Combinatorial Relationships in Transcription Factors

Jack Middleton
BIOC 218 final project
March 10, 2003

Abstract

Determining the location of transcription factor binding sites is an important initial step in solving the complexity of genetic networks. Numerous computational methods have been proposed to assist biologists in finding putative transcription factor binding site locations. These tools are typically good at finding the signal of an individual binding site within the DNA strand but have low specificity for finding actual know binding sites because they do not account fact that transcription factors interact with each other. This project examines how these tools can be used to account for the combinatorial nature of transcription factors and proposes some improvements.

Introduction

In order to understand the operation of the genetic networks of organisms we need to first understand the mechanisms by which genes are transcribed. Transcription is initiated through the interaction of various transcription factors and the RNA polymerase. An initial step in understanding this interaction requires locating the binding sites for these transcription factors. Because locating the transcription factor binding sites through experiments is expensive and time consuming, the use of computational techniques to find and characterize transcription factor binding sites can be a great aid to biologists. Numerous computational systems have been developed to locate putative transcription factor binding sites including BioProspector, MEME, AlignAce and MDSCAN. Typically these tools look for signals for binding sites from a set of sequences taken from the upstream region of genes that appear to be co-expressed from micro array experiments.

Multiple EM for Motif Elicitation (MEME) is based on expectation maximization (EM). The EM algorithm first makes an initial estimate of the alignment and uses it to create a scoring matrix that represents the distribution of bases in the motif. This matrix is compared to each sequence and the scoring matrix values are updated to maximize the alignment of the matrix to the sequences. This is repeated interactively until it converges. MEME uses a gradient decent algorithm which is prone to get stuck in a local maximum. MEME uses various heuristics to try to avoid this problem but it is not guaranteed to return the best motif. The advantages of MEME is that it can find multiple motifs which can occur multiple times and does not require that each sequence contain the motif. It does require that the user estimate the size of the motif .

BioProspector and AlignAce are extensions of Gibbs sampling with a number of improvements for searching for transcription factor binding sites. Gibbs sampling is a stochastic technique that works on a set of sequences which are believed to contain the motif. It starts by randomly removing one sequence and creating an initial motif by randomly aligning the remaining sequences and computing the base composition probabilities for each position. Next the optimal alignment is found by sliding the sequence back and forth to maximize the ratio of the motif probability to the background probability. Next the sequence which was originally left out is put back into the motif . The start position is estimated by scoring each segment of the sequence against the matrix. Weights for each segment are assigning using the motif. The start position is picked at random using the weights to bias the seclection. This is repeated until the residue frequencies in the columns do not change. BioProspector incorporates a number of improvements over the basic Gibbs sampling algorithm. First it does not require that every sequence contain the motif. It also handles multiple copies of the same motif within the sequence. BioProspector can also search for a two-block motif where the motif is separated into two parts with a short segment in-between which does not contribute the the motif. In order for BioProspector to find the motif it is important to have the correct background model. BioProspector generates a 3^{rd} order hidden markov model based on the background that it is given. It is also important that the correct background data be used. Since different genomes have different base compositions and can have sections of repeats the background data should be from the same genome as the sequences. And be from promoter regions which do not contain the same binding sites as the query sequences. BioProspector can get stuck in a local maxima and therefore must be run multiple times to find the true optimal alignment. It is also relatively slow and not tractable for searching whole genomes on the current class of cpu's. Gibbs sampling is an effective method for detecting weak and complex signals in a set of sequences which makes BioProspector and AlignAce very sensitive at finding signals of transcription factor binding sites.

