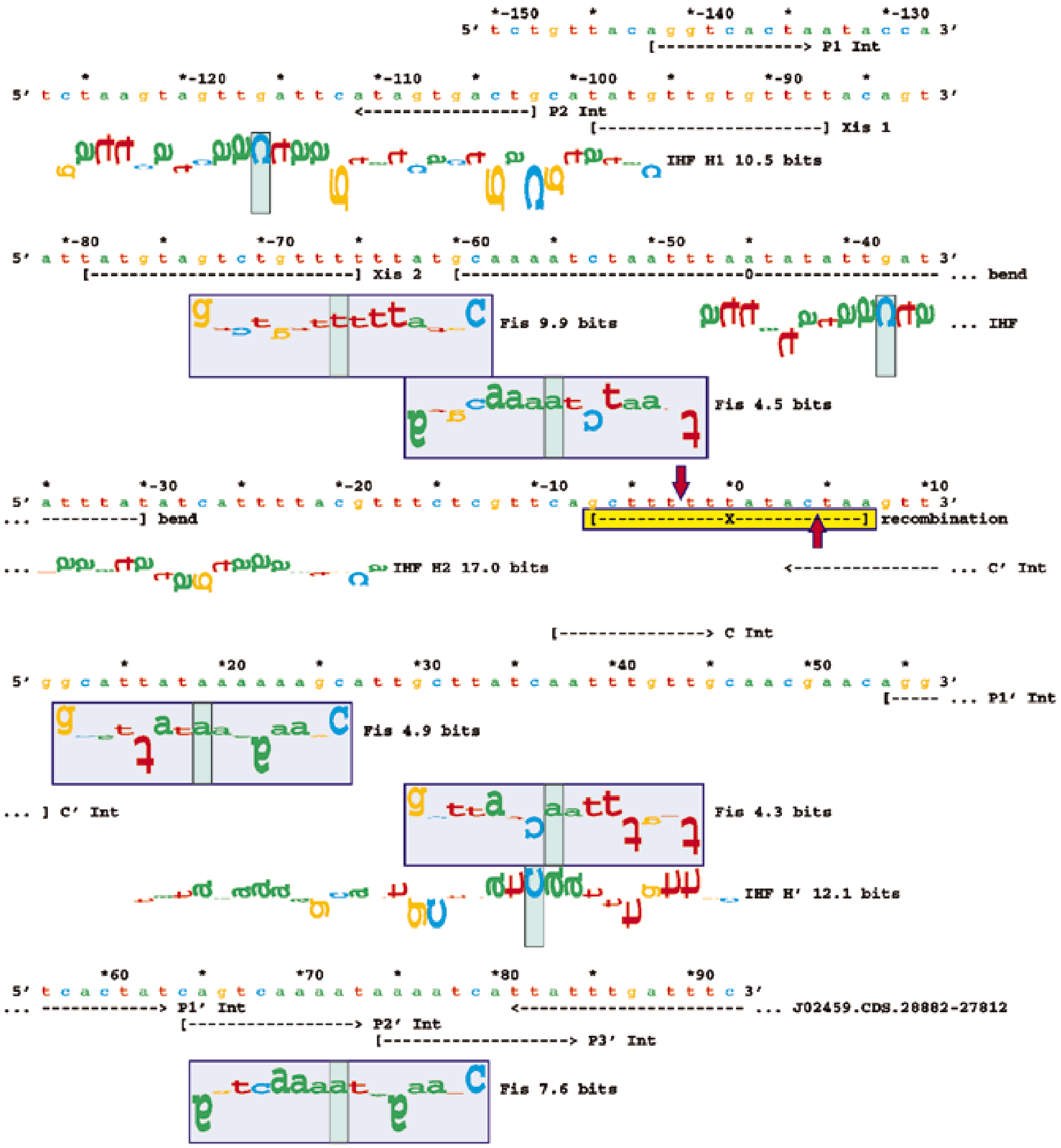

*-150 * *-140 * *-130
5' t c t g t t a c a g g t c a c t a a t a c c a 3'
[---------------> P1 Int
* *-120 * *-110 * *-100 * *-90 *
5' t c t a a g t a g t t g a t t c a t a g t g a c t g c a t a t g t t g t g t t t t a c a g t 3'
<-----------------] P2 Int
[-----------------------] Xis 1
IHF H1 10.5 bits
*-80 * *-70 * *-60 * *-50 * *-40
5' a t t a t g t a g t c t g t t t t t t a t g c a a a a t c t a a t t t a a t a t a t t g a t 3'
[----------------------------] Xis 2
[-----------------------------0-------------------- ... bend
Fis 9.9 bits
... IHF
Fis 4.5 bits
* *-30 * *-20 * *-10 * *0 * *10
5' a t t t a t a t c a t t t t a c g t t t c t c g t t c a g c t t t t t t a t a c t a a g t t 3'
... ----------] bend
[-------------X-------------] recombination
... IHF H2 17.0 bits
<--------------- ... C' Int
[---------------> C Int
* *20 * *30 * *40 * *50 *
5' g g c a t t a t a a a a a a g c a t t g c t t a t c a a t t t g t t g c a a c g a a c a g g 3'
[----- ... P1' Int
Fis 4.9 bits
...] C' Int
Fis 4.3 bits
IHF H' 12.1 bits
*60 * *70 * *80 * *90
5' t c a c t a t c a g t c a a a a t a a a a t c a t t a t t t g a t t t c 3'
... ------------> P1' Int
[-----------------> P2' Int
[-------------------> P3' Int
<------------------------ ... J02459.CDS.28882-27812
Fis 7.6 bits

Figure 3. A complex **lister** map for the λ *attP* region. Coordinate zero was renumbered by **delila** from GenBank J02459, coordinate 27731 (35). The sequence, coordinates and most sites are the same as in figure 2 of Landy's review (45). The common core region ('recombination') between the bacterial and phage *att* sites (36) is boxed in yellow. Red arrows mark the staggered sites of strand exchange on top (–2.5) and bottom (+4.5) strands. The IHF model uses sequences from ref. 34. For IHF sites, the direction one would read the letters 