Xlandscape: a tool for the graphical display of word frequencies in sequences

S. Levy, L. Compagnoni, E.W. Myers and G.D. Stormo

July 23, 1997

Abstract

Motivation: To provide a graphical interface for the generation, display and manipulation of a sequence landscape that will run on all X-windows based UNIX workstatations. *Results:* The sequence landscape approach enables the representation of the frequency of occurence of all query sequence sub-words within a database. The landscape approach can detect tandem and other repeating word motifs, specific sub-words that are over–represented words in a particular database using Markov probability and the preference for sub-words belonging to either one of two databases. All these features aid in the classification of a query sequence and given the open text format for sequences and databases the XLANDSCAPE tool can be applied to a wide range of problems including sequence classification. *Availability:* XLANDSCAPE is freely available by anonymous ftp to beagle.colorado.edu, directory, pub/Landscape/xland.v1.tar.Z. *Contact:* Email address, Samuel.Levy@colorado.edu or Gary.Stormo@colorado.edu.

1 Introduction

The ability to compare and classify biological sequences is an important prerequisite to understanding biological function. The simpliest form of comparison involves matching the sub-words of a query sequence against another source sequence or collection of source sequences. If a set of long query words occurs with a high frequency in the source sequence then this would suggest relatedness between the query and source sequences. The frequencies of mono- and di-nucleotidies are used as part of the first order Markov chain analysis that enabled the WORDUP program to identify functionally significant promoter words (Pesole *et al.*, 1992). Indeed, word frequency data of di- and tetranucleotides provides the basis of the ability to discriminate between different bacterial genomes (Karlin *et al.*, 1997) and eukaryotic organisms (Karlin and Burge, 1995). Word frequencies can also be used to identify tandem or inverted repeating elements within a sequence (Lefevre and Ikeda, 1993), (Milosavljevic and Jurka, 1993), (Cox and Mirkin, 1997). Tandem repeating DNA or protein elements have been implicated in nervous system development and related disorders (Karlin and Burge, 1996), (MacDonald and Gusella, 1996), (Zoghbi, 1997), (Nelson, 1993).

Word frequency data of a sequence compared with databases representing defined sequence classes can provide a useful first step in the classification of a sequence into functional domains. Claverie and Bougueleret (Claverie and Bougueleret, 1986) calculated frequencies of words up to 4- and 9-long found in protein and nucleic acid sequence databases, respectively, and then used this information to find words occuring in a query sequence. Combining information from word frequencies and discriminant analyses it was possible to identify intron/exon regions in DNA and functional domains in proteins. In a similar manner, hexamer or octamer word frequencies provide some of the data that can be combined with other types of protein coding sequence data to aid in identifying protein coding regions of genes (Uberbacher and Mural, 1991), (Snyder and Stormo, 1993), (Snyder and Stormo, 1995), (Solovyev *et al.*, 1994) or for the identification of over-represented words in promoter regions (Prestridge, 1995), (Hutchinson, 1996), (Chen *et al.*, 1997).

One potential extension to this approach is to use frequencies of all word lengths found in the query sequence that occur in sequence class databases. The result of this analysis can be represented as a *sequence landscape* or a matrix array of frequency values above the query sequence (Clift *et al.*, 1986). This approach previously illustrated the presence of repeating motifs and underepresented words in bacteriophage genomes. However, the sequence landscape is a general approach for the analysis of any textual alphabet, and thus can be applied to both DNA and protein sequences. Therefore, in order to facilitate the use of word frequency information for word matching, sequence comparision and classification we present an X-windows implimentation of the sequence landscape program and some examples of enhancements in landscape visualiza-

tion.

2 System and methods

The Xlandscape program comprises approximately 13,000 lines of C code that was developed on Sun, DEC and Hewlett Packard Unix workstation operating under UNIX. The program employs X-windows calls (X11 Release 5 and the Athena widget set) to allow the user access to sequences and databases as well as controlling different aspects of landscape display and filtering. The source code can be obtained at ftp://beagle.colorado.edu/pub/Landscape/xland.v1.tar.Z and in theory can be built for any Unix workstation with a C compiler, X11 Release 5 or higher and the Athena widget toolkit. The program produces an X-windows graphical display of a landscape which includes zooming, scaling and word searching capabilities. Word frequencies of a query sequence in a database can be represented by user defined color coded rectangular cells or may contain the numeric values themselves. A clickable interface enabled on the landscape cells themselves allows the user to access information about the length, position, and frequency value, ratio of frequency values and Markov probability of any word.