In contrast to BioProspector, MDscan uses a deterministic algorithm which always converges to the same result for a given set of data. MDscan works best when sequences can be separated into two groups. One group contains sequences believed to contain the motif while the other group contains sequences which do not contain the motif. MDscan starts with a word-enumeration strategy to look for w-mers in the top sequences. It then enumerates each non redundant w-mer (seed) and searches for all w-mers in the top sequences with at least m base pairs matching the seed. For each seed the top sequences are used to form a motif weight matrix. The weight matrix is evaluated by a maximum posterior scoring function which uses a measure of how often the matrix appears, how well the matrix is conserved and the probability of finding the motif by chance. The top 10-50 matrices are used to scan the remaining sequences. A new w-mer is added or removed from the weight matrix if doing so increases the score of the matrix. The algorithm usually stabilizes in around 10 interactions and the top candidate motifs are reported. The advantage of MDscan is that it is much faster than other methods such as BioProspector and is tractable for searching entire genomes. The search time increases only quadratically with respect to the total number of bases in the top sequences and linearly with respect to the number of bases in the remaining sequences (Liu 2001).

MatInspector differs from the first three tools described in that it looks for the signal of previously know transcription factors in order to find new binding sites for these transcription factors. The previous tools are used to look for sites in sequences which are believed to contain sites in common while MatInspector is from a class of tools which looks for occurrences of a know motif in individual sequences which may or may not contain the binding site. MatInspector builds a library of descriptions for binding sites which contain a nucleotide distribution matrix and an vector of conservation factors computed from the amount that each position is conserved in the motif. The library is constructed from literature references and entries in the TRANSFAC database (Matys 2003) which have been experimentally verified to be transcription factor binding sites. MatInspector searches for matches in two passes. The first pass computes a core similarity for the ratio of the sums of the score of the sequence to the maximum possible sore for that position. If the core similarity is above the threshold the matrix similarity is computed by recomputing the ratio while weighting each position by the conservation factor. The binding site for the transcription factor is described as a position weight matrix (PWM). The weight matrix has a matrix position for each base and each position in the binding site representing the probability of that base occurring in that position. The score for any particular site is the sum of the matrix values for that site's sequence. The advantage of using a tool like MatInspector is that TRANSFAC represents the binding site of transcription factors as a PWM. This allows MatInspector to take advantage of previously found binding sites. However, the use of a PWM is poor for finding binding sites of unknown transcription factors. PWM's are a better representation of the binding site than consensus sequences because they allow for the fact that some positions in the binding site are more conserved than others and should be more important for the activity of the site. Also the determination of whether a sequence matches a consensus sequence is usually a binary decision while a PWM can give a better quantitative understanding of how well a new sequence matches the binding site. One of the difficulties in using this method that PWM's do not always properly describe the binding site. PWM's assume that each position in the site makes an independent contribution and does not have an affect on the other positions. This is not always true. Attempts have been made to try different representations of PWM's such as di and tri nucleotide tables, neural networks and hidden markov models. Each of these different representations have been found to have some success in certain circumstances but none have been shown to be superior for all cases (Stormo 2000)

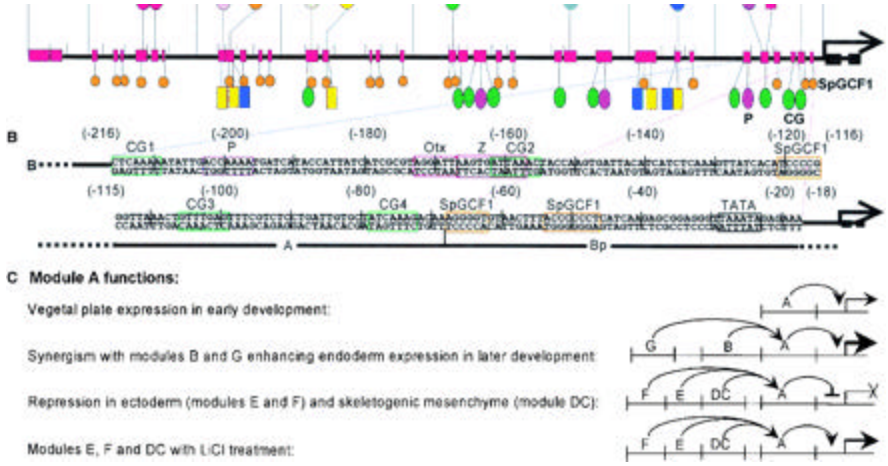
Within the last few years the performance of tools for finding transcription factor binding sites has improved but is still plagued by a low specificity causing too many false positives and in some cases low selectivity with too may false negatives (Fickett 1997). The problem is not that the tools are unable to detect signals but that the signals themselves are weak and occur frequently by random chance. Transcription factor binding sites are typically short (5-15bp) and are many times not well conserved. This causes many instances of the consensus motif of the site to occur by random chance causing noise that obscures the signals from the true binding sites. Two possible reasons for this are that the packaging of DNA into chromatin could cause some potential binding sites to be obscured in a way the prevents transcription factors from finding them. A