'downwards' indicates the orientation of the entire asymmetric site. There are not enough examples to make reliable models, so the Int and Xis sites are marked according to Landy (45). To keep the figure small, only Fis sites >4 bits (blue boxes) (23), and IHF >10 bits are shown. A DNA curvature is at –45 ± 15 (46). The CDS ending at 81 and named with λ coordinates is the 3′ end of the *int* gene.

(ii) When the weight matrix is applied to the same sequences it was derived from, a set of numbers is generated. The distribution of these numbers is approximately Gaussian. The most frequent bases in a binding site select the highest weight matrix values, so a consensus sequence attains the highest possible individual information content. The correspondingly large standard deviation indicates that the consensus sequence has a low probability of being a natural binding site. For example, a human acceptor splice junction only has a 3.7×10^{-3} probability of being the strict consensus, in which only the most frequent base is used (18). This means that walkers will rarely show a 'consensus' sequence at natural binding sequences.

(iii) Functional sites can be distinguished from non-functional sites by the individual information method. Three observations suggest that functional sites have positive values (18). First, a thermodynamic argument shows that individual information values above zero should be binding sites (within the accuracy of the matrix). Second, the majority of ribosome binding sites and splice junctions have sites with positive values. Third, negative values have frequently identified errors in databases. Thus walkers with information values higher than zero bits are expected to be at functional binding sites.

(iv) That there can be a sharp cutoff at zero is implied by the channel capacity theorem, which says that molecular binding can