3 Algorithm

3.1 Landscape description

A landscape of a simple 15 base query sequence, using the same sequence as a database, shows the basic elements of a landscape display (Figure 1). Every cell in the landscape, $L(j,k)$, indicates where a word begins,j, and its length, k, and the value of a cell is the frequency of that word in the database. For example, the frequency value of 2 at $L(3,3)$ indicates that the 3-base word *tcc* starting at position 3 occurs twice in the sequence, the other location being at position 8. The landscape also shows frequencies of two- and one-letter words in rows 2 and 1 respectively, or $L(*,2)$ and $L(*,1)$. The diagonal lines found at $L(4,2)$ and $L(9,2)$ indicate that the word *cc* occurs within the context of word "pointed" to by the diagonal line, *ie. tcc*. In addition, this sub-word occurs at the same frequency, which in this case is 2. Thus, the landscape represents the frequency of all words in the query sequence that can be found in a database. Furthermore, the peaks indicate the presence of relatively long matches in the query sequence to words in the database, diagonal and vertical lines indicate word implications and troughs indicate the presence of words within the query sequence that occur infrequently in the database.

3.2 Landscape production

The basic outline of landscape production is shown in Figure 2. In order to enable on-line word searching of a query sequence, a suffix array data structure (Manber and Myers, 1993) is created of the database(s) which takes $O(N \log N)$ time for a database of length N. Subsequently, searching for query sequence words in the indexed database(s) can occur in $O(P + \log N)$ time for query sequences of length P. In the case where a query sequence is searched against a single database ($x = h$ in Figure 2) the frequency values of sub-words $L(j,k)$ are displayed.

In order to determine the preference of a query sequence word for occurring within one database, say D_h, over another, say D_i, ($x = h, i$ in Figure 2) a ratio landscape can be produced. This is essentially the result of generating two landscapes of the same query sequence by using two different databases and dividing all the cells in one landscape by the corresponding cells of the other. Thus, we define each cell in the landscape as a log ratio of frequency values,

$$R_{h/i}(j,k) = \ln \frac{L_h(j,k)+1}{L_i(j,k)+1} + \ln \frac{|D_i|}{|D_h|} \tag{1}$$

Therefore, if a word occurs with a higher frequency in database D_h then $R_{h/i}(j,k) > 0$. Alternatively, if a word occurs with a higher frequency in database D_i then $R_{h/i}(j,k) < 0$. If a word occurs with an equivalent frequency in either database then $R_{h/i}(j,k) = 0$. The addition of 1 to the $L(j,k)$ terms serves as a small sample size correction and avoids any potential erroneous attempt to evaluate $\ln 0$. The last term in equation (1) accounts for differences in database size such that $R_{h/i}(j,k)$ represents a quantitative measure of a word's preference in any particular database, regardless of database size. This so-called ratio landscape can be generated on-line and displayed by simple menu operations.

The landscape contains a large amount of information and in order to extract more meaningful information a filtering operation can be employed that expresses each word as a ratio of observed:expected probabilities. A Markov model of degree $l-1$, where l is the word length, is applied to derived expected values for each word in the landscape. Thus, the observed and expected values for each word can be obtained from the landscape cells $L(j,k)$. The filtered landscape values can be represented thus,

$$M(j,k) = \ln \frac{Obs}{Exp} = \ln \frac{L(j,k) \times L(j+1,k-2)}{L(j,k-1) \times L(j+1,k-1)} \tag{2}$$

when $k = 1$ then $M(j,k) = 0$ and when $k = 2$ then the value of $L(j+1, k-2) = |D|$ in equation (2).