second reason is that many transcription factors interact with each other and the combinatorial effect has an influence on their ability to bind to the DNA strand. For example, the strength of the signals on the DNA strand may not be strong enough for either of two transcription factors to bind to it individually. However, if the two proteins are interacting with each other the combined signal from the two sites may be enough for the complex of the two proteins to bind to the DNA strand. The same holds true for multiple transcription factors. None of the computational tools discussed so far take into account the effect of the interaction of transcription factors have on their ability to bind to the DNA strand. In order for computational tools for finding transcription factor binding sites to improve these combinatorial relationships must be taken into account.

To date few computational tools have addressed this area. One example is COBIND (GuhaThakurta 2001) which looks for putative binding sites of two cooperatively binding transcription factors. The algorithm works by maximizing the joint likelihood of two binding site motifs. The binding sites are represented by two position weight matrices which are computed with an objective function derived from the thermodynamics of protein-DNA binding. The drawback to COBIND is it only checks for two bindings sites when sometimes multiple transcription factors are interacting and in contact with the DNA strand. Also, COBIND's ability to find the binding sites decreases as the distance between the individual binding sites increase. The use of an objective function based on the thermodynamics of protein-DNA binding has the advantage of giving a quantitate feel for the binding affinity but may be ignoring some of the signal that is encoded in the DNA sequence. The algorithm that COBIND uses is different than the two-block algorithm used in BioProspector which was designed to look for the binding sites of individual transcription factors that contact the DNA strand in two different locations. When the two-block option is used in BioProspector the highest scoring motif is removed from the matrix and then the next highest scoring motif is found. The combination the two motifs are reported.

Little work has been done in studying the combinatorial effect or transcription factors. One group analyzed micro array data to generate motif-association maps based on the combinatorial nature of their expression patterns (Pilpel 2001). Another group scanned the TRANSFAC and TRASCompel databases to find pairs of transcription factors known to interact and used a measure on their preference to co-localize at specific distances to predict new binding sites in the human genome. (Hannenhalli 2002). Another team did a study to try to characterize the positioning of TFs within regulatory regions to look for relationships which suggest that the proteins might be interacting (Wagner 1999)

One of the reasons so little has been computationally using the combinatorial effect of transcription factors is that the mechanism is not well understood. Difficulty in crystallizing the transcription initiation complex has made it difficult to understand exactly how transcription factors work together with the polymerase II to initiate transcription. Typically we only have solved structures for the segment of the transcription factors which are in contact with the DNA strand. This allows us to study how the transcription factor interacts with the DNA strand but we do not know what the



conformation of the rest of the protein or how it interacts with the other molecules in the transcription initiation complex(Brandon 1999)

One area that shows promise for studying the combinatorial relationship of transcription factors is using the fact that they tend to cluster together into modules (Davidson 2001). As an example, Figure 1 shows the upstream coding region of the endo16 gene in Stongylocentrotus, a sea urchin. This particular gene plays an important role in controlling the development of the Stongylocentrotus embryo. The figure show a 2300 bp section of DNA. Protein binding sites are shown as boxes. 13 different proteins bind with high affinity to 38 different sites in this region. The bindings sites tend to cluster into six groups or modules labeled A-G. The tendency for transcription factors to cluster into groups called cis-regulatory modules (CRM) has been observed in many eukaryotes. Numerous studies have used this concept to aid in locating and studying transcription factor binding sites in, for example, muscle specific genes (Wasserman 1998) and Drosophila melanogaster development (Berman 2002) (Markstein 2002).