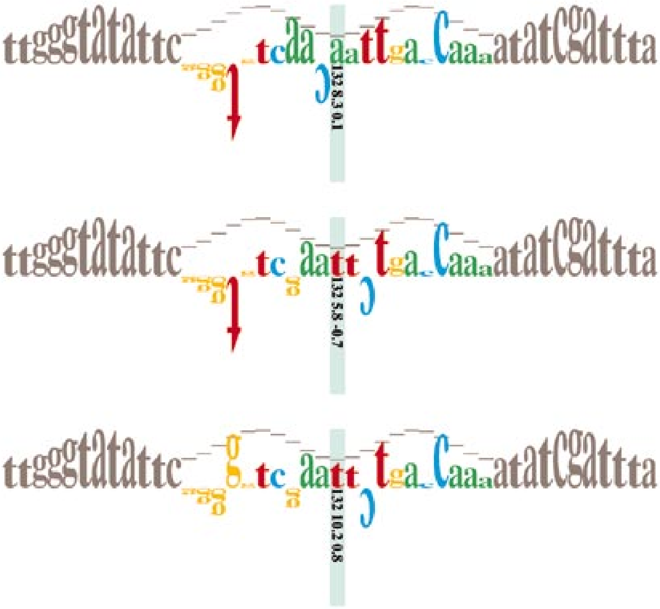


Figure 4. Engineering a binding site by **makewalker**. A *hin* proximal site (top walker) was engineered to contain an *Eco*RI restriction site from position –3 to +2 (middle walker) and an *Eco*RI restriction site with a base change at position –7 (bottom walker).

be precise, independent of the binding energy (28,29). The theorem shows that even small but positive individual information values can indicate functional sites, while small negative values could be distinctly non-functional. When a walker has low but positive conservation, the site is still likely to be functional.

(v) From an observed sequence conservation, it is not possible to reliably infer the binding energy, since sequence conservation and binding energy are not strictly proportional. The second law of thermodynamics defines the maximum information that can be gained for a given energy dissipation (30). Using the second law, it is presently not possible to predict the binding energy more than to say that the specific binding energy must be at least $k_bT\ln2$ multiplied by the information (k_b, Boltzmann's constant; T, absolute temperature). Two binding sites may have similar information content but different binding energies, and sites that differ in information may even have reversed energetics. Still, since energy dissipation tends to be minimized by loss of specific contacts, larger individual information values should generally correspond to higher binding energy. As a consequence, walkers (and any other purely sequence based analysis method) may not directly show values proportional to binding energy, although we expect a walker with higher information to be bound more strongly.

Walkers

A sequence walker graphically shows how a particular sequence is evaluated by an individual information weight matrix. Figure 1 depicts five horizontal rows of characters that represent a DNA sequence, and the sequence is the same in each row. Each row represents the placement of the individual information weight matrix for the Fis protein (23) at a particular position on *Salmonella typhimurium hin* DNA. Matrix weights are represented in colored letters and a set of contiguous ones is called a 'walker'. As one proceeds down the figure, the walker is stepped one position to the right on the DNA sequence so that the figure shows the frames of a 'movie'. Normally this is displayed on a computer screen and only one row is needed since the user controls the display in real-time. A movie of a walker is shown at http://www-lmmb.ncifcrf.gov/~toms/walker/movie/index.html.