4 Implementation and results

The Xlandscape program was developed and tested extensively on a Sun Sparcserver 10/51. The initial overhead for the Xlandscape windowing interface required approximately 2 megabytes of RAM. Subsequently, generating a suffix array of each database required ten times as much RAM as the original database size, which is platform independent. On the Sun Sparcserver 10/51, this indexing was accomplished at a rate of approximately 13 kbases/second. Therefore, a one megabase database required approximately 76 seconds for indexing and 10 megabytes of RAM for storing the index in memory. Once a database had been indexed, word searching of a query sequence can occur on-line with an additional overhead in RAM.

The detection of tandemly repeating DNA sequences in genes is readily performed by using the same DNA sequence as both the query and database and by generating the landscape of $L(j,k)$ values. Triplet repeating elements such as CAG have been implicated in diseases such as muscular atrophy and Huntington's. This repeating element can be detected in the Huntington's disease related gene (GenBank locus HUMHDA) (Huntington's disease collaborative research group, 1993) by generating a landscape using the gene itself as the query sequence and database (Figure 3). The tandemly repeating element is resolved as a large peak in the landscape of a small region of the gene since frequently occurring larger words can be found that are built made up of the core CAG word. For example, the cursor is situated on the 50-long word $CAG_{16}CA$ can be found 6 times within the larger 65-long word $CAG_{21}CA$. The CAG repeat expansion by virtue of encoding a poly-glutamine stretch, appears to be the most likely cause for Huntington's (MacDonald and Gusella, 1996), the landscape also detects an adjacent CCG_7 repeat (smaller peak in Figure 3) or unknown function.

The landscape can be used to determine the extent of similarity between two sequences. The landscape frequency values are the result of a comparison of all query sub-words to an indexed sequence. Therefore, the number and length of peaks provide a quantitative measure of similarity between these two sequences. This approach has been implimented previously by applying a minimal–length encoding algorithm (Milosavljevic and Jurka, 1993). The extent of similarity between sequences can be measured by comparision of the degree of compression between sequences (Milosavljevic, 1994). The cells that constitute the top edge of a landscape provide the same information as the minimal–length encoding algorithm. Using a landscape from a pairwise comparison of two sequences permits matching of words that essentially allows transposition events in the matching process. This factor permits an additional flexibility in the sequence matching scheme over existing alignment based methods.

The ratio landscape of an intron/exon boundary region in the human bone marrow serine protease gene (GenBank locus HSMED) on comparsion with exon (D_e) and intron (D_i) databases can be seen in Figure 4. The ratio frequency

values, $R_{e/i}(j,k)$, have been color coded to reflect positive (yellow) and negative values (blue) which indicate a preference for words in exon and intron databases respectively. From the overall distribution of the shaded cells it is evident that exon words occur preferentially in the known exon domain which starts at position 1786, whilst the intron words are more prevalent in the intron domain ending at position 1785.

The calculation of a Markov landscape permits the enquiry whether a sequence contains words that are either over- or under-represented in a database according to Markov probability. In the case of word over–representation, the word in question may be regarded as containing sequence particularly relevant to the functional class that the database represents. This can be illustrated by calculating the Markov landscape of a putative promoter region of the HSMED gene (Figure 5) using a database of 430 unique eukaryotic vertebrate promoter sequences prepared from eukaryotic promoter database (version 48) ((Bucher, 1996)). The red cells indicate words that occur at least two-fold more than expected in the promoter database. The eleven-letter word highlighted at position 1229 contains the sequence *ctataagagga* which is a good candidate for the TATA box sequence. In addition, the nine–letter words indicated by the five landscape cells upstream from position 1229 are all identified as potential transcription factor binding sequences when this region is compared to the TRANSFAC database (version 3.2) using MatInspector 2.1 (Quandt *et al.*, 1995). Thus, the Markov landscape is very efficient in detecting functionally relevant sequences in this particular example.