Fig. 1 From Yuh(1998)

Knowing that binding sites tend to cluster into CRM's can be used to aid in studying the combinatorial relationships among transcription factors. In this project a set of known CRM's were targeted to look for relationships that could be used to develop computational tools and to gauge how well our current tools work in finding combinatorial relationships.

Methods

A set of 19 known cis-regulatory modules (CRM) from a study by Berman et al. (Berman 2002) were used as the test set. This data was analyzed by a combination of computational and manual inspection methods to look for patterns in the CRM's themselves. The motivation was to determine to what extent we can find combinatorial patterns within the CRM's. If we can find these patterns they could potentially be incorporated into our set of computational tools and used as an aid to both locate unknown CRM's and to suggest relationships which would help elucidate how they work. The CRM's were determined by first compiling at list of know transcription factors which were know to be involved in the same genetic pathways. Literature searches were done to find a list of DNA sequences that the transcription factors bind to. Next MEME

was used to align the sequences and the results were used to compute position weight matrices using the PASTER program. The authors then developed a web based tool called CIS-ANALYST to search for CRM's using the PWM's in a similar way that the MatInspector program does. From this they identified 19 CRM's .

We did an analysis of the data (from Berman 2002) which was used by PASTER to construct the PWMs. The purpose was to find relationships between different positions in the binding sites which could not be captured in a PWM. The Results showed that there were some preferred inter depencies among the positions in the binding sites. For example, in the binding sites for the hunchback transcription factor, if an A occurred in position 8 an A also occurred in position 7 82% of the time compared to 52% of the time for all sites. When an A appeared in position 8 an A always occurred in position 7 in the upstream regions of the eve, hb, ubx, en and kni genes. A similar pattern was found in the binding sites of the bicoid. When position 5 was a G, a T was in position 7 53% of the time compared to 43% for all sequences. Also in examining the PWM's themselves, in each matrix there were some positions which were highly conserved and others that were not. This supports the idea that PWM's are often superior to a consensus sequence in expressing the preference of a transcription factor because some positions are more important than others.

Next the sequences for the CRMs were run through MEME and AlignAce. While these sequences are different that what these programs are normally run on (upstream sequences of co-expressed genes) the goal was to see if we could detect signals that implied some kind of relationships. In examining the output it was apparent that there are fairly long segments that appear to be conserved within the CRM that did not correspond to the known binding sites. This needs to be taken into account when scanning for binding sites within CRM's because these motifs were always returned as the best motifs. However, MEME did find the motif for the hunchback transcription factor with a consensus sequence of: ATTTTTTATGG as it's second highest scoring motif. And AlignAce found the motif for Bicoid as its 8th highest scoring motif.

Searches were submitting multiple times to BioProspector and Mdscan but the results never returned. Cause is unknown.

Also a multiple sequence alignment was done on the CRM sequences to to see if there were conserved patterns within the CRM sequences. Also, large conserved sections of the sequences could be make it very difficult for any promoter find algorithm of find short motifs representing binding sites. The BCM Search Launcher from the Baylor College of Medicine was used to perform a ClustalW multiple sequence alignment. It was difficult to see any relationships in the alignment.

To look for relationships among the spacing and orientation for the binding sites the CIS-ANALSYT(http://www.fruitfly.org/cis-analyst) was used to display the locations of predicted binding sites. The program has a web interface and asks which transcription factors to look for in either a single gene upstream region or the entire genome. It then uses the PWM's for the know transcription factors to search for putative binding site

locations and display an annotated plot of the binding site locations. The results are show in Figures 2-5 in Appendix I. The results showed that the binding sites definitely tend to cluster and there are some preferences. Some TFs, such as hunchback, tend to form a triplet of two sites relatively close together and a third slightly further away. In some cases it appears that if a TF only appears twice in a cluster they are spaced close together but when three are in a cluster the third is further from the pair. It was, however difficult to find any common patterns or preferences which could characterize each CRM's as a whole.