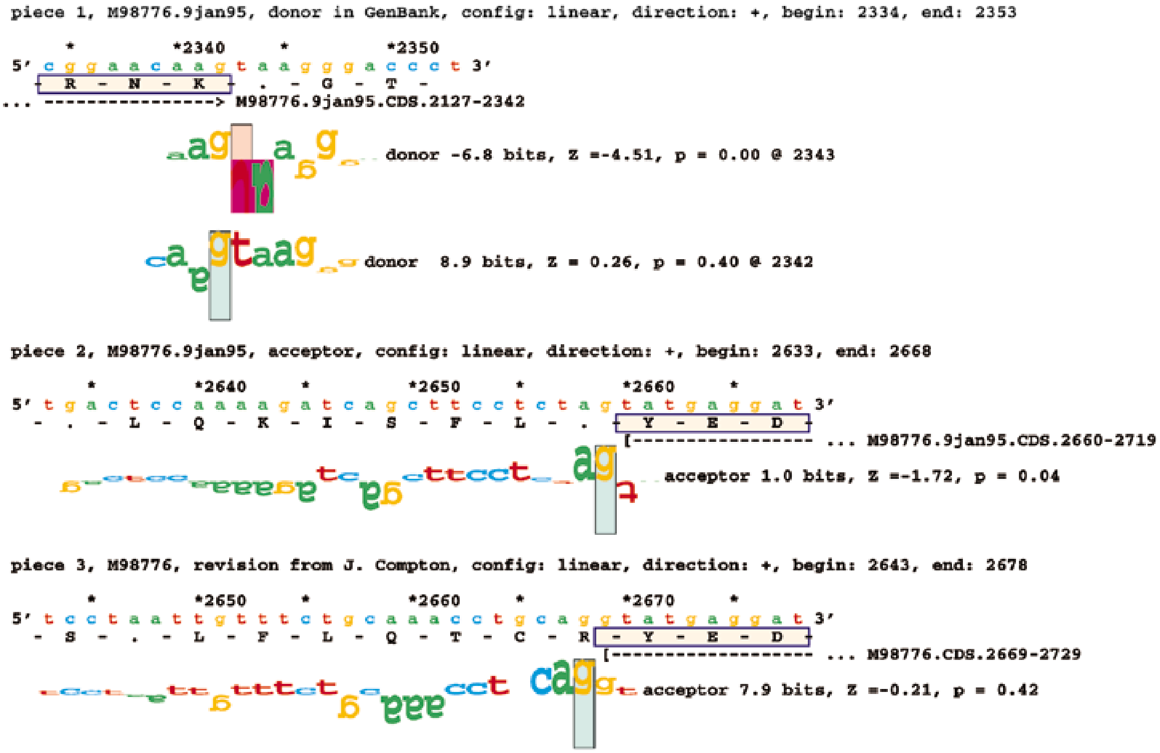
The heights of grey letters to the left and right of the walker indicate the orientation of a B-form DNA helix, with the high points of the sine wave representing the major groove facing the protein (2,4). Horizontal grey bars are used within the region of the walker. As the 'movie' proceeds, the sine wave shifts to represent the rotation of the DNA relative to the protein. A colored vertical box represents the 0 coordinate of the information weight matrix. This defines the position of the Fis protein on the DNA. The box is also a scale, with its lowest edge at –4 bits (an arbitrary lower bound set by the user since there is no lowest possible weight value) and its upper edge at +2 bits (the natural upper bound of weight values for nucleotides).

The walker itself is shown by colored letters, with heights determined by the individual information weight matrix. Letters extending upwards represent favorable DNA contacts (the individual information weight matrix value is positive), while upside-down ones extending downward represent unfavorable contacts (the individual information weight matrix value is negative). If a contact is more unfavorable than –4 bits, the letter is surrounded by a purple box (an example is shown in the 4th row of Fig. 1). If a contact has never been observed at a position in the weight matrix, it is given a black box (no examples are shown in this figure).

Three numbers, or as much of them as is possible, are reported in the vertical box above or below the zero line opposite to the base. The first is the position of the box on the sequence. The second is the sequence conservation of the entire binding site, given in bits. This is obtained by adding together the heights of all the letters in the walker. The third number is the Z score for this evaluation, assuming that individual information values form a Gaussian distribution. The Z score conveys the probability that a particular sequence is a member of the sites used to create the matrix. It is calculated by subtracting the mean (R_{sequence}) from the particular individual information content and dividing by the standard deviation of the distribution. If the Z score is below a given threshold (set by the user) and the evaluation is positive (or greater than some value set by the user) then the box is green to indicate that a binding site has been located. Otherwise the box is pink.

Because a walker shows the relationship between a sequence and the conserved bases at a binding site, we tend to interpret it as representing how a protein reacts to the DNA rather than how the DNA changes because of the protein contact. Both models are feasible in the case of Fis binding (23). However, the walker is valid for either model because information measures do not depend on the physical mechanism that determines the probabilities (15).