The "troughs" or "valleys" in a landscape indicate words that occur at a reduced frequency to adjacent words of the same length. This can be seen in Figure 3 in the landscape cells 429,5 and 441,5 which contain the frequency of 7 for these two words compared to the other adjacent 5-letter words which occur more than 10 times (marked by "*" in row 5 of Figure 3). The Markov landscape of a sequence also can indicates under–represented words and this is illustrated in the HSMED gene in Figure 5. The words indicated by the green cells occur at least two–fold less than expected compared to all promoter sequences. The occurrence of under–represented words in a region of sequence that potentially contains transcription factor binding site may serve to increase the specificity of transcription factor binding. The landscape approach previously identified the under-represented methylation sequence tetramer, GATC, in an analysis of the bacteriophage T7 genome (Clift *et al.*, 1986). Using a similar Markov statistic, it was possible to identify GATC as a relatively under-represented sequence in most bacteriophage genomes (Karlin *et al.*, 1994).

5 Discussion

The Xlandscape program is a graphical interface operating on all X-windows based Unix workstations that generates a "landscape" of word frequency val-

ues of all sub-words occurring within a query sequence when compared to a database. This graphical interface is an enhancement of the text-based interface that was developed previously (Clift *et al.*, 1986). A sequence landscape can be produced in real-time due to the indexing of the database and maintaining the suffix array of the database in RAM, enabling rapid querying of all sub-words contained within a query sequence.

In the simplest implimentation a sequence landscape can be used to determine whether a query sequence sub–word occurs in a database. The occurrence of a tandem repeat sequence can be detected in a landscape as a large "peak" with many word implications due to the overlapping sub–words that occur within the longest word constituting the extent of the tandem repeat (Figure 3). Repeating motifs of this kind can be detected within any biological sequence, DNA or protein, when the query sequence is also used as a database.

The Markov landscape approach provides another means by which a specific word or group of words can be detected in a feature database. The application of this filter to the upstream region of the transcription start site of the GenBank locus HSMED versus a promoter database reveals a potential TATA box and several sites that could be potential transcription factor binding sites (Figure 5). These putative transcriptional promoters were also detected by MatInspector version 2.1 (Quandt *et al.*, 1995), indicating that using all word frequencies can provide as useful a diagnostic feature as have been found for weight matrices for transcription factor binding sites (Prestridge, 1995) (Chen *et al.*, 1997) (Hertz and Stormo, 1996).

The addition of the Markov and ratio landscape methodologies has extended the capabilities of the sequence landscape approach as a classification tool. Several approaches to gene finding employ hexamer or octamer word frequencies from exon regions as one diagnostic feature to predict *de novo* intron/exon boundaries (Uberbacher and Mural, 1991), (Snyder and Stormo, 1993), (Snyder and Stormo, 1995), (Solovyev *et al.*, 1994)). This frequency data is equivalent to using the ratio of word frequency values, either $R(*, 6)$ or $R(*, 8)$, *ie.* only row 6 or 8 in Figure 4. However, using all word lengths in the landscape essentially includes the most relevant words for determining the differences between different sequence classes. The ratio landscape is a convenient way to examine whether a query sequence belongs to one of two different sequence classes. In an attempt to extend this classification to include more than two sequence classes, a scoring algorithm can be applied to ratio landscapes of a query sequence compared to several sequence class databases. This approach was taken to parse a sequence into promoter, exon and intron domains and preliminary data suggests that this may be a viable approach to sequence classification (Levy and Stormo, 1997).

Acknowledgements

This research is supported by ....

References

Bucher, P. (1996) The eukaryotic promoter database EPD, EMBL nucleotide sequence data library, release 48.

Chen, Q., Hertz, G. and Stormo, G. (1997) PromFD 1.0: a computer program that predicts eukaryotic pol II promoters using strings and IMD matrices. *Comput. Appl. Biosci.*, 13, 29–35.

Claverie, J. and Bougueleret, L. (1986) Heuristic informational analysis of sequences. *Nucl. Acids Res.*, 14, 179–196.

Clift, B., Haussler, D., McConnell, R., Schneider, T. and Stormo, G. (1986) Sequence landscapes. *Nucl. Acids Res.*, 14, 141–158.

Cox, R. and Mirkin, S. (1997) Characteristic enrichment of dna repeats in different genomes. *Proc. Natl. Acad. Sci*, 94, 5237–5242.