Discussion

In conclusion there are some preferences for combinatorial relationships among cooperatively interacting transcription factors but it is difficult to detect any absolutes. One of the obstacles in studying combinatorial relationships is the lack of data. To date, most of the evaluations of promoter finding tools and the studies of combinatorial relationships have used data from the Saccharomyces cerevisiae genome. This is due in some degree to fact that the yeast genome was one of the first eukaryotes genomes to be sequenced, is relatively small, and there is an abundance of micro-array data available for it. However, because the upstream regulatory region is relatively short (~600bp) compared to higher eukaryotes (~10k bp) yeast is probably not the best eukaryotic organism to screen for CRMs or cooperative interactions among known TFs (Wagner 1999). Given that these relationships are not easy to discern we need to first develop tools which will can generate and sift through the data to help infer relationships which can then be verified through experiment.

As we begin to find and characterize CRM's we will be able to extract more information from the DNA strand. For example, studies have shown that within CRM there are rules which govern the spacing and arrangement of the transcriptional regulatory elements. Papatsenko observed that the majority of high-affinity binding sites in fly enhancers are spaced at distances in 10bp increments on the same side of the DNA helix. This may indicate the positions of nucleosomes or other DNA-protein complexes in CRMs (Papatsenko 2002).

More rigorous statistical analysis needs to be done on the CRM's that we are aware of so that they can be characterized. Given the complexity of the combinational interaction of transcription factors it is unlikely that any single method will be able to accurately predict the locations of binding sites with a high degree of specificity and sensitivity. What is needed is a set of computational tools, which detect and characterize the signals on the DNA strand and then look for relationships. Once these relationships are better understood it may be possible to integrate important techniques into a suite of integrated tools or into one multidimensional tool.

First we need a more flexible way of expressing the binding sites. It was apparent from the data for the individual sites that a PWM is superior in expressing the affinity of a binding site as opposed to a consensus sequence. However, there are cases were the individual nucleotide positions exhibit preferences on one other which implies that they exert some influence on one another. HMM or neural nets can do a better job of expressing the affinity of the binding site for some cases. These richer expression models need to be integrated into a motif search tool to detect new sites for know transcription factors. We also need statistical tools which can capture the preferences for binding sites such as spacing, order, and orientation. All of the tools need to be integrated into a statistically rigorous system such as Bayesian framework so that both the probabilities and the heuristics can be integrated. Also, the tools for searching for potential new binding sites need to take into account the combinatorial affect of interacting transcription factors in a way similar to COBIND but be able to account for multiple transcription factors and flexible spacing between them.

The current set of tools used to locate putative binding sites appear to be able to detect the signal from binding sites of some of cooperatively operating transcription factors. The biggest problem is separating the true binding sites from the noise. The current set of algorithms are designed to detect similar patterns between sequences of characters. The problems is that many times the strongest signals are coming from patterns that do not appear to represent true binding sites. We need to be able to filter these signals based on the governing relationships of transcription factor interaction. We need to identify more potential CRMs and start to characterize them. In addition to looking for inter genetic regions with clusters of know binding sites, phyogeintic footprinting can be used to locate conserved regulatory control regions across genomes. Also Chromatin immunoprecipitation studies can yield another source.

References:

Baily, T. and Elkan, C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. IMSB Aug, 1994 pp. 28-36.

Berman, P. et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattering formation in the Drosophila genome. PNAS, vol. 99, no. 2, pp 757-762.

Brandon, C and Tooze, J. (1999) Introduction to Protein Structure, 2nd Edition Garland Pub.

Davidson, E. (2001) Genomic Regulatory Systems. Academic Press.

Fickett, J. and Hatzigeorgiou, A. (1997) Eukaryotic Promoter Recognition Cold Sprint Harbor Lab. Press 7:861:878

GuhaThakurta, D and Stormo, G. (2001) Identifying target sites of cooperatively binding factors. Bioinformatics, vol. 17, no. 7, pp. 608-621.

Hannenhalli, S. and Levy, S. (2002) Prediction transcription factory synergism. Nucleic Acids Research, vol, 20, No. 19, pp 4278-4284.

Liu, X. Brutlag, D. and Liu, J (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed Genes. Pacific Symposium on Biocomputing 6:127-138.
Liu, X. Brutlag, D. and Liu, J (2001) An algorithm for finging protein-DNA binding sites with applications to chromatin-immuoprecipitation microarray experiments. Nature Biotechnology. vol. 20 pp. 835-839.