When walkers are created for the original set of sequences used to make the individual information weight matrix, the position-by-position average of the walker heights will, by definition, be the height of the sequence logo at that position (18). Thus walkers and logos are intimately related.



```
piece 1, M98776.9jan95, donor in GenBank, config: linear, direction: +, begin: 2334, end: 2353

      *         *2340     *         *2350
5' c g g a a c a a g t a a g g g a c c c t 3'
  + R - N - K + . - G - T -
... -----------------> M98776.9jan95.CDS.2127-2342

                    donor -6.8 bits, Z =-4.51, p = 0.00 @ 2343

                    donor  8.9 bits, Z = 0.26, p = 0.40 @ 2342

piece 2, M98776.9jan95, acceptor, config: linear, direction: +, begin: 2633, end: 2668

      *         *2640     *         *2650     *         *2660     *
5' t g a c t c c a a a a g a t c a g c t t c c t c t a g t a t g a g g a t 3'
 - . - L - Q - K - I - S - F - L - . + Y - E - D +
                                       [----------------- ... M98776.9jan95.CDS.2660-2719
                                          acceptor 1.0 bits, Z =-1.72, p = 0.04

piece 3, M98776, revision from J. Compton, config: linear, direction: +, begin: 2643, end: 2678

      *         *2650     *         *2660     *         *2670     *
5' t c c t a a t t g t t t c t g c a a a c c t g c a g g t a t g a g g a t 3'
 - S - . - L - F - L - Q - T - C - R + - Y - E - D +
                                      [------------------ ... M98776.CDS.2669-2729
                                         acceptor 7.9 bits, Z =-0.21, p = 0.42
```

Figure 5. Detection and visualization of an anomaly in GenBank by a **lister** map. The splice junction models come from data sets previously described (5). Yellow boxes mark the coded peptides. The **exon** program reads GenBank entries and produces a feature file for **lister** containing coding sequence (CDS) locations. These are named by their piece name and coordinates. The sequence M98776.9jan95 was obtained from GenBank 98.0, while the revised sequence M98776 was obtained from GenBank 99.0. The walker at 2343 of piece 1 was 'forced' to be there by scanning a small sequence fragment (**delila** instructions: get from 2343 –3 to same +6;) with a low R_i cutoff, and then joining this to the other features in the figure found by a scan of the entire piece with a higher cutoff.

Graphical searching

How does the individual information weight matrix react as it passes by a binding site during a search? The walkers in Figure 1 show both positive and negative contributions to binding. The majority of the letters are downwards but many 'flip up' when the walker lands on a known Fis binding site at coordinate 180. The last frame at the bottom of the figure shows that bases –7 and +7 (relative to coordinate 182) are strongly preferred by Fis, yet because other positions are unfavorable the overall sum is below zero and this position is not expected to be a binding site. Other subtle details can be observed in these walkers. In the top three 'frames', there are five A or T bases with heights that follow the sine wave. This effect, which has been also observed in sequence logos, probably reflects the cylindrical geometry of DNA/protein interactions (2,4).

Viewing medium sized regions of sequence

The **scan** program is used to locate likely binding sites according to three cutoff criteria as set by the user: (i) individual information, (ii) absolute value of the Z score (to eliminate both abnormally low and high values) and (iii) probability of obtaining that Z score or lower. The site locations for one kind of binding site can be joined to those of others, and features defined by hand can be added. When these are graphed by the **lister** program, one can see more than one binding site at a time, but the display is not interactive.

lister can help to identify the locations of binding sites by fitting the individual information model to biochemical data. For example, Slany and Kersten detected two Fis sites in the *tgt/sec* promoter region and identified one at coordinate –58 (31). A **lister** 'map' of the region (Fig. 2) shows this site and the likely location of the second Fis site. Both sites are stronger than the average Fis site, which is about 8 bits (23). This result is consistent with the likely size of the DNase I protected region, DNase I hypersensitive sites at –78.5 and –79.5 (which are found in the region from ±7.5 to ±4.5 of a Fis site (23,32,33), two gel electrophoresis band-shifts observed with DNA fragments spanning the region from –137 to +12 but only one for the region from –68 to +12, and effects on transcription *in vivo* (31).

Displaying complex genetic regions

Figure 3 shows a **lister** map of the λ *att* region, containing Fis and IHF walkers. IHF binds DNA asymmetrically. To represent this, **lister** can be told to rotate letters 90° according to the orientation of the binding site. For IHF site H′ (at coordinate 36), the letters are rotated counter-clockwise so the direction that one reads down through the walker letters is the standard 5′→3′ direction of the site (34). IHF sites H1 (at –116) and H2 (at –38) are oriented in the opposite direction so that their letters are rotated clockwise. To make their pattern correspond to H′, the bases are for the complementary strand. On this **lister** map, the four letters take on all four orientations without any ambiguity.