Hertz, G. and Stormo, G. (1996) *Escherichia coli* promoter sequences: analysis and prediction. *Methods Enzymol.*, 273, 30–42.

Huntington's disease collaborative research group, T. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72, 971–983.

Hutchinson, G. (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comp. Appl. Biosci.*, 12, 391–398.

Karlin, S. and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Gen*, 11, 283–290.

Karlin, S. and Burge, C. (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci.*, 93, 1560–1565.

Karlin, S., Ladunga, I. and Blaisdell, B. (1994) Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci.*, 91, 12837–12841.

Karlin, S., Mrazek, J. and Campbell, A. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol*, 179, 3899–3913.

Lefevre, C. and Ikeda, J. (1993) Pattern recognition in dna sequences and its application to consensus foot-printing. *Comput. Applic. Biosci*, 9, 349–354.

Levy, S. and Stormo, G. (1997) DNA sequence classification using DAWGs. *Lecture Notes in Computer Science*, page In Press.

MacDonald, M. and Gusella, J. (1996) Huntington's disease: translating a CAG repeat into a pathogenic mechanism. *Curr Opin Neurobiol*, 6, 638–643.

Manber, U. and Myers, G. (1993) Suffix arrays: A new method for on-line string searches. *SIAM J. Comput.*, 22, 935–948.

Milosavljevic, A. and Jurka, J. (1993) Discovering simple DNA sequences by the algorithmic significance method. *Comp. Appl. Biosci.*, 9, 407–411.

Milosavljevic, A. (1994) Sequence comparisons via algorithmic mutual information. *Ismb*, 2, 303–309.

Nelson, D. *Genome rearrangement and stability*, volume 7, chapter Six human genetic disorders involving mutant trinucleotide repeats: similarities and differences, pages 1–24. Cold spring Harbor Laboratory Press, (1993) .

Pesole, G., Prunella, N., Liuni, S., Attimonelli, M. and Saccone, C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucl. Acids Res.*, 20, 2871–2875.

Prestridge, D. (1995) Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, 249, 923–932.

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) Matind and matinspector - new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nuc. Acids Res.*, 23, 4878–4884.

Snyder, E. and Stormo, G. (1993) Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucl. Acids Res.*, 21, 607–613.

Snyder, E. and Stormo, G. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, 248, 1–18.

Solovyev, V., Salamov, A. and Lawrence, C. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of splicable open reading frames. *Nucl. Acids Res.*, 22, 5156–5163.

Uberbacher, E. and Mural, R. (1991) Locating protein coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.*, 388, 11261–11265.

Zoghbi, H. (1997) The expanding world of ataxins. *Nature genetics*, 14, 237–238.

Figure legends

Figure 1. Landscape of 15-base sequence showing frequencies of all words that occur in this sequence.

Figure 2. Flow chart of Xlandscape program operation. A landscape of a sequence is generated by first producing a suffix array index of the source sequence or source database of concatenated sequences. All the words of an unindexed query sequence can then be compared to the indexed source and the X-windows display outputs the landscape cell values in either a numeric or color-coded fashion. The values outputted depend on whether the query sequences is compared to one or two database. These are the L values, for exact word frequencies, M values, for Markov probability of occurrence and R, for a log-ratio value representing a comparision of a query sequence with two databases.

Figure 3. Landscape of a small region of the first exon of the Huntington's disease gene (GenBank locus HUMHDA) showing CAG_{21} tandem repeats starting at position 367 as a larger peak. The top region of the peak is truncated in the figure but frequency values of 2 are found for longer words containing the CAG tandem repeat. The "*" represent single cells in which frequency values are greater than 9.

Figure 4. Screen image showing X-windows interface for displaying a sequence landscape. Displayed in the window is a ratio landscape of the intron-exon boundary (marked with cursor above sequence line at position 1786) for the GenBank locus HSMED showing preferentially occurring exon words in yellow and intron words in blue. Note that more exon words occur to the right of position 1786 (marked with a blocked cursor in red on sequence row) where an exon domain is known to exist.