Markstein, M. et al.(2002) Genome-wide analysis of clustered Dorsal sites identifies putative target genes in the Drosophila embryo. PNAS, vol. 99, no. 2, pp 763-768.

Matys, v. et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Research, vol, 31, No. 1, pp 374-378.

Ohler, U. and Niemann, H. (2000) Computational identification and analysis of eukaryotic promoters: new algorithms on the traces of gene regulation. Trends Genet. 2001 Feb;17 (2):56-60.

Pipel, Y. et al. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. Nature Genetics, vol. 29

Papatsenko (2002) http://homepages.nyu.edu/~dap5/statement/statement.html

Pedersen, A. et al. (1999) The biology of eukaryotic promoter prediction--a review. Comput Chem. 1999 Jun 15;23(3-4):191-207.

Stormo, G. (2000) DNA binding sites: representation and discovery. Bioinformatics, vol. 16, no. 1, pp. 16-23.

Wagner, A (1999) Genes regulated cooperatively by one ore more transcription factors and their identification in whole eukaryotic genomes Bioinformatics vol. 15 no. 10 pp. 776-784.

Wasserman, W. and Fickett, J., (1998) Identification of regulatory regions which confer muscle-specific gene expression. Journal Molecular Biology vol. 278, pp. 167-181.

Wasserman, W. and Fickett, J. (2000) Discovery and modeling of transcriptional regulatory regions.Curr Opin Biotechnol. Feb;11(1):19-24.

Yuh, C., Bolouri, H. and Davidson, E. (1998) Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene. Science 279, 1896

Appendix I CRM Plots

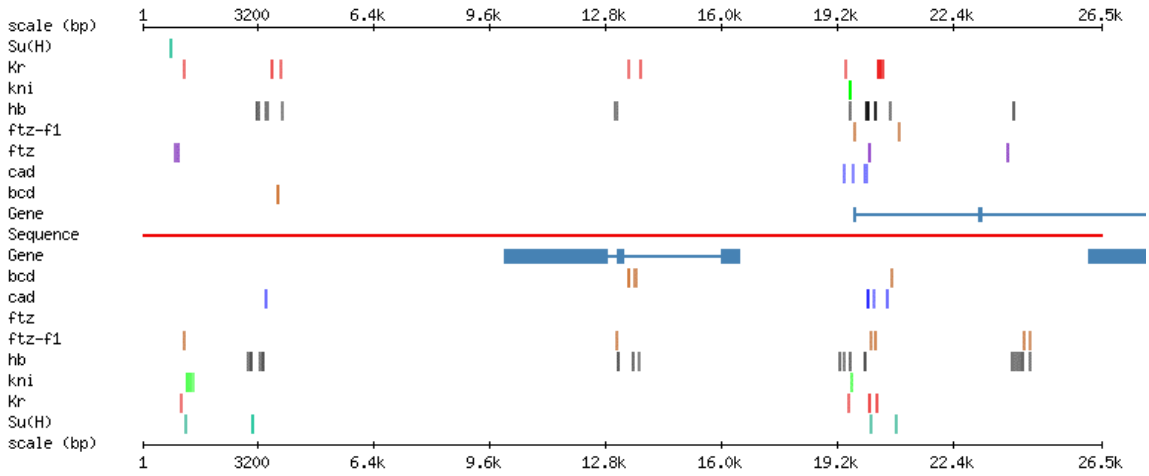


Fig 2. (contains genes hb, CG8112, jockey{}1268)

Cluster #1 [654, 1302] has 11 sites (**Kr**=2, **Su(H)**=3, **ftz**=2, **ftz-f1**=1, **kni**=3)
Cluster #2 [2785, 3765] has 14 sites (**Kr**=2, **Su(H)**=1, **bcd**=1, **cad**=1, **hb**=9)
Cluster #3 [12949, 13646] has 11 sites (**Kr**=2, **bcd**=3, **ftz-f1**=1, **hb**=5)
Cluster #4 [19174, 20809] has 34 sites (**Kr**=7, **Su(H)**=2, **bcd**=1, **cad**=7, **ftz**=1, **ftz-f1**=4, **hb**=10, **kni**=2)
Cluster #5 [23786, 24435] has 10 sites (**ftz**=1, **ftz-f1**=2, **hb**=7)