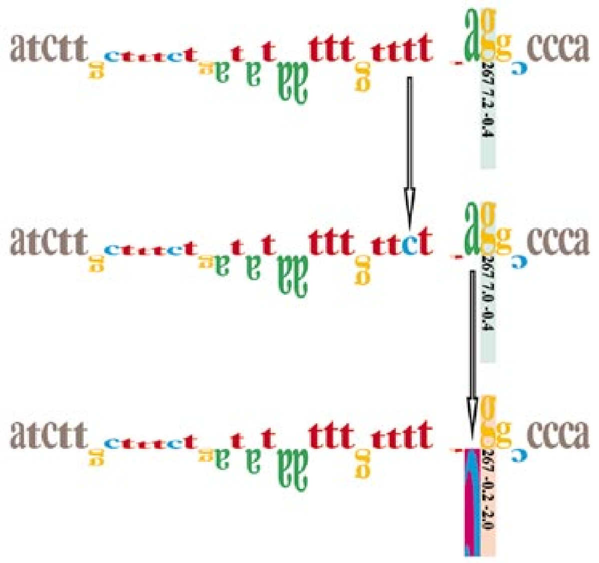


Figure 6. Distinguising mutations from polymorphisms with **makewalker**. The top walker shows a wild type human MSH2 splice junction (GenBank accession U41218 coordinate 267). The middle walker shows a sequence change proposed to cause familial non-polyposis colon cancer (41). This change is found in 2 of 20 normal individuals (43) and therefore is a polymorphism. The bottom walker shows a hypothetical strong mutation. The weight matrix was created from the data set described in ref. 5.

The previously identified 9.9 bit Fis site at –66 overlaps Xis site 2 (35–37) and a second weaker 4.5 bit Fis site at –55 is predicted to be 11 bases away (23). Fis is involved in both integrative and excisive recombination and will stimulate excision when the concentration of Xis is low (37–40). Fis binding to the 9.9 bit site excludes Xis binding at Xis 2 and stimulates Xis binding at Xis 1 (37). Walkers show that at least four more Fis sites may be involved in λ *att* site specific recombination.

Quantitative genetic engineering

Figure 4 shows how a binding site can be engineered. The top walker is the *hin* proximal site at coordinate 132 of GenBank entry V01370. It is an 8.3 bit site only 0.1 standard deviations above the mean. The middle walker shows the effect of introducing 5′-GAATTC-3′ into the sequence at positions –3 to +2 to create an *Eco*RI restriction site. The binding site is expected to be slightly weaker at 5.8 bits and 0.7 standard deviations below the mean. This can be compensated for by noting that the T at –7 is not good for binding. In the bottom walker this has been changed to a G and the adjustment raises the information content to 10.2 bits, which should be stronger than the original site.

Detecting database anomalies

lister has also helped to identify misplaced splice junctions. For example, in Figure 5, piece 1 is the donor region for human epidermal keratin intron 2. The location of the zero coordinate of a donor is, by definition, on the first base of the intron (coordinate 2343) (5). A walker set at this position gives –6.8 bits, which should be non-functional ($P < 2 \times 10^{-6}$, indicated by the pink box), therefore something is amiss. The corresponding acceptor site is also poor, as shown in piece 2. The GenBank entry recorded the problem by stating that the intron ‘does not fit consensus’. The **scan** program reveals a strong donor just inside the coding region at 2342. If this donor were used, the coding frame would be shifted by one base to the 5′, destroying the reading frame. Upon contacting the authors, it was learned that the gene had been resequenced and corrections were found but these had not yet been reported to GenBank. In the revised sequence the donor region is unchanged, but the acceptor region is altered (piece 3) so that the acceptor is now a ‘healthy’ 7.9 bits. The acceptor has also shifted one base to the 5′ so the reading frame of the polypeptide, RNK|YED is preserved by using the predicted donor at 2342.