Figure 5. Markov landscape of the HSMED gene, upstream from the transcription start site (position 1287–not shown), compared to a promoter database. Red cells indicate words that occur two-fold more times in the promoter database than expected. The 5 adjacent red cells to the left represent words that contain predicted transcription factor binding sequence.

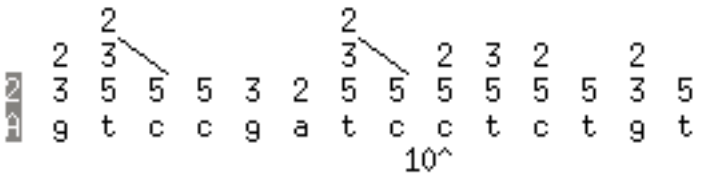

Figure 1:

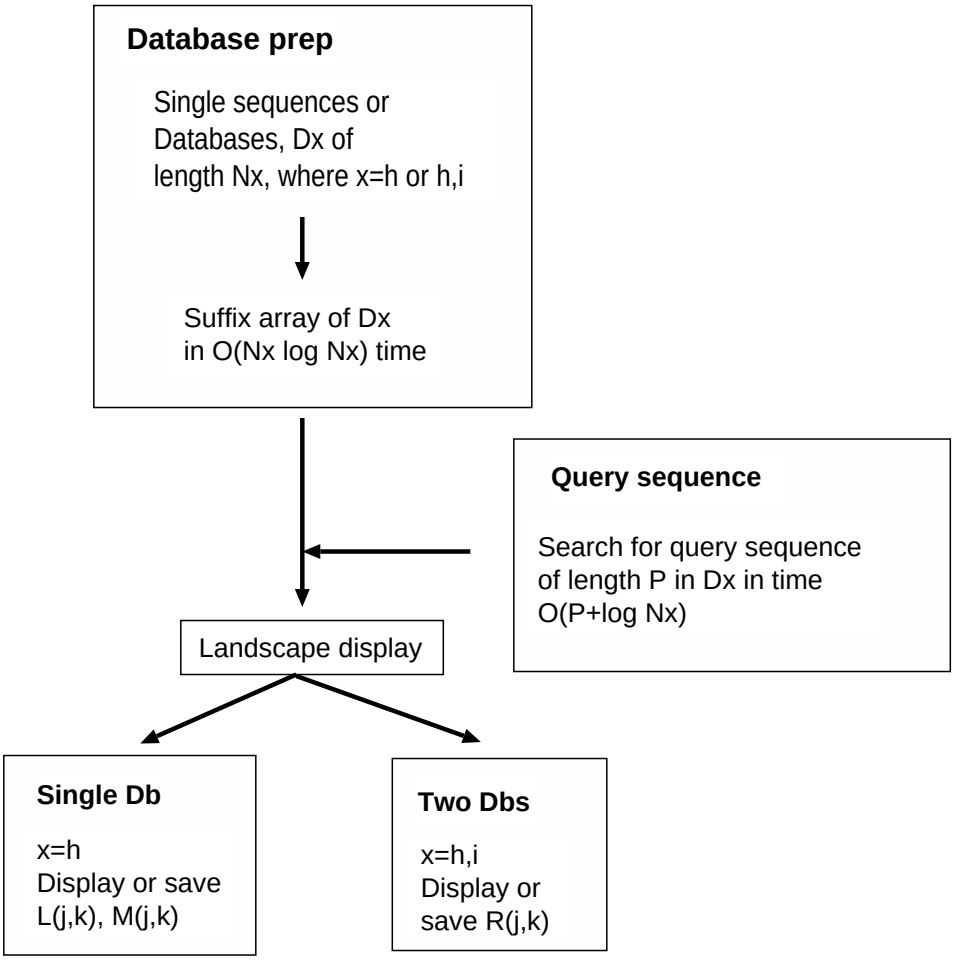
Database prep
Single sequences or
Databases, Dx of
length Nx, where x=h or h,i
Suffix array of Dx
in O(Nx log Nx) time
Query sequence
Search for query sequence
of length P in Dx in time
O(P+log Nx)
Landscape display
Single Db
x=h
Display or save
L(j,k), M(j,k)
Two Dbs
x=h,i
Display or
save R(j,k)