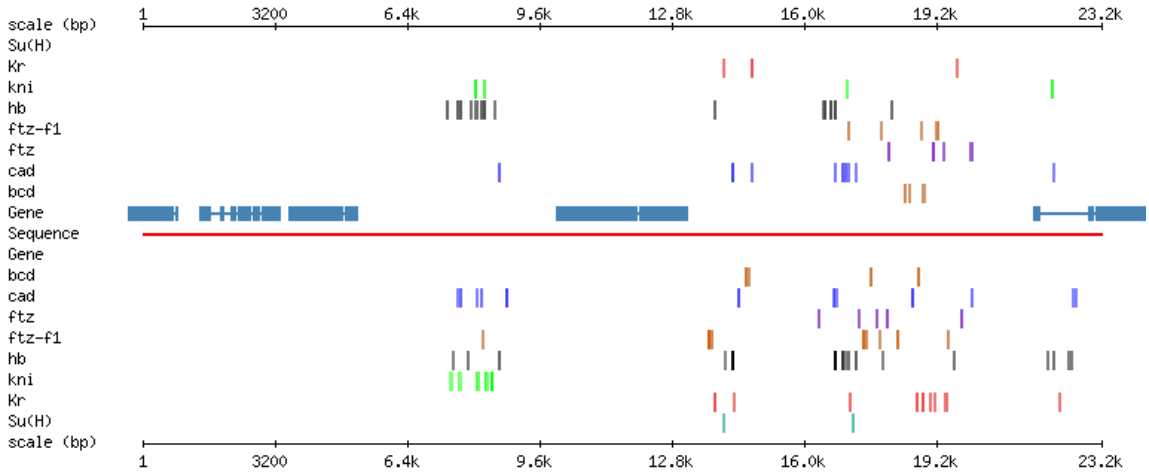


Fig. 3 (contains genes CG12133, Adam, CG12134, eve, TER94)

Cluster #1 [7271, 8730] has 30 sites (**cad**=6, **ftz-f1**=1, **hb**=12, **kni**=11)
Cluster #2 [13612, 14662] has 15 sites (**Kr**=4, **Su(H)**=1, **bcd**=2, **cad**=3, **ftz-f1**=2, **hb**=3)
Cluster #3 [16248, 18038] has 34 sites (**Kr**=1, **Su(H)**=1, **bcd**=1, **cad**=8, **ftz**=5, **ftz-f1**=5, **hb**=12, **kni**=1)
Cluster #4 [18165, 19981] has 25 sites (**Kr**=7, **bcd**=5, **cad**=2, **ftz**=5, **ftz-f1**=5, **hb**=1)
Cluster #5 [21791, 22503] has 10 sites (**Kr**=1, **cad**=4, **hb**=4, **kni**=1)

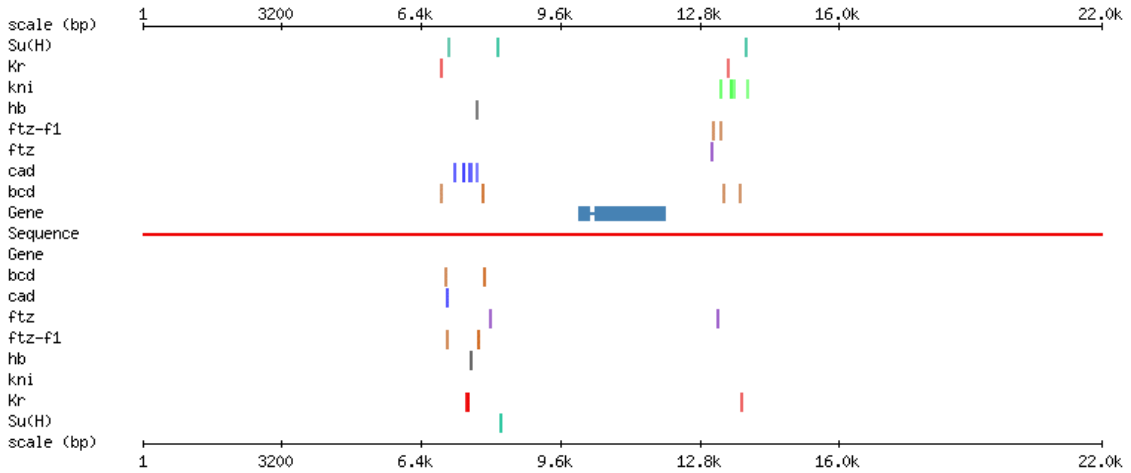


Fig 4 (contains genes tll)

Cluster #1 [2419, 3400] has 17 sites (**Kr**=2, **Su(H)**=2, **cad**=3, **ftz-f1**=4, **hb**=2, **kni**=4)
Cluster #2 [11780, 12667] has 19 sites (**Kr**=1, **Su(H)**=1, **bcd**=1, **cad**=4, **ftz-f1**=1, **hb**=9, **kni**=2)
Cluster #3 [13701, 15560] has 40 sites (**Kr**=8, **Su(H)**=3, **bcd**=4, **cad**=8, **ftz**=1, **ftz-f1**=1, **hb**=14, **kni**=1)
Cluster #4 [19191, 19861] has 12 sites (**Kr**=1, **Su(H)**=2, **cad**=2, **ftz**=4, **ftz-f1**=1, **kni**=2)
Cluster #5 [20269, 21024] has 12 sites (**cad**=3, **ftz**=1, **ftz-f1**=3, **hb**=2, **kni**=3)
Cluster #6 [22219, 22910] has 11 sites (**Su(H)**=4, **bcd**=1, **ftz**=1, **ftz-f1**=3, **hb**=1, **kni**=1)

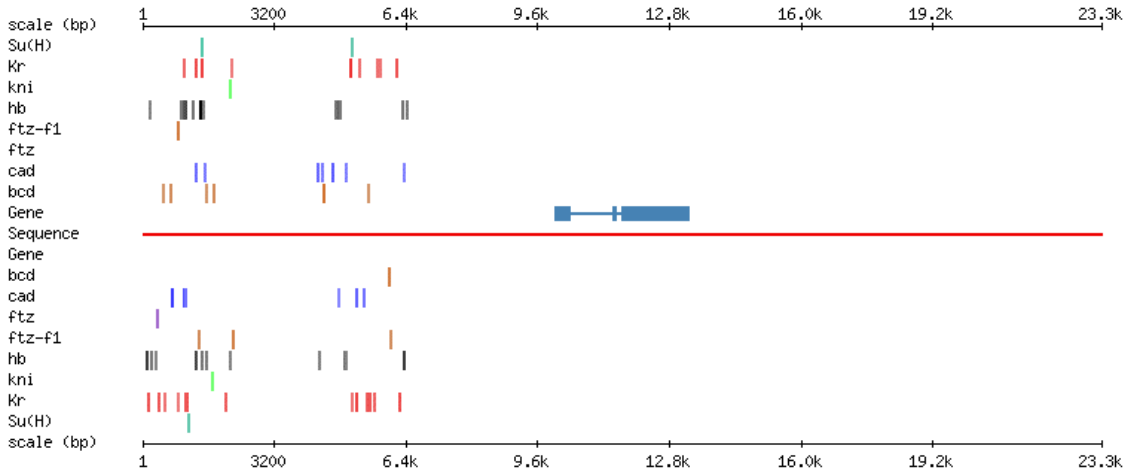


Fig. 5 (contains genes h)

Cluster #1 [-4, 2112] has 44 sites (**Kr**=11, **Su(H)**=2, **bcd**=4, **cad**=6, **ftz**=1, **ftz-f1**=3, **hb**=15, **kni**=2)

Cluster #2 [4178, 6344] has 37 sites (**Kr**=13, **Su(H)**=1, **bcd**=3, **cad**=8, **ftz-f1**=1, **hb**=11)