Distinguishing mutations from polymorphisms

A clinical application of individual information analysis is shown in Figure 6. The top sequence is a human splice acceptor site found in normal colon tissue. The middle sequence is an alteration found in a sporadic colorectal tumor. The T→C change at position –5 was proposed to be the cause of the cancer (41), but the walker shows that this change is not significant since the individual information only changes by 0.2 bits and the number of standard deviations from the mean is still below 1. Thus, this change should represent a polymorphism and not a splice junction mutation (42). Indeed, a C appears in 2 of 20 normal people (43). The true mutation lies elsewhere or this alteration represents a disruption in the binding site for some molecule other than the spliceosome. The bottom row shows the effect of altering the sequence in the top row: when position –1 is changed to a C, the individual information becomes negative and the Z score approaches significance ($P < 0.02$). Such a mutation in the spliceosome recognition site would be severe and might predispose a person to colon cancer.

NOTE ADDED IN PROOF

Dr Paul N. Hengen has demonstrated by gel shift analysis that the predicted *tgt/sec* Fis site at –73 binds Fis (23).

ACKNOWLEDGEMENTS

I thank Paul N. Hengen for suggesting the sine wave on walker DNA, Seth Taylor for suggesting that walkers could be outside the DNA sequence; Denise Rubens for suggesting the mutation function, Stacy Bartram for writing **dnaplot**; Doris Schneider for pointing out that walkers show total protein/nucleic acid interactions even though only the letters change; Rick Wolf, Ken Kraemer and John Compton for sharing SoxS and splice junction data before publication; Paul N. Hengen, Pete Rogan, Denise Rubens, Stacy Bartram, Howard Young, and Kenn Rudd for comments on the manuscript; and the Frederick Biomedical Supercomputing Center for computer resource support.

REFERENCES

1 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
2 Papp,P.P., Chattoraj,D.K. and Schneider,T.D. (1993) *J. Mol. Biol.*, **233**, 219–230.
3 Papp,P.P. and Iyer,V.N. (1995) *J. Mol. Biol.*, **246**, 595–608.
4 Schneider,T.D. (1996) *Methods Enzymol.*, **274**, 445–455. ftp://ftp.ncifcrf.gov/pub/delila/oxyr.ps
5 Stephens,R.M. and Schneider,T.D. (1992) *J. Mol. Biol.*, **228**, 1124–1136.

6 Schneider,T.D. (1993) In Keller,P.R. and Keller,M.M. (Eds) *Visual Cues - Practical Data Visualization*. IEEE Press, Piscataway, NJ, USA. p. 64.
7 Pietrokovski,S., Henikoff,J.G. and Henikoff,S. (1996) *Nucleic Acids Res.*, **24**, 197–200.
8 Pietrokovski,S. (1996) *Nucleic Acids Res.*, **24**, 3836–3845.
9 Blom,N., Hansen,J., Blaas,D. and Brunak,S. (1996) *Protein Sci.*, **5**, 2203–2216.
10 Shaner,M.C., Blair,I.M. and Schneider,T.D. (1993) In Mudge,T.N., Milutinovic,V. and Hunter,L. (Eds) Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences. Volume I: Architecture and Biotechnology Computing. IEEE Computer Society Press, Los Alamitos, CA, USA. http://www-lmmb.ncifcrf.gov/~toms/paper/hawaii/ pp. 813–821.
11 Staden,R. (1984) *Nucleic Acids Res.*, **12**, 505–519.
12 Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) *Nucleic Acids Res.*, **10**, 2997–3011.
13 Stormo,G.D., Schneider,T.D. and Gold,L. (1986) *Nucleic Acids Res.*, **14**, 6661–6679.
14 Stormo,G.D. (1990) *Methods Enzymol.*, **183**, 211–221.
15 Pierce,J.R. (1980) *An Introduction to Information Theory: Symbols, Signals and Noise*. Second Edition. Dover Publications, Inc., New York, NY, USA.
16 Sloane,N.J.A. and Wyner,A.D. (1993) *Claude Elwood Shannon: Collected Papers*. IEEE Press, Piscataway, NJ, USA.
17 Schneider,T.D. (1995) Information Theory Primer. http://www-lmmb.ncifcrf.gov/~toms/paper/primer/
18 Schneider,T.D. (1997) *J. Theor. Biol.*, In press.
19 Schneider,T.D., Stormo,G.D., Haemer,J.S. and Gold,L. (1982) *Nucleic Acids Res.*, **10**, 3013–3024.
20 Schneider,T.D., Stormo,G.D., Yarus,M.A. and Gold,L. (1984) *Nucleic Acids Res.*, **12**, 129–140.
21 Schneider,T.D. and Mastronarde,D. (1996) *Discrete Appl. Math.*, **71**, 259–268. ftp://ftp.ncifcrf.gov/pub/delila/malign.ps
22 Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) *J. Mol. Biol.*, **188**, 415–431.
23 Hengen,P.N., Bartram,S.L., Stewart.L.E. and Schneider,T.D. (1997) *Nucleic Acids Res.*, In press.
24 Jensen,K. and Wirth,N. (1975) *Pascal User Manual and Report*. Springer-Verlag, New York, NY, USA.
25 Adobe Systems Incorporated (1985) *PostScript Language Reference Manual*. Addison-Wesley Publishing Company, Reading, MA, USA.
26 Adobe Systems Incorporated (1985) *PostScript Language Tutorial and Cookbook*. Addison-Wesley Publishing Company, Reading, MA, USA.
27 Shannon,C.E. (1948) *Bell System Tech. J.*, **27**, 379–423 and 623–656.
28 Schneider,T.D. (1991) *J. Theor. Biol.*, **148**, 83–123. http://www-lmmb.ncifcrf.gov/~toms/paper/ccmm/
29 Schneider,T.D. (1994) *Nanotechnology*, **5**, 1–18. http://www-lmmb.ncifcrf.gov/~toms/paper/nano2/
30 Schneider,T.D. (1991) *J. Theor. Biol.*, **148**, 125–137. http://www-lmmb.ncifcrf.gov/~toms/paper/edmm/
31 Slany,R.K. and Kersten,H. (1992) *Nucleic Acids Res.*, **20**, 4193–4198.
32 Finkel,S.E. and Johnson,R.C. (1992) *Mol. Microbiol.*, **6**, 3257–3265.
33 Finkel,S.E. and Johnson,R.C. (1992) *Mol. Microbiol.*, **6**, 1023.
34 Goodrich,J.A., Schwartz,M.L. and McClure,W.R. (1990) *Nucleic Acids Res.*, **18**, 4993–5000.
35 Sanger,F., Coulson,A.R., Hong,G.F., Hill,D.F. and Petersen,G.B. (1982) *J. Mol. Biol.*, **162**, 729–773.
36 Landy,A. and Ross,W. (1977) *Science*, **197**, 1147–1160.
37 Thompson,J.F., deVargas,L.M., Koch,C., Kahmann,R. and Landy,A. (1987) *Cell*, **50**, 901–908.
38 Ball,C.A. and Johnson,R.C. (1991) *J. Bacteriol.*, **173**, 4027–4031.
39 Ball,C.A. and Johnson,R.C. (1991) *J. Bacteriol.*, **173**, 4032–4038.
40 Numrych,T., Gumport,R.I. and Gardner,J.F. (1992) *EMBO J.*, **11**, 3797–3806.
41 Fishel,R., Lescoe,M.K., Rao,M.R.S., Copeland,N.G., Jenkins,N.A., Garber,J., Kane,M. and Kolodner,R. (1993) *Cell*, **75**, 1027–1038.
42 Rogan,P.K. and Schneider,T.D. (1995) *Human Mut.*, **6**, 74–76.
43 Leach,F.S., Nicolaides,N.C., Papadopoulos,N., Liu,B., Jen,J., Parsons,R., Peltomäki,P., Sistonen,P., Aaltonen,L.A., Nyström-Lahti,M. *et al.* (1993) *Cell*, **75**, 1215–1225.
44 Hübner,P. and Arber,W. (1989) *EMBO J.*, **8**, 577–585.
45 Landy,A. (1989) *Annu. Rev. Biochem.*, **58**, 913–949.
46 Thompson,J.F., Fong,H., Mark,L., Franz,B. and Landy,A. (1988) In Olson,W.K., Sarma,M.H., Sarma,R.H. and Sundaralingam,M. (Eds) *Structure and Expression*. Volume 3: DNA Bending and Curvature. Adenine Press, New York, NY, USA. pp. 119–128.