Figure 2:

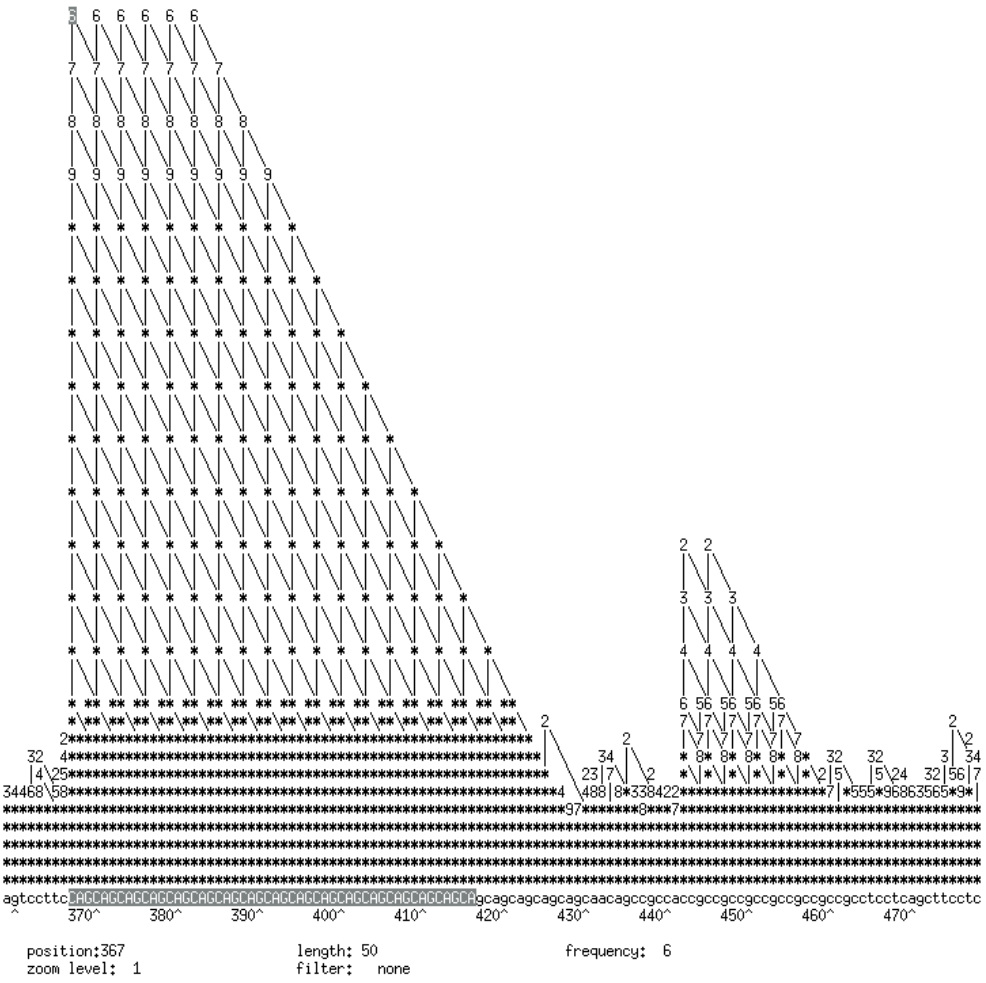
agtccttcCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAgcagcagcagcagcaacagccgccaccgccgccgccgccgccgcctcctcagcttcctc
370^ 380^ 390^ 400^ 410^ 420^ 430^ 440^ 450^ 460^ 470^
position:367 length: 50 frequency: 6
zoom level: 1 filter: none

Figure 3:

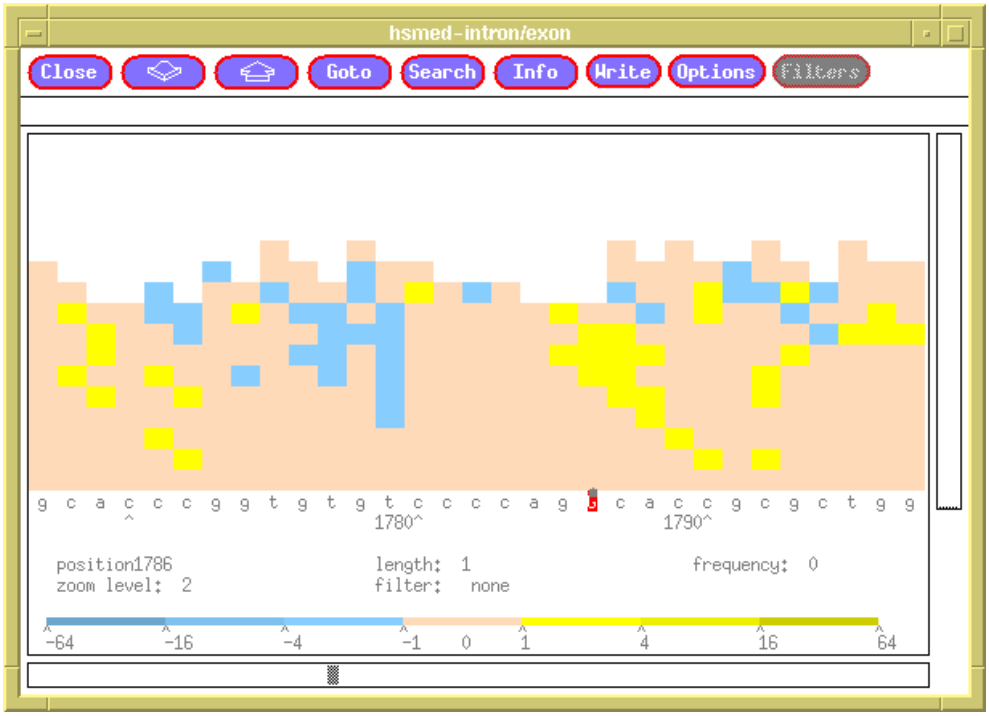
hsmed-intron/exon
Close
Goto
Search
Info
Write
Options
Filters
g c a c c c g g t g t g t c c c c a g g c a c c g c g c t g g
1780^
1790^
position1786
zoom level: 2
length: 1
filter: none
frequency: 0
-64
-16
-4
-1
0
1
4
16
64

Figure 4:

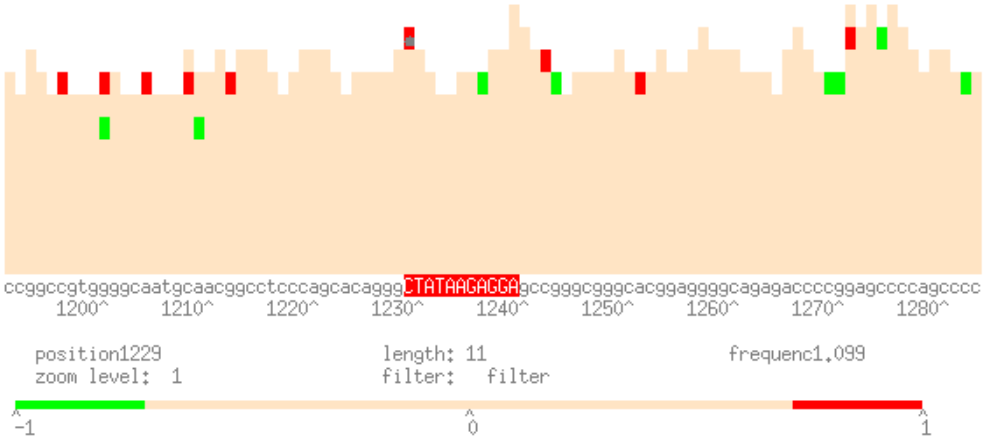
ccggccgtggggcaatgcaacggcctcccagcacagggCTATAAGAGGAgccgggcgggcacggaggggcagagaccccggagccccagcccc
1200^ 1210^ 1220^ 1230^ 1240^ 1250^ 1260^ 1270^ 1280^
position1229
zoom level: 1
length: 11
filter: filter
frequenc1.099
-1
0
1

Figure 5: