THE UNIVERSITY OF CHICAGO

EVOLUTIONARY ANALYSES OF TRANSCRIPTIONAL CONTROL SEQUENCES IN DROSOPHILA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY

CASEY M. BERGMAN

CHICAGO, ILLINOIS

AUGUST 2001

To my collaborators.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank Marc Halfon, Miki Fujioka, Jim Langeland, Misha Ludwig, Etsuko Moriyama and Charles Sackerson for providing unpublished sequence data and clones, without which I would not have been able to obtain the results presented here. I am also indebted to the many authors from around the world who supplied unpublished sequence data and binding site coordinates in Tables 1, 6, 7 and 8. I thank Oliver Hohman, Naomi Pierce and Jean Gladstone for technical assistance in various phases of this project. I thank Mark Biggin, Peter Bouman, Steve Dorus, Carrie Grimsley, Winship Herr and members of the labs of Harinder Singh and Marty Kreitman for helpful comments and feedback during the initial drafts of the ideas presented here. I thank the National Science Foundation for funding me and this research through these years. I also thank the members of my dissertation committee: Marty Kreitman, Manyuan Long, Wen-Hsiung Li, Misha Ludwig and Harinder Singh.

Specifically, I thank Eli Stahl for advice, discussion, and friendship during my years in the Kreitman Lab, as well as for upholding true interest and collegiality in the work of others. I also thank Mark Jensen for guidance and an open ear during the darker days of my third year, and for giving perspective on the role of graduate school in the big picture. Special thanks goes to Josep Comeron for many hours of animated discussion, debate, statistical advice and for being a model of how to consider difficult problems

thoroughly. I would like to particularly thank Harinder Singh for allowing me access to resources in his lab, and for engaging in collaboration which provided direction at a critical time in my graduate career; I also thank Eric Bertellino and Hyun-Jun Li for technical advice during this period. I also wish to thank Nipam Patel for giving me the opportunity to collaborate, teach and watch a true professional in action. I thank Marty Kreitman for giving me the space and computer resources to do my research, and for giving me the opportunity to learn about grant writing, the mechanics of running a lab, and the academic lifestyle. I thank Misha Ludwig for countless hours of advice (technical, personal, political), hypothesizing, story- and joke-telling, comraderie, strategizing, quarreling, mentorship, and for an unorthodox perspective on modern science. Without your intuition about what is good science and stimulation about how to survive in science, I would not have started nor completed the work presented here.

Lastly, I thank the friends and family whom I have come to know and love better over these five years for support and affirmation. Brian Bettencourt, Alex Bick, Jim Colliander, Justin Fay, Erin Himmelberger, Jordan Karubian, Jill Mann, Jordan McIntyre, Kate McGurn, Robin O'Keefe, Nate Pearson, Lea Pinsky, Penelope Spain, and Eli Stahl have helped keep me sane and each played important roles in my development here as a person. Finally, I thank my extended nuclear family for relaxing weekends and the opportunity for me to reconnect as an adult that being in the midwest has provided. Special thanks go to my wonderful sister and brother-in-law, Harley and Jud Beck. You have been a constant in Chicago and generously provided more than I could have ever asked for. I promise to splice that T. Rex for you as soon as I can.

INTRODUCTION

Complex eukaryotic genomes are composed of large amounts of DNA that does not code for protein. The presence of non-protein-coding DNA in introns and intergenic regions represents one of the major differences in genome structure between prokaryotes and eukaryotes, and is likely responsible in part for the major transition in complexity between these groups of organisms. Little is known about the structure or function of non-protein-coding DNA, although it is thought that some fraction contributes to the regulation of gene expression. Gene regulation has been speculated to play a major role in the evolution of animal and plant morphology, and thus finding the keys to understand non-protein-coding DNA evolution may be an important step in understanding the morphological diversity of life on earth. Towards this end, this thesis focuses on the development of methods for the molecular evolutionary analysis of non-protein-coding sequences.

In the first two chapters I develop approaches to analyze the static pattern of pairwise *cis*-regulatory conservation in Drosophila. In Chapter I, my focus is the structural and evolutionary properties of conserved non-coding sequences, also known as phylogenetic footprints. I study the density and length distribution of phylogenetic footprints, as well as the substitutional properties that operate within them. In Chapter II, I study the structural and evolutionary properties of sequences complementary to

phylogenetic footprints, which I define as spacer intervals. My premise in this chapter is that these oft-neglected sequences may be rich with information concerning the spacing of sequence-specific elements important in transcriptional regulation. In this chapter I present a hierarchical model of *cis*-regulatory structure to interpret the pattern of conservation observed in the non-coding regions of developmentally regulated genes.

The last two chapters describe approaches to study the dynamics of *cis*-regulatory molecular evolution in multi-species data sets of homologous sequences. In chapter III, I address the nature of analyzing *cis*-regulatory sequences under multiple alignment framework. Then, using multiple alignment and classical molecular evolutionary approaches, I test predictions in the pattern of enhancer divergence derived from a model of stabilizing selection. Finally in Chapter IV, I elaborate an alignment-free approach to analyze *cis*-regulatory sequences based on technology developed to predict binding sites *in silico*. In this chapter, I outline several methodological considerations and describe the pattern of *cis*-regulatory molecular evolution in a model eukaryotic enhancer revealed by this approach.

The molecular evolutionary analysis of non-coding and *cis*-regulatory DNA is only in its infancy. This is partly a historical and technological artifact, but largely because the problem is extremely difficult, as I have learned over the last five years. Thus, I beg of the reader to view this work as an early foray into an emerging field. Though many of the claims presented here may be qualitative, I hope that the insights and approaches gained from reading this thesis can contribute to further quantitative understanding of the evolution of genome evolution and gene regulation.

CHAPTER I

STRUCTURE AND EVOLUTION OF CONSERVED NON-CODING DNA IN DROSOPHILA

A. Introduction

The functional annotation of eukaryotic genomic sequences represents one of the greatest challenges in modern biology. Thus a diversity of approaches have emerged to identify genes and the *cis*-regulatory sequences controlling their expression. A promising class of methods for both gene and *cis*-regulatory prediction are based on comparative sequence analysis (Batzoglou, et al. 2000; Loots, et al. 2000). These approaches work because functionally constrained sequences are often conserved in evolution, much more so than non-functional sequences. Although why comparative sequence analysis enhances functional predictions is widely recognized, the link between molecular evolution and functional constraint is rarely made explicit. This link is most clearly formulated under the neutral theory of molecular evolution, which relates functional constraint with the rate and pattern of sequence evolution (Gillespie 1991; Kimura 1983). Thus, acknowledging this framework implies that constructing models of molecular evolution should complement the development of models that predict function from comparative genomic sequence data.

I am specifically interested in modeling the molecular evolution of *cis*-regulatory sequences controlling developmentally regulated gene expression in Drosophila. Drosophila is an excellent model system to explore the link between comparative and functional representations of *cis*-regulatory sequences. First, *Drosophila melanogaster* is a complex animal with a compact, completely sequenced genome with excellent physical and genetic maps (Adams, et al. 2000; Hoskins, et al. 2000). With such rapid progress in the completion of the *D. melanogaster* genome, sequencing of additional Drosophila genomes for comparative analysis is a distinct possibility. Second, this species has a rapid and cost-effective transgenic system that can be adapted for rescue, reporter, misexpression or knockout studies to test the function of predicted *cis*-regulatory sequences (Ashburner 1989; Rong 2000; Rorth, et al. 1998). Furthermore, recent developments in Drosophila transgenic vectors allow the possibility of reciprocal, cross-species analysis (Horn and Wimmer 2000). Third, the molecular genetics of many developmentally important *cis*-regulatory regions and pathways are well understood, providing the necessary functional context to test predictions based on comparative sequence analysis (Lawrence 1992). And finally, the phylogeny and evolutionary genetics of the genus Drosophila present a well-described range of divergence times to calibrate comparative and predictive models (Powell 1997). Thus in Drosophila, all of the tools are in place to critically test *cis*-regulatory structural and functional predictions based on comparative sequence data.

For these reasons, the use of comparative sequencing has become a common technique in the analysis of *cis*-regulatory structure/function in Drosophila.

Unfortunately, the utility of such data in predicting *cis*-regulatory function is limited, since little is known about the expected features of *cis*-regulatory molecular evolution (Stern 2000). A major difficulty impeding the quantitative analysis of *cis*-regulatory sequence evolution is the lack of a framework for the *a priori* statistical interpretation of non-coding DNA, akin to the genetic code for protein coding sequences. Empirically, however, the pattern of *cis*-regulatory molecular evolution in Drosophila and other species can be described qualitatively by highly conserved blocks of non-coding DNA, also known as phylogenetic footprints, separated by unalignable gaps (Tagle 1988). Phylogenetic footprints likely result from the sequence-specific constraints of DNA-protein interactions, although the correspondence between functionally characterized binding sites and conserved sequences is not exact (Dickinson 1991). Remarkably for many sequences assayed, the result of this mode of molecular evolution does not lead to drastic changes in the pattern of gene expression, as assayed by interspecific transgenic analysis (Tautz 2000). These seemingly paradoxical observations have led to a model in which stabilizing selection acts on the phenotype of gene expression, allowing a flux in the composition of the underlying *cis*-regulatory sequence (Carroll, et al. 2001; Ludwig, et al. 2000) (Chapter IV). Despite these insights into the mode of *cis*-regulatory molecular evolution, the comparative analysis of non-coding DNA has yet to be placed in a quantitative framework.

In this chapter I describe constraints acting on Drosophila non-coding sequences to gain insight into the expected features of *cis*-regulatory molecular evolution. Specifically, I use pairwise sequence analysis of non-coding DNA between *Drosophila*

melanogaster and *Drosophila virilis* to study molecular evolutionary constraints acting on over 100 kb of non-coding DNA sampled from 40 loci scattered throughout the Drosophila genome. These two species have been separated for approximately 40 million years, a divergence which is approximately equal to that between human and mouse, and more than sufficient to discern functional constraint in non-coding sequences (Blackman and Meselson 1986; Hartl and Lozovskaya 1994; Kwiatowski, et al. 1994; Russo, et al. 1995). I focus my attention primarily on non-coding regions that have been shown functionally to contain *cis*-regulatory activity in at least one of the two species. Both intergenic and intronic DNA is surveyed to analyze the effects of transcription on non-coding molecular evolutionary constraints. This work addresses the following questions about the structure and evolution of conserved non-coding DNA: 1) what is the fraction and density of phylogenetic footprinted sequences; 2) what is the length distribution of phylogenetic footprints; 3) what is the rate and pattern of point substitution in phylogenetic footprints; and 4) what is the rate and pattern of insertion/deletion (indel) substitution in phylogenetic footprints. I compare constraints in Drosophila non-coding DNA to that of other species, as well as to other types of sequences in the Drosophila genome. I evaluate my results using several tools for genome alignment to substantiate my findings and benchmark automated approaches. Finally, I suggest future prospects for the analysis of conserved non-coding DNA.

B. Materials and Methods

A clone of the *D. virilis* hairy 5' region contained in a P-element vector was obtained from J. Langeland (Langeland and Carroll 1993). The insert was digested using *Not* I and *Asp* 718 shotgun sequenced as previously described (Andolfatto, et al. 1999). A PCR product derived from *D. virilis* genomic DNA (Pasadena, CA strain 1052) was amplified using the Expand system (Roche Molecular) using primers designed from Genbank accession M87885 and the homologous region to Genbank accession S78746 (Langeland and Carroll 1993). The latter fragment is derived from the same parental plasmid as the 5' clone and has been submitted to Genbank under the accession number AF329639. The PCR primers for this reaction are vir_h_2322U24: 5'-CCATCTCGCGAGCGTGTCCAAAGC-3' and vir_h_6547L24: 5'-GTATTGGGCACCGCTGTCGTCTCC-3'. The reaction used 1.5 μl of genomic DNA from *D. virilis* strain 1048 Pasadena (Ashburner 1989). The cycling conditions on an MJ Research PTC-200 for this reaction are an initial denaturation at 92°C for 2 min, 10 rounds of denaturation at 92°C for 10 sec, annealing at 65°C for 30 sec, and extension at 68°C for 3 min 45 sec, followed by 19 rounds of the same conditions adding 20 sec per round to the extension time, terminated by a 68°C incubation for 7 min. This long distance PCR product has been submitted to Genbank under the accession number AF329640. The three *D. virilis hairy* 5'

fragments were joined into one contig for the final comparative analysis with *D. melanogaster*.

I attempted to generalize the pattern of sequence conservation derived from preliminary analysis of the *hairy* region by searching PubMed and Genbank for entries which contained *D. virilis* homologues of sequences with known or suspected *cis*-regulatory function in *D. melanogaster*. Where possible sequences were downloaded from Genbank; additional sequences were obtained by personal communication or transcribed from figures in the primary reference (Table 1). *D. melanogaster* sequences were obtained and oriented from the BDGP database *via* preliminary BLAST analysis with the *D. virilis* homologue. Sequences were edited so that the beginning and end of each region would correspond to conserved blocks.

My interest lies in the molecular evolutionary analysis of *cis*-regulatory sequences involved in transcriptional regulation, so I focused whenever possible on 5' and 3' non-transcribed, non-coding sequences with experimentally verified *cis*-regulatory transcriptional function. Transcriptional regulatory elements are however found in the introns of genes involved in many developmentally regulated genes in *Drosophila* (e.g. Ubx, eyeless, B-tubulin), so I also chose to include long introns (> 1 kb) or those with known or suspected regulatory function in the dataset. Coding and non-coding exons were excluded from my analysis to the limits of resolution of the annotated transcript structure in Genbank or the primary reference. In general, a given pair of sequences is terminated by either the transcription initiation site for 5' intergenic sequences, or the first block

Table 1. Names, cytological positions in *D. melanogaster*, and accession numbers (references) of non-coding sequences used in this study. An (*) indicates that more than one pair of contigs were surveyed for that locus.

Intergenic Region	Cytological position	*D. melanogaster*	*D. virilis*
achaete-scute *	1B1	AL024453 / AL023873	AF060507 / AF132809
Antennapedia	84B1	AC001655	M95827
bride of sevenless	96F9	AC019700	L08132
brown	59E2-3	AC005639	L37035
decapentaplegic *	22F1-2	AC004369 / AC019923	U95037 / X81976
dopadecarboxylase *	37C1	AC007176	X05065 / Johnson, et al. (1989)
D-mef2	46C1-2	AC014124 / AC014416	Cripps, et al. (1998) / Gajeweski, et al. (1997)
e74	74F1	AC019594	X59493
engrailed	48A2	AC020381	Kassis, et al. (1989)
frmf-amide	46C1-2	AC015179	AH000028
fused	17C5-7	AE003509	U20586
fushi tarazu	84B1	AE001573	Schier, et al. (1993)
glass	91A1	AC014473	U39746
hairy *	66D10	AC014797	AF329639 / AF329640 / Langeland, et al. (1993)
hunchback	85A6-11	U17742	S70575
knirps	77E2	AC020104	L36177
paramyosin	66D14	AE003554	AJ243069
prospero	86E3	AC013194 / AC012748	AF190404
runt	19E2	AE003570	Wolff, et al. (1999)
tailless	100B1	AC014779	AF019361
teashirt *	40A1-4	AC006467	McCormick, et al. (1995)
tinman *	93E1	AC020256	Lee, et al. (1997) / Xu, et al. (1998)
troponin T	12A1-4	AC014434	AJ002263
twist	59C3	AC005975	Pan, et al. (1994)
wingless	27F1-2	AC017528	AF046865
zerknullt	84A5	AC002512	L17339

Table 1. Continued.

Intronic Region	Cytological position	*D. melanogaster*	*D. virilis*
Antennapedia	84B1	AC001655 / AC020267	M95827 / M95828
bride of sevenless	96F9	AC019700	L08132
corkscrew	2D3	AC017610	U22356
decapentaplegic	22F1-2	AC019923	U63855
engrailed	48A2	AC020381	Kassis, et al. (1986)
glass	91A1	AC014473	U39746
Gpdh	26A7-9	AC017294	D10697
hunchback	85A6-11	U17742	X15395
knirps	77E2	AC020104	L36177
miniparamyosin	66D14	AE003554	AJ243070
myosin light chain	98A6	AC013071	L08053
paralytic	14D1-E1	AC014944	U26718
pdm-2	33F1-2	AC006470	U14723
prospero	86E3	AF190403	AF190405
rough	97D5	AC014838	M35372
sevenless	10A2	AC017584	M34544
single minded	87E1	AC020412	AF071932
Staufen	55B4-5	AC004336	AF225924
tinman	93E1	AC020256	Yin, et al (1997)
trithorax	88B1-2	AC013943	Z50038
vestigial	49D3-4	AC014851	Williams, et al. (1994)

downstream of reported polyadenylation site for 3' intergenic sequences. For intronic sequences, the first and last blocks are contained entirely within the intron.

I performed an manual analysis of homologous pairs of sequences in the data set using the Filtered DotPlot implementation in the MegAlign Program (DNAStar) (Maizel and Lenk 1981). This type of pairwise sequence analysis affords an interactive and exhaustive search for conservation that can be directly visualized. The parameters used in the initial search were percent match: 70%; minimum window: 1; window size: 10. I filtered top-scoring segments using a locus-to-locus heuristic threshold based on the shape of the tail of the distribution of segment scores. I then chose a colinear path of phylogenetic footprints to generate a set of local alignments spanning the entire region of homology (available on-line at http://www.genome.org). In an attempt to ensure that the substitutions observed in my data are truly within conserved sequences, I trimmed the phylogenetic footprints in my data set so that at least three nucleotides of identity flanked each local alignment to avoid spurious inclusion of nucleotides around the core of conservation. In general, off-main-diagonal high scoring segments were omitted because they were due to simple sequence repeats in which one species had a higher-scoring match elsewhere closer to the main diagonal. For all loci the counter-diagonal was also analyzed, but high scoring segments in the opposite orientation were also generally restricted to simple repeats. I justify this method based on previous estimates of what would be significant local alignments between *D. melanogaster* and *D. virilis*, in

conjunction with the established pattern of colinear conservation in *cis*-regulatory sequences (Hartl and Lozovskaya 1994; Jareborg, et al 1999).

During the course of this study I was provided an automated alignment tool called Lamark based on a dotplot algorithm developed by S. Shabalina and A. Kondarashov (NCBI, personal communication) which I used to help to evaluate my choice of phylogenetic footprints. The parameters of the Lamark search were 6 matches in a window size of 7, with each significant block requiring a segment score of 10 (i.e. 10 contiguous windows offset by one base of 6/7 matches). I imposed colinearity of the local alignments from the output of Lamark and compared the results of the automatic and manual analyses. Stimulated by similarities and differences among these two methods I chose to investigate the effects of alignment methods on my data set. Thus I also employed default parameters of the DNA block aligner (DBA), a finite state/hidden markov algorithm available in the Wise package to evaluate my choice of phylogenetic footprints (Jareborg, et al. 1999). Both Lamark and DBA generate a set of local alignments rather than a true global alignment. I also utilized default parameters of the DiAlign v2.1 (T = 0 with regions of maximum similarity denoted by five "*") alignment method (Morgenstern 2000). Finally, I submitted my dataset to the global VISTA genomic alignment tool using a window size of 10 and a percent identity of 70 (Dubchak, et al. 2000). I attempted to choose parameters for Lamark and VISTA that were comparable to those used in the filtered dotplot analysis.

Phylogenetic footprint sequences in both *D. melanogaster* and *D. virilis* sequences were parsed using helper applications written in the C programming language. For each phylogenetic footprint, identical and variant nucleotides were counted relative to the plus strand of the local transcription unit in *D. melanogaster*. Insertions causing phylogenetic footprints that are contiguous in one species to be separated in the other species were treated as insertion/deletion (indel) events . G-tests and sign tests were used to evaluate the difference in the percent of conserved sequences among transcription classes due to changes in the lengths of spacer intervals. Non-parametric tests were used to evaluate differences in length distributions of phylogenetic footprints and indel substitutions by transcriptional class and species since the data are not distributed normally. Goodness of fit to various expectations was evaluated using χ^2 tests. Tests were considered significant if the χ^2 statistic had probability less than (0.05/number of tests) to correct for multiple testing.

C. Results

The results of my survey for Drosophila non-coding regions exhibiting primary sequence conservation identified the 40 loci in Table 1. The loci are scattered among all five major chromosomal arms in the genome of *D. melanogaster*, and thus reflect a sample that is more or less random with respect to positional influences. Some loci have data for both intergenic and intronic regions, thus the total number of loci surveyed is smaller than the total number of regions. Several loci fitting my criteria for inclusion in the dataset had no conserved non-coding blocks distinguishable from background similarity using my methods, and were thus excluded from further analyses (*Adh*, *Amy*, *Gld*, *RP140*, *sisA*, *Sxl*, *Rh1-4*, *elav*, *su(s)*). For those loci that did show substantial conservation by my methods, initial attempts to find parameters of Wilbur-Lipman, Needleman-Wunsch, and pairwise BLAST algorithms that consistently gave comparable results were unsuccessful, frequently failing to recover easily identifiable dot-matrix local alignments. This is likely a result of the fact that alignment methods designed for coding sequences perform poorly when stretches of homology are short and gaps are frequent and variable, as is true for non-coding DNA [see (Jareborg, et al. 1999) for further discussion]. Using the combined output of several tools specifically developed for non-coding or genomic alignment, however, I was able to recapitulate results similar to those obtained by filtered dotplot.

Even by my relatively stringent criteria, I find that substantial amounts of intergenic and intronic non-coding DNA in Drosophila are subject to primary sequence constraint. When pooled across loci, 29,915 bp (20,501 bp intergenic, 9,414 bp intronic) are contained within conserved block sequences, out of 114,015 bp (79,874 bp intergenic, 34,141 bp intronic) and 138,831 bp (95,592 bp intergenic, 43,239 bp intronic) of DNA surveyed in *D. melanogaster* and *D. virilis*, respectively (Table 2). I note that each species has the same amount of DNA in the conserved block component of the data set by definition. Although the fraction of conserved block sequence ranges considerably across regions as a consequence of variation in the length of unalignable DNA, the average fraction of DNA under primary sequence constraint appears to differ little between intergenic and intronic DNA. Globally, there is a significantly higher percent of conserved non-coding DNA in the *D. melanogaster* genome relative to *D. virilis* for both intergenic ($G = 431.65$, 1 d.f., $p < 10^{-12}$) and intronic regions ($G = 347.24$, 1 d.f., $p < 10^{-12}$). These results are expected since *D. melanogaster* is known to have a smaller genome size than *D. virilis* (Powell 1997).

I studied the density and length distribution of ungapped conserved blocks which are important and yet unknown features of non-coding sequences that can be used to increase the sensitivity of genomic alignment and prediction tools. I observed 1225 (825 intergenic, 400 intronic) conserved non-coding blocks which I used to estimate these features empirically. The number of conserved blocks observed in intergenic and intronic DNA fit expected proportions based on the total amount of intergenic and intronic DNA surveyed in *D. melanogaster* ($\chi^2 = 4.28$, 1 d.f., $p < 0.039$) and *D. virilis* (χ^2

Table 2. Number of nucleotides surveyed, conserved, and percent sequence conservation of intergenic and intronic regions analyzed in this study. The species with the larger regional size is shown in bold.

	D. mel	*D. vir*		*D. mel*	*D. vir*
Intergenic	bp surveyed	bp surveyed	bp conserved	% conserved	% conserved
achaete-scute	2151	**5434**	559	26	10
Antennapedia	**3894**	3624	799	21	22
bride of sevenless	1458	**1730**	194	13	11
brown	269	**286**	113	42	40
decapentaplegic	12271	**15562**	4173	34	27
dopadecarboxylase	**590**	523	178	30	34
D-mef2	10117	**14083**	1407	14	10
e74	294	**331**	120	41	36
engrailed	2222	**2716**	738	33	27
frmf-amide	2072	**2623**	438	21	17
fused	**347**	275	179	52	65
fushi tarazu	385	**464**	135	35	29
glass	**1779**	1281	488	27	38
hairy	11942	**12673**	3549	30	28
hunchback	3423	**3624**	808	24	22
knirps	**1631**	1559	327	20	21
paramyosin	1444	**1803**	295	20	16
prospero	5525	**7714**	2053	37	27
runt	5995	**6170**	498	8	8
tailless	3646	**3906**	1095	30	28
teashirt	787	**846**	416	53	49
tinman	**1531**	1213	445	29	37
troponin T	**427**	389	64	15	16
twist	**1322**	1153	301	23	26
wingless	3979	**5213**	973	24	19
zerknullt	373	**397**	156	42	39

Table 2. Continued

Intronic	*D. mel* bp surveyed	*D. vir* bp surveyed	bp conserved	*D. mel* % conserved	*D. vir* % conserved
Antennapedia	2648	**2861**	881	33	31
bride of sevenless	1308	**1951**	399	31	20
corkscrew	521	**1102**	153	29	14
decapentaplegic	832	**929**	435	52	47
engrailed	1026	**1271**	416	41	33
glass	68	**111**	36	53	32
Gpdh	1448	**2346**	222	15	9
hunchback	2630	**3574**	892	34	25
knirps	614	**705**	273	44	39
miniparamyosin	1002	**1563**	251	25	16
myosin light chain	**1223**	1187	400	33	34
paralytic	**173**	171	80	46	47
pdm-2	**708**	642	215	30	33
prospero	4603	**5511**	1327	29	24
rough	2659	**4239**	627	24	15
sevenless	2266	**2606**	262	12	10
single minded	3119	**4777**	851	27	18
Staufen	1170	**1628**	349	30	21
tinman	309	**381**	158	51	41
trithorax	**5162**	4988	871	17	17
vestigial	652	**696**	316	48	45
Intergenic total	79874	95592	20501	26	21
Intronic total	34141	43239	9414	28	22
Grand total	114015	138831	29915	26	22

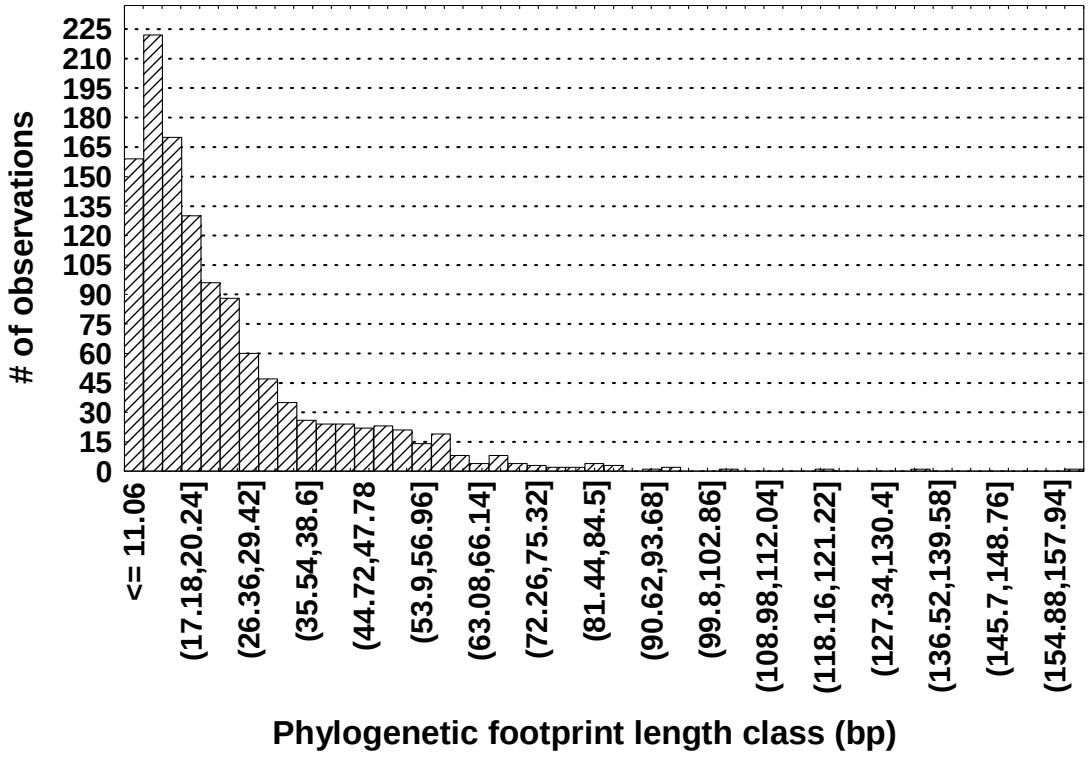


Figure 1. Length distribution of phylogenetic footprints in Drosophila. Since the distributions of intergenic and intronic phylogenetic footprint lengths do not differ significantly they are plotted together. The distribution is truncated at the lower extreme as a result of the criteria used to define phylogenetic footprints (see Materials and Methods).

= 1.30, 1 d.f., $p < 0.254$). Thus, from the total data set, I estimated the average density of ungapped conserved non-coding blocks to be 10.7 conserved blocks per kb in *D. melanogaster* and 8.9 conserved blocks per kb in *D. virilis*. Furthermore, no significant differences were observed between the length distribution of intergenic versus intronic conserved blocks (Kolmogorov-Smirnov test, $p > 0.10$). Since the density and length distribution of blocks does not appear to differ substantially among intergenic and intronic DNA the data were pooled into one frequency distribution (Fig. 1). The

distribution is highly skewed towards conserved blocks shorter than the mean (24.4 bp), with median and modal block lengths of 19 and 11 bp, respectively. Approximately 95% of conserved non-coding blocks are distributed between 10-71 bp, and only one intergenic block and three intronic blocks are greater than 100 bp in length. Using a linear transformation of the data (variate = length - 8), to correct for truncation of the distribution at the lower extreme, I could not reject that my data are obtained from a discrete approximation to a lognormal distribution (mean 2.376, variance 0.926; Kolmogorov-Smirnov test, $p > 0.10$), although other continuous (normal, gamma, chi-square) and discrete (binomial, Poisson, geometric) distributions could be rejected.

For each position in each conserved block, the nucleotide of both species was counted in order to derive a match-mismatch matrix for conserved non-coding blocks in Drosophila (Table 3). This information is critical for understanding the molecular evolutionary dynamics of conserved block substitutions, as well as the statistical evaluation of ungapped local alignments using extreme value (Karlin and Altschul 1990) or Bayesian theory (Zhu, et al. 1998). As expected, the majority of sites remain unchanged between *D. melanogaster* and *D. virilis*, and have a base composition ([AT] $\approx$ 60%) typical of non-coding regions in Drosophila (Moriyama and Hartl 1993). I observe 2,157 (1,503 intergenic, 654 intronic) nucleotide sites that differ within conserved block sequences. The number of substitutions in intergenic and intronic conserved blocks fit expected proportions ($\chi^2 = 1.32$, 1 d.f., $p < 0.250$), indicating that similar rates of substitution are observed in both types of conserved blocks. From the

Table 3. Nucleotide match-mismatch table for conserved non-coding blocks between *D. melanogaster* and *D. virilis*. Each cell represents the observed numbers of matched and mismatched bases for nucleotide positions aligned in intergenic, intronic and total conserved blocks. Rounded frequencies of observations are in parentheses. Asterices (*) represent significant differences between reciprocal substitutions at $p < 0.05/18 = 0.00278$.

D. melanogaster		*D. virilis* A	C	G	T
A	intergenic	5501 (0.268)	86 (0.004)	150 * (0.007)	82 (0.004)
	intronic	2738 (0.291)	39 (0.004)	68 (0.007)	31 (0.003)
	total	8239 (0.275)	125 (0.004)	218 * (0.007)	113 (0.004)
C	intergenic	108 (0.005)	3803 (0.186)	92 (0.004)	241 * (0.012)
	intronic	55 (0.006)	1676 (0.178)	45 (0.005)	72 (0.008)
	total	163 (0.005)	5479 (0.183)	137 (0.005)	313 * (0.010)
G	intergenic	212 * (0.010)	107 (0.005)	3937 (0.192)	78 (0.004)
	intronic	97 (0.010)	36 (0.004)	1613 (0.171)	51 (0.005)
	total	309 * (0.010)	143 (0.005)	5550 (0.186)	129 (0.004)
T	intergenic	96 (0.005)	150 * (0.007)	101 (0.005)	5755 (0.281)
	intronic	60 (0.006)	71 (0.008)	29 (0.003)	2733 (0.290)
	total	156 (0.005)	221 * (0. 007)	130 (0.004)	8488 (0.284)

total data set, I estimate that approximately 7.2% of the nucleotide sites in conserved non-coding blocks are substituted between *D. melanogaster* and *D. virilis*.

Not only the rate of substitution per bp, but the entire structure of the match-mismatch matrix is similar for intergenic and intronic conserved blocks (Table 3). The frequency of a given observation differs between intergenic and intronic blocks by 2.3% or less for identities and 0.9% or less for substitutions. But in contrast to similarity of substitution pattern across classes of DNA based on transcriptional state, there appears to be differences in the substitution pattern across species. This is especially apparent for reciprocal intergenic transitions, which show significant differences among observed counts of *mel* A: *vir* G (150) versus *mel* G: *vir* A (212) ($\chi^2 = 10.62$, 1 d.f., $p < 1.1 \times 10^{-3}$) and *mel* C: *vir* T (241) versus *mel* T: *vir* C (150) ($\chi^2 = 21.18$, 1 d.f., $p < 4.0 \times 10^{-6}$). Curiously, there is a trend across all reciprocal mismatch cells for substitutions to make *D. virilis* more AT rich for both intergenic ($\chi^2 = 20.51$, 1 d.f., $p < 6.0 \times 10^{-6}$) and intronic ($\chi^2 = 9.50$, 1 d.f., $p < 2.1 \times 10^{-3}$) conserved blocks. Unlike reciprocal substitutions, there are no significant differences observed in the pattern of complementary substitutions (i.e. *mel* A: *vir* G versus *mel* T: *vir* C) across species.

Though I can observe species differences in base usage at variable sites in conserved blocks, I am unable from pairwise sequence data to polarize the direction of substitutions. Considering this and the similarity of the substitution pattern in intergenic and intronic regions, I collapsed the data in Table 3 by summing the total numbers of reciprocal substitutions within the matrix (i.e. *mel* A: *vir* T + *mel* T: *vir* A = A↔T) to derive the relative rates of substitution among bases contained in conserved non-coding

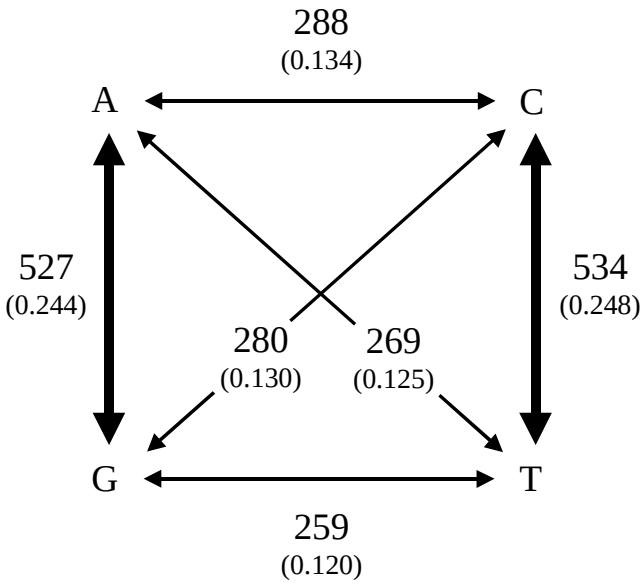


Figure 2. Relative rates of point substitution in phylogenetic footprints. Frequencies of observations are in parentheses below total observed counts. Reciprocal substitutions were pooled (i.e. *mel* A: *vir* T + *mel* T: *vir* A = A↔T) since divergence data is pairwise, and thus the polarity of the change is ambiguous. Bold arrows indicate transition substitutions.

blocks in Drosophila (Fig. 2). This view of the data shows that the relative rates of purine and pyrimidine transition substitutions are equal to each other, as are relative rates for each of the four different transversion substitutions. Relative rates of individual transitions are twofold greater than relative rates of individual transversions, indicating transition bias in these sequences. This twofold bias towards transitions leads to equal numbers of observed transition and transversion substitutions, since there are twice as many possible transversions as transitions. I caution that symmetry in the relative rate matrix should not be interpreted as stationarity in the substitution process in light of the lineage effects detected above.

In addition to the rate and pattern of point substitution, I can study the properties of conserved non-coding block indel substitution when two blocks are contiguous in one species but interrupted by an insertion in the other species. There are 96 observations of this kind that can be ascribed to *de novo* indel events, although these events cannot be distinguished as insertions or deletions from pairwise data alone. *D. melanogaster* has insertion of this kind relative to *D. virilis* 49 times (32 intergenic, 17 intronic), and *D. virilis* has an insertion relative to *D. melanogaster* 47 times (31 intergenic, 16 intronic). Moreover, the length distribution of inserted sequences does not differ significantly between *D. melanogaster* and *D. virilis* (Kolmogorov-Smirnov test, $p > 0.10$). From these observations, I can infer that the apparent rate and pattern of indel substitution in conserved non-coding blocks do not show lineage effects. The number of indel events in intergenic versus intronic blocks fits the expected proportions based on the total amount of conserved block DNA in each class ($\chi^2 = 0.376$, 1 d.f., $p < 0.539$). Additionally, no significant differences in the indel length distribution were observed between intergenic versus intronic regions (Kolmogorov-Smirnov test, $p > 0.10$). Thus, both the rate and pattern of indel substitution in Drosophila conserved non-coding blocks are similar in intergenic and intronic sequences.

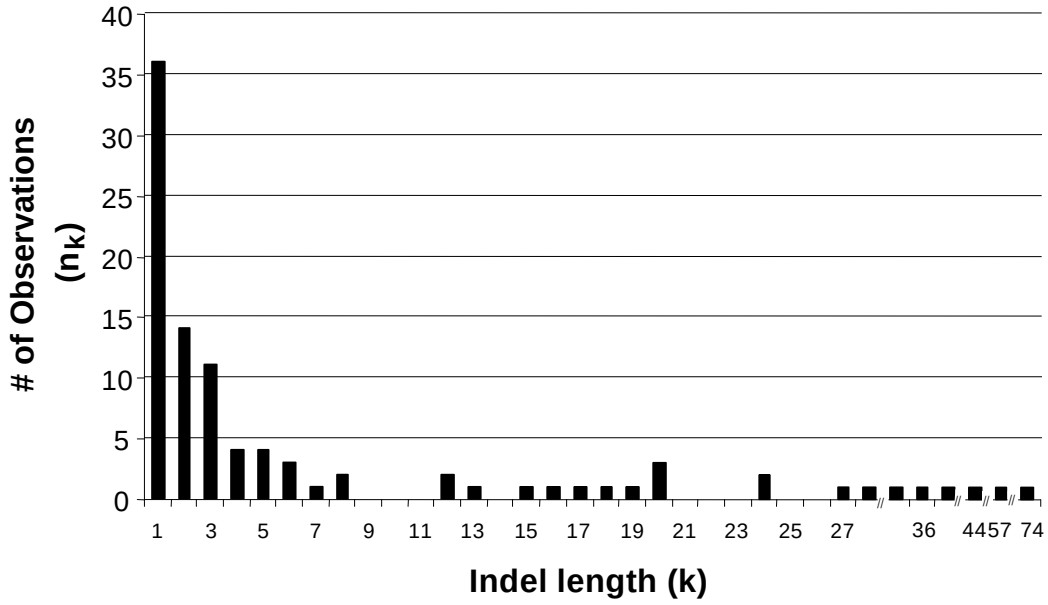


Figure 3. Length distribution of indel substitutions in phylogenetic footprints. Since the length distributions of indels for intergenic and intronic phylogenetic footprints do not differ significantly they are plotted together.

Since the indel length distribution does not differ significantly across species or transcriptional state, and since I cannot discriminate insertions from deletions, indel substitutions were pooled for further analysis. The total rate of indel substitution in conserved block DNA is estimated to be 0.32% indels per site, a rate more than twenty-fold less than point substitution. The length distribution of indel substitutions is skewed towards small (1-5 bp) sequences with a long tail of larger indels (Fig. 3). The mean and median indel lengths are 7.73 bp and 2 bp, respectively. The relationship between the natural log of indel size and the natural log of indel frequency for the combined data set is linear and highly correlated (Spearman's coefficient of rank correlation; $R = -0.740$, $p < 2.4 \times 10^{-5}$) (Fig. 4). This pattern has been found for an analysis of indel substitutions in

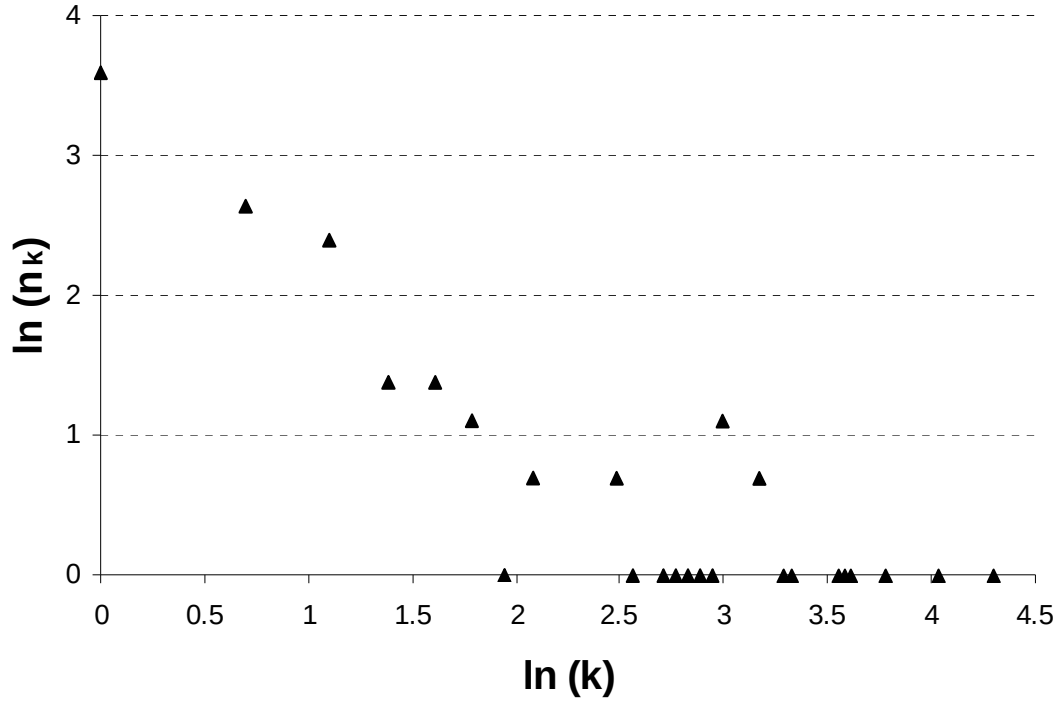


Figure 4. Correlation between length and frquency of indel substitutions. Both length (k) and frquency (n_k) axes have been transformed using into natural log scale. The Spearman coefficient of rank correlation between ln(n_k) and ln(k) is R = -0.740 ($p < 2.4x10^{-5}$).

a variety of data sets and suggests that the frequency distribution of indels follows a zeta distribution (Gu and Li 1995; Johnson and Kotz 1969). I attempted to fit my data to this function by using the maximum likelihood estimate, ρ, and approximate variance of the zeta distribution (pp. 241-243, (Johnson and Kotz 1969). I estimate ρ (± one standard deviation) for indels in Drosophila conserved non-coding blocks to be 0.6 ± 0.06, a parameter estimate similar to that found for indels in organellar and mammalian nuclear DNA [Note that 1+ ρ = **b** of (Gu and Li 1995)].

The results of all molecular evolutionary analyses are dependent on the underlying sequence alignments used, and my study is no exception. To substantiate the validity of my alignments and to benchmark tools designed for the alignment of non-

coding genomic regions, I performed a compatibility analysis of my results with several recently developed alignment platforms: DiAlign, DNA Block Aligner (DBA), VISTA and Lamark (Dubchak, et al. 2000; Jareborg, et al. 1999); S. Shabalina and A. Kondarashov, personal communication). I considered a block to be compatible if conserved block sequences defined by my method were also contained within an alignment block produced by an automated method. Since alignment is expected to differ among methods, particularly around block edges, I did not require exact correspondence of alignments. Of the 1225 blocks identified by my filtered dotplot analysis, 888 (72.4%), 1030 (84.0%), and 1074 (87.6%), 1122 (91.5%) blocks were identified by DBA, VISTA, Lamark and DiAlign respectively. 1209 (98.6%) conserved blocks identified by my method were identified by at least one of the four automated tools, 1158 (94.5%) by at least two, 1018 (83.0%) by at least three, and 745 (60.3%) were identified by all four.

D. Discussion

My finding that 22-26% of Drosophila non-coding sequences are highly constrained is similar to published estimates of the percent sequence conservation in non-coding regions of other complex eukaryotes. An analysis of conservation between *C. elegans* and *C. briggsae* using dynamic programming alignment methods reveals that for both intergenic and intronic regions the percent of non-coding sequence constraint is at least 18% for both species (Shabalina and Kondrashov 1999). A similar approach comparing 100 complete non-coding regions in mammals revealed that 33% and 27% of the mouse and human genomes, respectively, can be aligned (Shabalina, et al. 2001). The percent of conserved sequences in upstream and intron regions in the mouse genome has also been estimated to be 36% and 23%, respectively, using a HMM approach to non-coding alignment (Jareborg, et al. 1999). Similarly, a comparison of both intergenic and intronic muscle-specific regulatory regions between human and mouse using a Bayesian alignment procedure reveals that 19% of human sequences are highly conserved (Wasserman, et al. 2000). These two estimates may in fact be compatible given that the fraction of conserved sequences can vary across species because of changes in genome size, as is observed in Drosophila and mammals (Shabalina, et al. 2001). Despite differences in organismal complexity, sampling and alignment methods, these different analyses give remarkably similar values. In sum, preliminary estimates suggest that

~20-30% of nucleotide sites may be expected to be conserved in functionally constrained non-coding regions of eukaryotic genomes.

I found that average density and length distributions of conserved blocks are statistically indistinguishable between intergenic and intronic regions in my sample. The length distribution of conserved blocks pooled over intergenic and intronic regions reveals that the majority of non-coding sequence constraints act continuously over only short stretches of DNA: the median block length (19 bp) is small, and the mass of the distribution lies between 8 and 75 bp. I was not able reject the hypothesis that the pooled length distribution of conserved blocks is generated by a lognormal function using the best-fit parameter estimates from the data, although I could reject other continuous and discrete distributions (Fig. 1). Although the blocks of homology I detected are in general quite small, they are on average longer than the length of a single transcription factor binding site, and therefore likely correspond to the module level of *cis*-regulatory structure (Arnone and Davidson 1997). It is also interesting to note that the distribution of conserved block segment lengths is quite different from the distribution of exon lengths in Drosophila; conserved non-coding blocks are much shorter on average than the expected length of Drosophila exons [141.1 bp, (Deutsch and Long 1999)]. This observation should help the construction or parameterization of alignment and prediction algorithms that discriminate non-coding from coding DNA.

In addition to block length, the rate and pattern of point substitution also did not differ statistically between intergenic and intronic blocks. I estimate that ~ 7% of nucleotide sites are substituted in conserved non-coding blocks between *D.*

melanogaster and *D. virilis*, a value similar to one obtained for a sample of loci between mouse and human (Wasserman, et al. 2000). Substituted sites within highly constrained non-coding sequences showed two noteworthy features in the relative rates of point substitution: transition bias and lineage effects in base composition (Table 2, Fig. 2). I observed a 2:1 bias towards transition substitutions in my data, which is similar to estimates based on the divergence of coding sequences and non-functional dead-on-arrival retroelements (Moriyama and Powell 1997; Petrov and Hartl 1999). Polymorphisms in Drosophila non-coding and four-fold degenerate coding sites also show a 2:1 transition rate bias (Moriyama and Powell 1996). Thus a 2:1 bias towards transitions may be a general feature of molecular evolution throughout the Drosophila genome. The relative contributions of mutation, selection and other evolutionary forces to generating this pattern, however, remain unclear. Evidence for the second notable feature, lineage-specific changes in base composition, has also been observed for synonymous substitutions in Drosophila coding sequences (Moriyama and Hartl 1993; Rodríguez-Trelles, et al. 2000). Thus, changes in base composition among different lineages may ramify throughout both non-coding and coding regions of the Drosophila genome. Transition bias operating in conjunction with changes in base composition indicate that a nonstationary, nonhomogeneous model is necessary to adequately describe the subtleties of conserved non-coding block point substitution in Drosophila (Galtier and Gouy 1998).

In contrast with point substitution, rates of indel substitution are 20-fold less frequent in conserved non-coding blocks and appear to show no lineage effect. This

decrease in rate of indel substitution in non-coding regions may reflect differential mutation rates or more severe selective constraints on indel substitution relative to point substitution. Order-of-magnitude differences in the rates of point and indel substitution have been observed previously in comparative analyses of mammalian non-coding DNA (Saitou and Ueda 1994). Like point substitution, however, the rate and pattern of indel substitution is similar for intergenic and intronic sequences. The pooled length distribution of indels in conserved non-coding blocks is skewed towards short sequences (Fig. 3), as has been noted for *de novo* indels in inactive retroelement sequences in both *D. melanogaster* and *D. virilis* and in analyses of polymorphism and divergence in the *D. melanogaster* species group (Comeron and Kreitman 2000; Petrov and Hartl 1998). This skew is sufficient to produce a negative correlation between frequency and length for indel substitutions in Drosophila non-coding conserved blocks (Fig. 4). This result adds to a variety of different data sets which suggest that the zeta distribution can describe indel substitution, and that a logarithmic gap penalty is appropriate for the alignment of neighboring conserved non-coding blocks in Drosophila (Gu and Li 1995).

A major conclusion of my findings is that most features of non-coding DNA conservation are indistinguishable between intergenic and intronic regions. This is true for the average density and length distribution of conserved blocks, the rate and pattern of point substitution, as well as for the rate and pattern of indel substitution. I suggest that the similar properties of intergenic and intronic conserved blocks reflect similar mechanistic constraints operating on these sequences, and that transcription *per se* does not substantially influence major features of non-coding sequence evolution in

Drosophila. This finding has an important implication that can substantially reduce the complexity of large-scale comparative sequence analyses in Drosophila. Namely, my results would indicate that a single model for the identification of conserved non-coding DNA is sufficient for both intergenic and intronic compartments of the Drosophila genome.

Since there is no reading frame to constrain alignments, results based on pairwise sequence comparisons of non-coding DNA are critically dependent on alignment methods and parameters. I have attempted to be relatively conservative in my criteria for including sequences in the conserved block component of my data set. The main reason for adopting stringent inclusion criteria was to ensure that the majority of substituted sites analyzed are contained within sequences under purifying selection. In my opinion, it is first necessary to understand the pattern and relative rates of substitution in conserved non-coding blocks in order to statistically identify the boundaries of conserved blocks (Karlin and Altschul 1990). I suspect, however, that there are sequences which have functional constraint that are not conserved at the level of greater than 70% nucleotide identity, especially nucleotides flanking the edges of conserved blocks. For this reason, the fraction of constraint in non-coding regions with known or suspected *cis*-regulatory function based on my analysis is likely underestimated. Conversely, it is clear that some non-coding regions exhibit little if any primary sequence constraint, and thus genomic averages of non-coding constraint may be lower than what I report for functional regions. Moreover, my block definition would tend to bias the length distribution of conserved blocks towards lower values and

increase the number of functionally independent blocks relative to true values. Using stringent block criteria also leads to underestimating the total rates, but not the relative rates, of point and indel substitutions relative to true values. In the absence of a good substitution model, problems such as these can only be ameliorated empirically by multiple-species sequence comparisons.

In order to evaluate potential biases in my methods, I compared my results to those derived from four independent automated genomic alignment tools. Such a comparison is helpful for substantiating the results of my filtered dotplot approach, as well as calibrating automated tools for large scale non-coding sequence analyses in the future. From the combined output of DiAlign, DBA, VISTA and Lamark, over 98% of blocks identified by my method can be automatically identified, although only 60% are identified by all four methods. These results indicate that on the order of only 2% or less of blocks in my data set have no evidence of being conserved using an automated method, even though they contain matches that meet my criteria. I analyzed the 60% of blocks that were compatible across all methods (the compatible set) for properties of non-coding conservation reported above. As expected, the estimated fraction of DNA conserved in non-coding regions is lower in the compatible set (*D. melanogaster*: 19.4%, *D. virilis*: 16.0%). Also, the length distribution of conserved blocks differs between the total data set and compatible set, in that fewer short blocks are included in the compatible set (Kolmogorov-Smirnov test, $p < 0.001$). This difference is reflected in an increase in the location of the length distribution of the compatible set: the mean and median block lengths are 29.8 bp and 25 bp, respectively.

Importantly, however, my conclusions about the rate and pattern of point substitution is not dependent on alignment method. In the compatible set I observe that 7.0% of conserved block nucleotide positions are substituted, compared with 7.2% of sites in the total data set. Moreover, the overall structure of the match-mismatch matrix does not differ between the total and compatible data sets ($\chi^2 = 7.85$, 15 d.f., $p < 0.93$), nor in the distribution of indel sizes (Kolmogorov-Smirnov test, $p > 0.10$). Despite similarity in the pattern of indel substitution in the total and compatible data sets does, the estimated rate of indel substitution in the compatible data set (0.15%) is two-fold less than the total data set (0.32%). Thus, differences in alignment procedures may affect inferences about the density and length distribution of conserved blocks, however inferences concerning substitution properties are not substantially affected.

The results of my computability analysis also indicate that the majority of discordant blocks are missed uniquely by only one of the three methods. For instance, DBA systematically neglects many of the shortest blocks in my data set, which represent the mass of the conserved block distribution in my analysis (Fig. 1). DBA also has the tendency to insert bases and gaps in the output local alignments that are not present in the input sequences; this is likely a consequence of finite symbol emission probabilities of the HMM architecture underlying the alignment algorithm. VISTA, on the other hand, tends to omit small blocks that flank longer, strongly conserved blocks, or omit small blocks with that lie between two larger blocks that have strongly conserved spacing. These effects are likely due to the global alignment nature of the algorithm that might compromise small local alignments at the expense of aligning larger regions.

Since Lamark was designed for a hierarchical search strategy, a single parameter search such as ours lead to many local alignments off the main diagonal which necessitated imposing colinearity on the output to filter real from additional alignments. DiAlign identified the highest proportion of blocks in my data set with the minimum number of spurious alignments using a single set of parameters. However, DiAlign occasionally excludes flanking nucleotides from 'regions of similarity' (i.e. conserved blocks) which are clearly aligned in the output. Since each method has characteristic difficulties, I conclude that the use of multiple non-coding alignment tools is currently advisable to identify conserved sequences in non-coding regions.

Finally, my discussion of non-coding constraints must consider the methodological limitations imposed by pairwise sequence analysis. In addition to making the definition of block edges problematic, pairwise data limits understanding of the constraints on non-coding sequences in a number of important ways. As noted previously, pairwise data cannot distinguish the polarity of evolutionary changes, and thus the relative rates for reciprocal substitutions, or for insertions versus deletions, cannot be estimated individually. Moreover, pairwise data does not allow the observation of multiple substitutions at the same nucleotide position, for which I have made no corrections in my analyses. Multiple substitution in conserved non-coding blocks may not be a serious concern, however, since variant sites in the Drosophila *even-skipped* stripe two enhancer are generally substituted only once on the phylogeny (Ludwig, et al. 1998). However, not detecting multiple hits reveals a more general limitation of pairwise sequence analysis: the inability to assess heterogeneity in the rate

and pattern of substitution across sites. It is well established in coding sequences, that both the rate and pattern of substitution vary across sites and lineages (Yang 1996; Yang 1996). Furthermore, variation in the rate or pattern of substitution can influence estimation of other evolutionary parameters such as transition bias (Huelsenbeck and Nielsen 1999; Wakeley 1994). Thus, proper estimation of a substitution model for conserved non-coding blocks will require multiple species sequence comparisons to address these limitations of pairwise data.

My results provide insight into the mode of molecular evolution for a subset of highly conserved non-coding sequences. Future analyses of multiple species comparisons will be necessary to evaluate the generality of my results and construct more realistic substitution models for highly conserved non-coding DNA. Multiple species comparisons within Drosophila are especially important since my results suggest that the pattern of substitution in conserved non-coding DNA may fluctuate across lineages. For the same reason, it is also worth investigating conserved non-coding block substitution models in other eukaryotic (e.g. mammalian, rhabdid) lineages. Multiple species comparisons may also allow for null models of sequence evolution to be applied to data in hopes of potentially identifying Darwinian selection operating on non-coding sequences. Finally, comparing sequences from species more closely related than *D. melanogaster* and *D. virilis* will allow for a more thorough understanding of the evolutionary dynamics of weakly constrained non-coding sequences, since their rate of evolution should be higher than sequences studied here. Modeling the molecular evolution of non-coding sequences in general will require much additional research,

since these results and others show that the majority of non-coding nucleotides are not under strong primary sequence constraint.

Just as more widespread taxonomic sampling aids molecular evolutionary modeling, whole genome comparative analyses should offer a wealth of information relevant to modeling the individual units of non-coding structure and function. For instance, a genomic database of conserved non-coding blocks will be particularly useful for modeling structural properties of DNA-protein interactions like transcription factor binding site specificity. With such a resource, models of binding site specificity can be inferred from similarities intrinsic to a database of conserved non-coding blocks and be confirmed, rather than developed, experimentally. As proof of this principle, Wasserman et al (2000) were able to reconstruct the usage matrices for three myogenic transcription factors using conserved blocks from a sample of skeletal muscle regulatory regions. It is generally appreciated that non-coding conservation can be used to locate regulatory sequences, and that conservation can be used in combination with binding site prediction to identify potential upstream regulators of these sequences (Duret and Bucher 1997; Fickett and Wasserman 2000). Using binding site prediction in conjunction with conservation in this manner is 'top-down' (*sensu* (Bucher 1999)) and requires detailed *a priori* knowledge about which sequences a particular factor binds, a step which precludes efficient whole genome analysis. A 'bottom-up' approach like clustering conserved blocks should rapidly provide many models of transcription factor specificity that can be used to make functional predictions.

Finally, future comparative analyses will also help understanding of the higher order structural organization present in non-coding regions of eukaryotic genomes. Advances in this direction will require linking the pattern of non-coding primary sequence conservation to higher order functional units through specific structural models. For example, functional analysis of enhancer structure points to the importance of hierarchical spatial constraints operating between sequence specific elements (Ondek, et al. 1988). Under such a model, analyzing the spatial organization of conserved non-coding blocks will potentially aid the functional annotation of enhancer sequences. Other functional non-coding sequences will certainly benefit from the reciprocal development of higher-order models of structure and molecular evolution as well. A hallmark of success for models of non-coding molecular evolutionary in the future will be their ability to make useful functional predictions in comparative genomic data. Free from the constraints of the genetic code, the analysis of non-coding DNA presents a unique opportunity to develop and test models of molecular evolution which interface with those that predict structure and function.

CHAPTER II

NON-RANDOM SPATIAL CONSTRAINTS BETWEEN PHYLOGENETIC FOOTPRINTS IN THE DROSOPHILA GENOME

A. Introduction

Automatically identifying genes and the *cis*-regulatory sequences that control their expression is a central issue in genome research. Towards this end, many algorithms have been developed to predict protein-coding sequences, nearly all of which are based on constraints deduced from the genetic code (Badger and Olsen 1999; Batzoglou, et al. 2000; Claverie 1997). In the absence of a known code which maps sequence to function, two structural principles guide *cis*-regulatory prediction: (1) sequence-specific protein-DNA interactions, and (2) spatial constraints among individual protein-DNA interactions. Of these two principles, however, only the first is generally explicit in the major approaches to *cis*-regulatory prediction -- binding site prediction and phylogenetic footprinting (Duret and Bucher 1997; Fickett and Wasserman 2000). Inclusion of spatial constraints can improve the ability of binding site prediction to detect functional *cis*-regulatory elements (Crowley, et al. 1997; Wagner 1997), and will likely benefit approaches based on phylogenetic footprinting as well. Thus, in this chapter I investigate spatial constraints between phylogenetic footprints in Drosophila non-coding

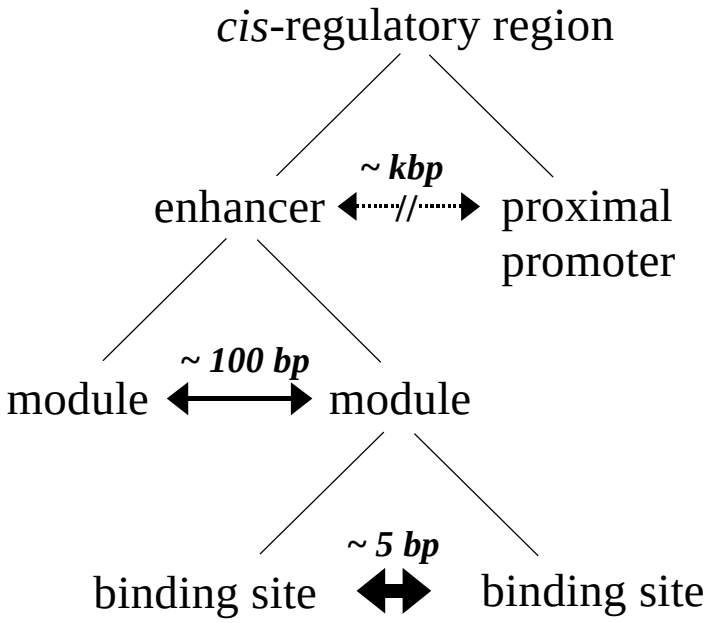


Figure 5. A hierarchical model of *cis*-regulatory spatial constraints (after Ondek, et al 1988). The fundamental level of structure is the binding site where sequence-specific DNA-protein interactions occur. Multiple clustered binding sites assemble to form modules, in which spatial interactions among transcription factors occur with strict spatial constraints over short distances. It is possible that single binding sites can act as modules as well (Fromental, et al 1988) . Enhancers are assembled from multiple modules which interact with relatively flexible spacing at intermediate distances. Enhancers can operate over long distances to stimulate transcription and are positionally independent by definition. Double-headed arrows indicate the length, and variation in length tolerated, among interactions at different levels of the hierarchy.

DNA, with the goal of developing a predictive model of *cis*-regulatory structure based on comparative genomic data.

Spatial constraints between sequence-specific *cis*-regulatory elements are deduced from functional analysis and define levels in the hierarchy of *cis*-regulatory structure (Fig. 5). For example, functional dissection of the SV40 enhancer, a prototypical eukaryotic *cis*-regulatory element, used mutations in spacing to define at least two different levels of enhancer structure (Fromental, et al. 1988; Ondek, et al.

1988). Under this model, enhancers are composed of subunits termed modules, which are in turn composed of smaller subunits -- transcription factor binding sites. The existence of distinct levels of enhancer structure is revealed by the differential functional effects of insertion mutations depending on where they occur in the hierarchy of enhancer structure. For example, an insertion of 10 bp between adjacent binding sites may entirely disrupt the function of an enhancer, whereas the same insertion may have only a quantitative effect on transcription when placed between neighboring modules. Further, this insertion should have no discernable effect when placed between two enhancers or between an enhancer and a promoter, since enhancers are defined operationally by their spatial independence (Khoury 1983). Thus functional analysis reveals that spatial constraints operate differentially on the distance, and variation in distance, between sequence-specific elements at different levels of the *cis*-regulatory hierarchy.

I propose an alternative approach to study spatial constraints between *cis*-regulatory elements based on phylogenetic footprinting. Specifically, I analyze properties of the distribution and divergence of spacer intervals, which are defined here as the sequences complementary to phylogenetic footprints in non-coding DNA (as defined in chapter I). Spacer intervals are typically ignored in non-coding sequence comparisons since they are by definition unalignable, even though functional and statistical considerations suggest they may be important predictors of *cis*-regulatory structure. Comparative genomic approaches to studying *cis*-regulatory structure are important because they potentially provide an abundant source of data on *cis*-regulatory

constraints *in vivo*. In addition, comparative genomic analyses offer independent means to test functionally-derived structural models. Moreover, to the extent structural models predict features of non-coding sequence evolution, they can be used for *cis*-regulatory annotation based on comparative genomic sequence data. Finally, comparative analysis of spacing between phylogenetic footprints is particularly important for the construction and parameterization of genomic alignment tools.

Here, I investigate constraints operating on individual spacer intervals between phylogenetic footprints from sequences with known or suspected *cis*-regulatory function in *D. melanogaster* and *D. virilis*. These two species split approximately 40 million years ago, a divergence time that is more than sufficient to discern functional constraint in non-coding sequences (Dickinson 1991; Kwiatowski, et al. 1994; Russo, et al. 1995). This comparison is also a relevant model system for mammalian comparative genomics since the amount of divergence between these species approximates that between human and mouse (Hartl and Lozovskaya 1994). First, using a data set of 40 non-coding regions, I estimate features of spacer intervals in the Drosophila genome. Second, I test and reject the null hypothesis that spacer interval lengths can be modeled randomly by an exponential distribution. Third, I analyze spacer interval divergence and show that spacer interval lengths are highly correlated across species despite genome size effects. Finally, I frame my results in the context of Drosophila genome evolution, interpret the data under a hierarchical model of *cis*-regulatory structure, and discuss implications for comparative genomic *cis*-regulatory prediction.

B. Materials and Methods

I surveyed PubMed and Genbank for entries which contained *D. virilis* homologues of non-coding sequences with known or suspected *cis*-regulatory function in *D. melanogaster* to analyze spatial constraints operating on these sequences. A detailed description of the resulting data set and methods of sequence analysis is presented elsewhere (see Chapter I, Materials and Methods). Briefly, I analyzed sequences using the Filtered DotPlot implementation in the MegAlign Program (DNAStar) (Maizel and Lenk 1981). The parameters used in the initial search were percent match: 70%; minimum window: 1; window size: 10 bp. I filtered top-scoring segments and then chose a colinear path of phylogenetic footprints spanning the entire region of homology. Spacer intervals sequences separating homologous phylogenetic footprints were extracted from both taxa and coded as paired data.

Biases and assumptions in my definition of phylogenetic footprints are discussed elsewhere but affect my current analyses (see Chapter I, Discussion). In particular, it is possible that my definition systematically omits small phylogenetic footprints, and that spacer intervals as I define them are actually sums of neighboring spacer intervals and smaller phylogenetic footprints. This effect may change the location spacer interval length distribution, but should not cause systematic deviation from the form of an exponential distribution. Summing neighboring intervals would not generate correlation across species but may affect the estimation of the correlation coefficient. More

importantly, I assume no regional heterogeneity in the distribution and divergence of spacer intervals across the genome, and treat all non-overlapping spacer intervals as independent. More data will be necessary to evaluate how these biases and assumptions affect analyses of spacer interval properties.

C. Results

My survey of evolutionary constraints operating on Drosophila non-coding DNA revealed many phylogenetic footprints of ~10-100 bp in size with ≥ 70% identity for all subsequences of 10 bp (Chapter I, Results). During the course of this study, I consistently observed a pattern of spatial constraints between phylogenetic footprints which would be predicted under a hierarchical model of enhancer structure (see below). Based on these initial observations, I attempted to quantify spatial constraints acting between phylogenetic footprints more closely using a data set of 40 homologous non-coding regions which display sequence-specific conservation totaling 114,015 bp and 138,831 bp in *D. melanogaster* and *D. virilis*, respectively. After excluding phylogenetic footprint sequences, 84,200 bp (59,373 bp intergenic; 24,727 bp intronic) of spacer interval DNA are observed in *D. melanogaster* and 108,916 (75,091 bp intergenic; 33,825 bp intronic) bp in *D. virilis* (Table II). This DNA is distributed among 1,164 spacer intervals (792 intergenic, 372 intronic), of which 1,068 are present in both species. 96 spacer intervals (63 intergenic, 33 intronic) are considered *de novo* insertions/deletion (indel) substitutions since they are unique to one species, and therefore not included in analyses presented here. I note that the majority of indel substitutions are < 10 bp, and occur at similar rates in intergenic and intronic DNA (Chapter I, Figs. 3 and 4).

The numbers of spacer intervals in intergenic and intronic DNA fit expected proportions based on total amounts DNA surveyed in both species (*D. melanogaster*: χ^2 = 1.64, 1 d.f., p < 0.200; *D. virilis*: χ^2 = 0.177, 1 d.f., p < 0.673). Furthermore, spacer interval lengths do not differ significantly between intergenic versus intronic sequences in either species (Mann-Whitney U-test: *D. melanogaster* p < 0.575; *D. virilis*, p < 0.576). The maximum spacer interval lengths observed in my sample are 1877 bp (1877 bp intergenic; 1464 bp intronic) in *D. melanogaster* and 1696 bp (1696 bp intergenic; 1276 bp intronic) in *D. virilis*. The mean spacer interval length is 78.5 bp (81.2 bp intergenic; 72.6 bp intronic) for *D. melanogaster* spacer intervals, and 101.6 bp (102.7 bp intergenic; 99.2 bp intronic) for *D. virilis* spacer intervals. The distribution of spacer interval lengths is highly skewed towards short lengths in both species (Fig. 6), thus I also report the median spacer interval length for *D. melanogaster* (38 bp total; 39 bp intergenic; 37 bp intronic) and D. *virilis* (47 bp total; 45 bp intergenic; 51 bp intronic). The fact that spacer intervals are on average longer in *D. virilis* relative to *D. melanogaster* is expected based on genome size differences between these species (Powell 1997). These results indicate that individual spacer intervals have similar properties in intergenic and intronic DNA but differ among species.

I used the sample of spacer intervals defined by comparative sequence analysis to investigate how a simple model of spacer interval lengths fits the data. Under the null hypothesis that spacing among phylogenetic footprints has no functional significance, spacer interval lengths should conform to a model in which phylogenetic footprints are distributed at random on a one-dimensional space. My null hypothesis takes the form of

the problem of randomly distributed breaks on the unit interval, which generates an expected set of segment lengths described by an exponential distribution (Feller 1966). Although DNA is discrete, I view the exponential model as a reasonable null model because it has well-known properties that can be exploited to capture spacer interval patterns. Moreover, attempts to identify spatially-structured clusters of predicted binding sites use the same null model (Wagner 1997; Wagner 1999; Wasserman and Fickett 1998). I estimate the rate parameter of the exponential distribution, λ, from the inverse of the mean spacer interval length. The observed distribution of spacer interval lengths can be tested for goodness-of-fit to the expected distribution under this parameter estimate (Sokal and Rohlf 1995).

Spacer lengths are distributed in a decreasing manner in both species, suggesting that the exponential distribution is a reasonable null hypothesis (Figs. 6 and 7). For the total data set I estimate λ to be 1.274×10^{-2} for *D. melanogaster* and 0.984×10^{-2} for *D. virilis*. However, using these parameter estimates and combining the data into 50 bins, I can reject the hypotheses that spacer lengths are randomly distributed according to an exponential distribution for both *D. melanogaster* ($\chi^2 = 199.8$, 8 d.f., $p < 1 \times 10^{-6}$) and *D. virilis* ($\chi^2 = 401.8$, 12 d.f., $p < 1 \times 10^{-6}$). Specifically, parameterizing an exponential distribution on average length predicts too few short intervals and too many intermediate intervals (Figs. 6 and 7). This deviation is exacerbated if I include spacer intervals defined by *de novo* indels, which are typically < 10 bp in length (Chapter I, Figs. 3 and 4). For both species, I can also reject the specific hypotheses that intergenic and intronic spacer intervals conform to exponential distributions (not shown). The data show

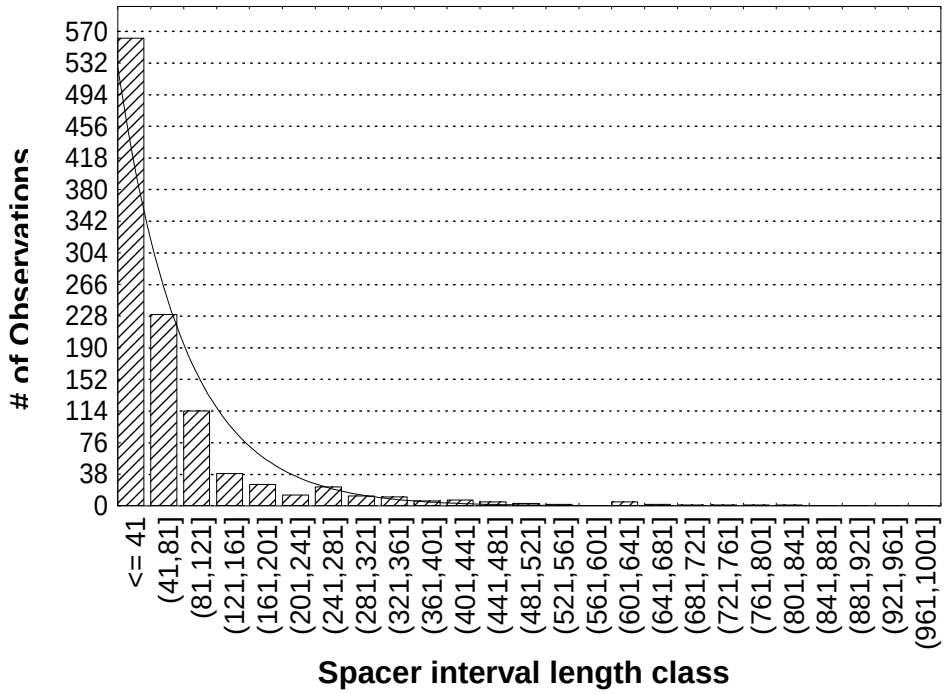


Figure 6. Length distribution of spacer intervals in *D. melanogaster*. Lengths of spacer intervals between phylogenetic footprints were plotted as histograms. The solid line represents the expectation under an exponential distribution using the parameter estimate of λ to be 1.274×10^{-2}.

substantially worse fit to other less likely distributions (geometric, Poisson, normal, beta; not shown). These results show that a simple model like the exponential distribution is not sufficient to describe spacer interval lengths, indicating that phylogenetic footprints are not randomly spaced in Drosophila non-coding DNA.

Further insight into spatial constraints between phylogenetic footprints can be gained by analyzing changes in spacer interval length across species (Fig. 8). Pairwise analysis shows that spacer intervals exhibit variation over the range of lengths, although the majority retain order-of-magnitude size relations across species. The Spearman coefficient of rank correlation for all interval lengths across species is 0.810 ($p < 1 \times 10^{-3}$),

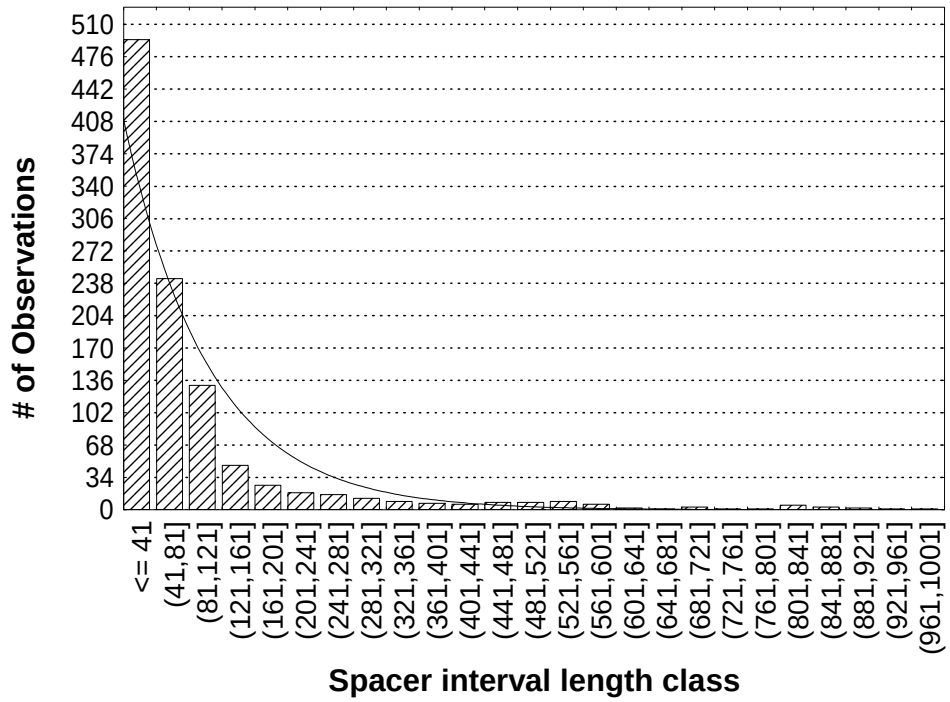


Figure 7. Length distribution of spacer intervals in *D. virilis*. Lengths of spacer intervals between phylogenetic footprints were plotted as histograms. The solid line represents the expectation under an exponential distribution using the parameter estimate of λ to be 0.984×10^{-2}.

demonstrating a significant correlation in spacer interval lengths across species. Variation in interval length across species does not differ for intergenic and intronic spacers, as reflected by individual coefficients of rank correlation for intergenic (0.813, $p < 1 \times 10^{-3}$) and intronic (0.803, $p < 1 \times 10^{-3}$) spacer intervals and similarity in the distribution of interval size ratios (Mann-Whitney U-test: $p < 0.183$). Log_{10} transformation of the data approximate normality sufficiently to estimate a regression coefficient using parametric techniques. I estimate the correlation coefficient for the total data set to be $r = 0.771$ ($p<10^{-3}$) and for intergenic and intronic data to be $r = 0.775$ ($p < 10^{-3}$) and $r = 0.762$ ($p < 10^{-3}$), respectively.

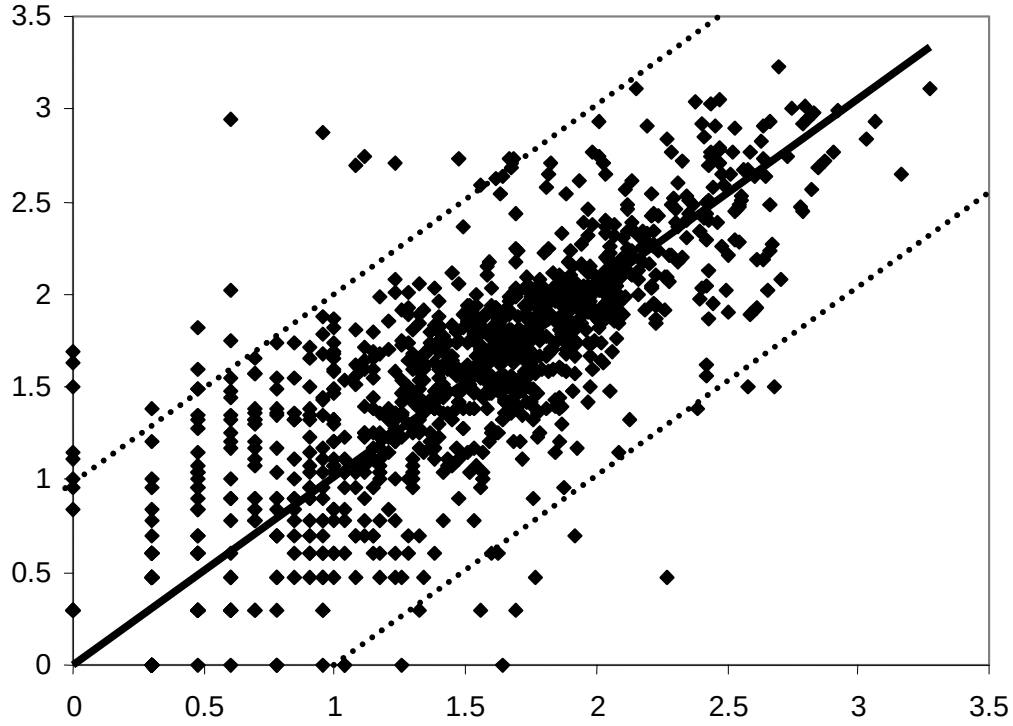


Figure 8. Correlation among homologous spacer interval lengths in Drosophila. Each point represents a pair of Log_{10} transformed lengths for a homologous spacer interval in *D. melanogaster* (x-axis) and *D. virilis* (y-axis). The solid line represents perfect spacer interval length conservation; the dashed lines represent order of magnitude size changes in spacer interval length across species. The Spearman coefficient of rank correlation for homologous spacer interval lengths across species is 0.810 ($p < 1x10^{-3}$)

As expected from euchromatin size differences among these species (Powell 1997), significantly more spacer intervals are longer in *D. virilis* (621) than *D. melanogaster* (416) (Sign test: $p < 1x10^{-6}$). The trend for *D. virilis* spacer intervals to be longer than *D. melanogaster* is observed for both intergenic (412 vs. 300; Sign test: $p < 3.2x10^{-5}$) and intronic (209 vs. 116; Sign test: $p < 1x10^{-6}$) sequences. The observation that the total length of introns in *D. virilis* tends to be longer than in *D. melanogaster* has been noted previously and extrapolated to support the claim that "mechanisms governing

the increase or decrease in size of DNA sequences operate more or less uniformly throughout the euchromatin" (Moriyama, et al. 1998). My results support this claim by providing direct evidence for a general increase in size in *D. virilis* intergenic regions relative to *D. melanogaster*, as was predicted from the pattern of intron size evolution between these species. Together, these results indicate that spacer interval length in both intergenic and intronic non-coding DNA is typically conserved across species, but that evolutionary changes tend to follow species-specific trends in genome size.

D. Discussion

The stable presence of spacer intervals over evolutionary time without sequence-specific constraint raises the question of why these sequences exist in the eukaryotic genome. One answer to this question is that many of these sequences are functionally constrained for their effects on the spacing of sequence-specific *cis*-regulatory elements. Consideration of two features of Drosophila genome evolution strengthens this position. First, *D. melanogaster* and *D. virilis* have been separated by approximately 40 mya, leading to substitutional saturation at alignable synonymous sites in coding sequences (Kwiatowski, et al. 1994; McVean and Vieira 2001; Russo, et al. 1995). Second, analysis of inactive retroelement divergence (Petrov and Hartl 1998; Petrov, et al. 1996) and indel polymorphism (Comeron and Kreitman 2000) shows that deletion bias operates in the Drosophila genome, a result of both inherent mutational biases towards deletions and possibly selection acting to reinforce this tendency. Thus given sufficient divergence time (i.e. saturation), deletion bias predicts that spacer intervals would ultimately disappear if unconstrained, and non-coding sequences would collapse to regions of sequence-specific conservation. Quantitative evaluation of this argument, however, is difficult in the absence of molecular evolutionary models of length substitution (Li 1997; Nei 1987). Nevertheless, the conserved presence and length of spacer intervals, despite substitutional saturation and deletion bias, argues strongly for spatial constraints in Drosophila non-coding DNA.

Analysis of spacer interval distribution and divergence also supports the claim that they may play a role in *cis*-regulatory function. First, spacer interval lengths do not conform to the simple null hypothesis of exponential distribution (Figs. 6 and 7). This result suggests that the spacing of phylogenetic footprints is not distributed at random in the Drosophila genome. Non-random distribution of phylogenetic footprints is expected if sequence-specific elements interact functionally with their neighbors. Second, there are evolutionary constraints on spacer interval length as shown by the strong correlation of spacer interval lengths across species (Fig. 8). Constraints on variation in spacer interval length may reflect conservation of functionally important spatial interactions (i.e. cooperativity, competition, quenching, etc.) over evolutionary time. Finally, the constraints observed likely act at the level of DNA (not RNA or protein), since the sequences I studied are non-coding, and since constraints are indistinguishable among non-transcribed and transcribed regions. In sum, my results argue for non-random spacing of phylogenetic footprints and conservation for spacer length with constraints operating at the level of DNA.

Though not exclusively, these characteristics of spacer intervals are expected under a hierarchical model of *cis*-regulatory structure (Fig. 5), which I propose may represent the form of spatial constraints operating between phylogenetic footprints. Under a hierarchical model of *cis*-regulatory structure, rejection of a simple distribution like the exponential is expected since spacer interval lengths should follow a composite distribution. Under such a model, spacing between elements at different levels of the hierarchy should have corresponding distributions of spacer interval sizes. Inter-

binding-site intervals should have the shortest and most frequent mode, with each higher level in the hierarchy of structure having longer and less frequent modes. Additionally, a positive correlation of spacer interval lengths across species is expected at all levels of the hierarchy because changes in spacer interval length should have functional consequences. The hierarchical model, however, predicts that changes in interval length across species should vary by functional class of the spacer interval: for example, inter-binding site spacing should be more conserved than inter-module or inter-enhancer spacing.

In support of this proposition, closer analysis of complex *cis*-regulatory regions in Drosophila directly reveals a hierarchical pattern of spacing between phylogenetic footprints, as illustrated by pairwise sequence comparison of the *dpp* 3' disk *cis*-regulatory region (Fig. 9). In such sequences phylogenetic footprints tend to be separated by multiple short or intermediate length spacer intervals, followed by a single longer spacer interval. This is the expected pattern of *cis*-regulatory spatial constraints if enhancers are composed of multiple phylogenetic footprints, as is true for two minimally defined enhancers in this region (Fig. 9) (Blackman, et al. 1991; Hepker, et al. 1999; Kopp, et al. 1999; Muller and Basler 2000). A hierarchy of constraints among phylogenetic footprints can also be inferred from the pattern of conservation revealed by electron microscopic analysis of interspecific heteroduplex DNA (Kassis 1985). Additionally, hierarchical spatial constraints can also be observed in complex *cis*-regulatory regions in mammals and can be used to make high quality predictions about the location of *cis*-regulatory sequences (Ishihara, et al. 2000). Thus this pattern may be

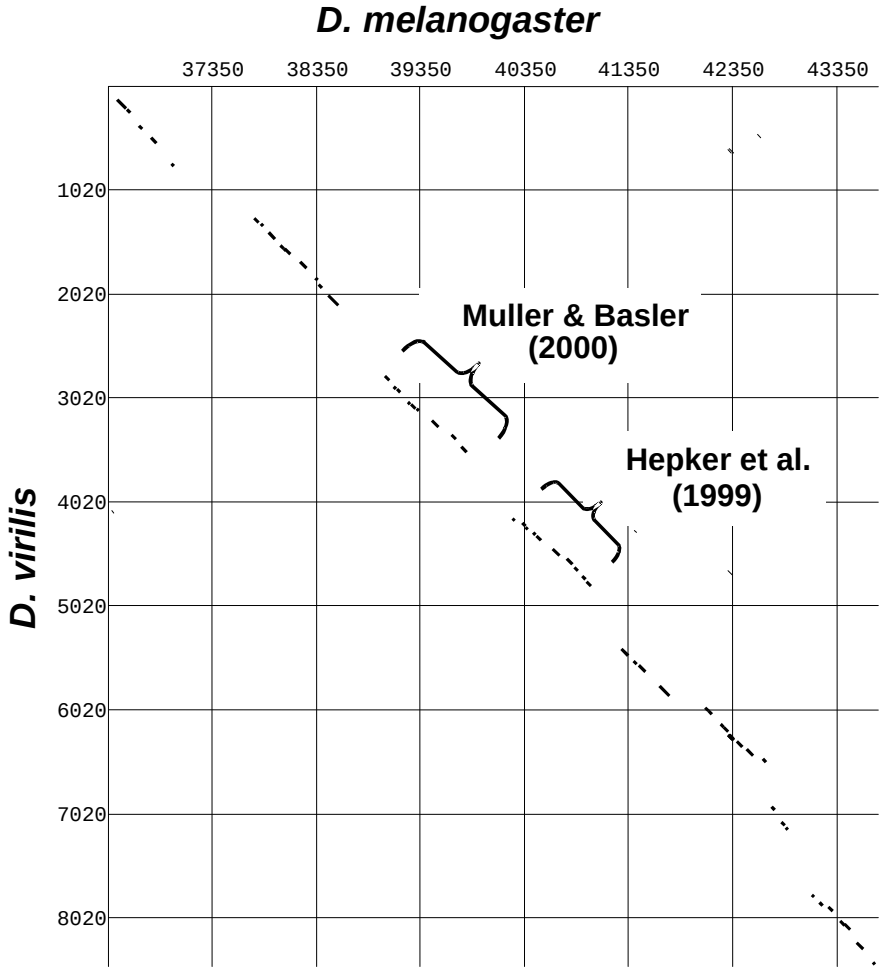


Figure 9. Conservation in the Drosophila *dpp* 3' disk *cis*-regulatory region. Homologous non-coding sequences from *D. melanogaster* (reverse complment of AC004369) and *D. virilis* (U95037) were analyzed comprehensively by dot matrix analysis. The top 80 of 115,232 high scoring segments were filtered for display. The boundaries of two functional characterized enhancers defined by minimization are bracketed (Hepker, *et al.*, 1999; Muller and Basler, 2000). Qualitative predictions of a hierarchical model of *cis*-regulatory constraints are observed in this region, assuming that phylogenetic footprints correspond to *cis*-regulatory modules.

a general feature of spatial constraints operating on *cis*-regulatory DNA and is unlikely

to be an artifact of my methods or specific to the Drosophila genome.

However if the hierarchical model is indeed true, some aspects of the data remain to be explained. For instance, the distribution of spacer interval lengths is not obviously multi-modal (Figs. 6 and 7). This may be result from sampling over different genomic contexts (e.g. recombinational, mutational) or different classes or architectures of *cis*-regulatory sequences (e.g. enhancers, promoters). Also in contrast to predictions of the hierarchical model, I observe that short spacers often display relatively extreme changes in length across species (Fig. 8). Under the hierarchical model, substitutions that change spacing among binding sites should have more severe functional effects than substitutions between modules or enhancers. This result is likely an artifact of my definition of phylogenetic footprints, rather than a true violation of the predictions of the hierarchical model (Fig 10). Highly conserved short spacers (say 5 bp) between two phylogenetic footprints may only accept one or two point substitutions between species, and thus the composite sequence is defined as one larger phylogenetic footprint (Fig. 10A). Thus, short spacers will only be identified when one species has an indel event in the spacer interval, giving rise to the observed short-in-one-species regime (Fig. 8, Fig. 10B). This interpretation may explain why, though expected to be the most frequent observation, highly conserved short spacer intervals appear to be underpopulated (Fig. 8). This ascertainment bias does not apply to longer spacer intervals, since two phylogenetic footprints are identified regardless of indel change (Fig. 10C).

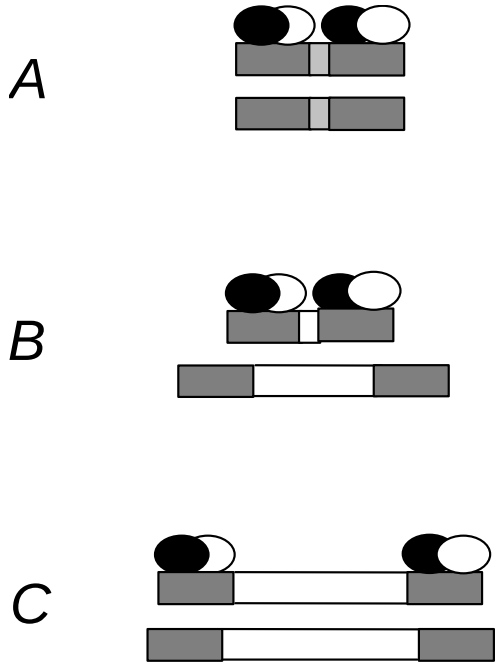


Figure 10. Evolutionary outcomes for spacer intervals of differing length. Ovals represent transcription factors, gray boxes represent phylogenetic footprints, and white boxes represent spacer intervals. (A) Highly conserved short spacer intervals can induce sequence-specific as well as spatial constraints (light gray boxes). This outcome leads to the definition of a one long phylogenetic footprint and thus short conserved spacer intervals are not observed. (B) Short spacer intervals are defined when they accept an indel event, causing an apparent lack of conservation on short intervals (Fig. 4). (C) Intermediate or long spacer intervals are defined regardless of length changes and thus represent an unbiased sample of length changes.

It is increasingly clear that non-coding sequence conservation provides a rapid and reliable means to the identification of *cis*-regulatory sequences in unannotated genomic DNA (Hardison 2000). However, full-genome *cis*-regulatory annotation based on comparative genomic data will require the development of quantitative models that predict *cis*-regulatory structure. To link such predictive models to *cis*-regulatory structure requires reconciling the basic units of *cis*-regulatory and comparative genomic data. I argue that individual phylogenetic footprints likely correspond to the module

level of *cis*-regulatory structure for three reasons. First, phylogenetic footprints are on average longer than length of single binding sites (Chapter I, Fig 1). Second, assuming my interpretation of pairwise conservation is true, inter-binding site intervals are rarely observed and thus the majority of spacer intervals separate modules. Finally, minimally-defined enhancers span several consecutive phylogenetic footprints and are bounded by longer flanking spacer intervals (Fig. 9). In sum, these facts suggest that phylogenetic footprints likely correspond to the module level of *cis*-regulatory organization.

If this link is accepted, then my results indicate that only certain features of *cis*-regulatory structure can be formulated and tested using comparative genomic data. For instance, it is unlikely that comparative genomic data will be useful for precisely defining the boundaries of individual transcription factor binding sites. Likewise, comparative genomic data may not contain easily extractable information concerning the spatial constraints operating between binding sites. These limitations are not severe if the goal of *cis*-regulatory prediction is to successfully identify the coordinates of full *cis*-regulatory sequences (i.e. enhancers) in genomic DNA, rather than constitutive elements (i.e. binding sites). Constructing and testing predictive models from the module level and above may be the relevant formulation of the problem, focussing *cis*-regulatory annotation on the appropriate scale and reducing the number of predictions to be tested. Further, many mechanistic issues regarding modules and module spacing remain open which can be addressed with comparative data (Gray 1994; Hewitt, et al. 1999; Small 1993).

The primary implication of my work is that spatial constraints between phylogenetic footprints may be key predictors of *cis*-regulatory structure. My approach shows that spatial information is readily extracted from comparative genomic data and that constraint exists in the distribution and divergence of spacer interval lengths. From this perspective, visualization tools should preserve spatial information in annotated alignments (Dubchak, et al. 2000; Gottgens, et al. 2001; Schwartz, et al. 2000). Moreover, comparative sequencing efforts should take advantage of natural variation in genome size across taxa to explore spatial constraints. Spatial constraints between phylogenetic footprints are complex; they are neither random nor strictly in accordance with predictions of a hierarchical model of enhancer structure. Future research will determine how spatial constraints between phylogenetic footprints make useful comparative genomic *cis*-regulatory predictions.

CHAPTER III.

PATTERNS OF ENHANCER DIVERGENCE UNDER STABILIZING SELECTION

A. Introduction

Despite increasing emphasis on the importance of *cis*-regulatory evolution, the majority of evidence for changes in *cis*-regulatory sequences is indirect (Carroll, et al. 2001). The direct analysis of *cis*-regulatory evolution has been approached experimentally in Drosophila providing examples of both *cis*-regulatory stasis and change. (Brady 1990; Bray and Hirsh 1986; Korochkin 1995; Krasney 1990; Langeland and Carroll 1993; Liu, et al. 1996; Ludwig, et al. 1998). Even when functional conservation of expression patterns are observed, the amount of sequence divergence at the DNA level can be considerable. Under a model of stabilizing selection on the pattern of gene expression, conserved sequences necessarily contribute to functional conservation, and divergent sequences contribute to function via compensatory changes (Ludwig, et al. 2000; Tautz 2000). This model describes constraints on the evolution of the phenotype of gene expression, however it does not make explicit predictions about the pattern of DNA sequence evolution. The lack of a quantitative framework for the analysis of *cis*-regulatory sequences stems from the fact that the rules governing the molecular evolution of these sequences are complex and cannot be derived from the genetic code.

The analytical framework to study *cis*-regulatory sequences therefore must be empirically derived, both through the structure-function analysis and elucidation of the pattern of *cis*-regulatory molecular evolution.

The empirical pattern of pairwise *cis*-regulatory conservation among distantly related species can be qualitatively described by blocks of highly conserved sequences (phylogenetic footprints) separated by unalignable spacer intervals (Blackman and Meselson 1986; Tagle 1988) (Chapters I and II). In general, approximately 20-30% of eukaryotic non-coding sequences in functionally constrained regions are contained in phylogenetic footprints, although some regions with *cis*-regulatory function show no detectable conservation (Shabalina and Kondrashov 1999; Wasserman, et al. 2000) (Chapter I). In Drosophila, phylogenetic footprints range from ~10-100 bp in length, have a base composition typical of non-coding DNA, and exhibit transition bias and lineage effects in base composition, although little is known about the structural and evolutionary properties of phylogenetic footprints in other taxa (Chapter I) The evolutionary properties of spacer intervals have been more difficult to characterize, since extensive point and indel substitution make such regions unalignable at evolutionary distances typically used to define phylogenetic footprints. Despite such high sequence turnover in spacer intervals, phylogenetic footprints tend to maintain colinearlity and relative spacing (Jareborg, et al. 1999) (Chapter II).

In addition to comparing only highly divergent taxa, most inference about the evolution of *cis*-regulatory sequence is derived from pairwise sequence analysis. These studies outline the structure of *cis*-regulatory conservation, but do not provide

opportunities to test the direction, tempo or mode of *cis*-regulatory evolution. To overcome these limitations, I present an attempt to use qualitative observations from divergent pairwise comparisons to make predictions about the pattern of *cis*-regulatory divergence among closely related sequences. The model system in which I evaluate these predictions is the 5' region of the Drosophila *even-skipped* gene (Kreitman and Ludwig 1996; Ludwig, et al. 1998). I have chosen this region for investigation since it has been extensively characterized for *cis*-regulatory function using *in vitro* mutagenesis and transgenic analysis (Fujioka 1996; Goto 1989; Harding 1989; Jiang 1991; Small 1992). Here I present contiguous sequence data for this intergenic region to address the pattern of *cis*-regulatory divergence for multiple enhancers in a complex genetic locus. Next I use a sample of *eve* stripe two enhancer sequences in Sophophora to determine the sampling framework appropriate for testing *cis*-regulatory divergence in multiple sequences related to a model system. Based on these considerations, I use stripe two sequences from the melanogaster species subgroup to test predictions of the null hypothesis of uniform substitution across the enhancer. In conclusion, I discuss the implications of my results for comparative approaches to *cis*-regulatory prediction.

B. Materials and Methods

I designed a long-distance PCR strategy to isolate homologous fragments of the *even-skipped* region in the genus Drosophila, based on conserved regions between species in the Drosophila and Sophophoran subgenera. The strategy employed both degenerate and species-specific primers to amplify three fragments approximately corresponding to bp coordinates (1) 131,623-133,994, (2) 126,076-131,937, and (3) 121,597-126,305 of the *D. melanogaster* scaffold sequence AE003831. Genomic DNA was prepared (protocol 48) from multiple individuals for each of the following lines: *D. erecta*, strain 1013; *D. pseudoobscura*, obtained from Soojin Yi (Yi and Charlesworth 2000); *D. virilis*, Pasadena strain 1048 (Ashburner 1989). Fragment 1 spans part of *eve* exon one, intron one, exon two, 3' UTR and a small amount of 3' untranscribed DNA, and was isolated using two universal primers, eve_1+ CCNCCNTCGCCAAATGGTA and eve_1- TCAAATCATAAAATGNTGCCACTT, which amplified both *D. pseudoobscura* and *D. virilis* as well as other species in the Drosophila and Sophophoran subgenera (not shown). In place of fragment 1 for *D. erecta*, I used a previously characterized sequence which covers the 5' non-coding sequences which are the focus of this study (Ludwig, et al. 1998). Amplified sequences were gel purified and shotgun sequenced as previously described (Andolfatto, et al. 1999).

Using the sequence of fragment 1, I designed species-specific primers to be used in conjunction with degenerate primers to isolate fragments 2 and 3. Fragment 2 covers

the autoregulatory region through the *eve* exon 1, and was isolated using the universal upper primer eve_2+ (U1+ of (Ludwig, et al. 1998)) ATTTGCTGCGGTNAGTCG and a species-specific lower primer designed from the sequence of fragment 1. The lower primer used to isolate fragment 2 from *D. erecta* was ere_eve_2-AACAATGGAACCCGAACCGT, *D. pseudoobscura* was pse_eve_2-GACGCTACCGAACCGCCTTCACACG, and from *D. virilis* was vir_eve_2-CAAAGTGTGGTTCCAGAATCGCCGCA. Fragment 3 extends from the 5' neighboring gene *Adam* to the autoregulatory region and was isolated using the universal upper primer eve_3+ TAAACAAGTGGGAGGGCGAGGACG in conjunction with species-specific lower primer designed from the sequence of fragment 2. The lower primer used to isolate fragment 3 from *D. erecta* was ere_eve_3-TAGTCCACCGCACTGACGAATCAC, *D. pseudoobscura* was pse_eve_3-CAGCAGTGTTTGTTGGCGAATCTCT, and from *D. virilis was* vir_eve_3-GTTTGTGCCGCTGTCTCCTGTTGCC. Long distance PCR was performed as described with the following annealing temperatures: fragment 1, 49°C; fragment 2, 56°C; fragment 3, 62°C (Chapter I, Materials and Methods).

eve stripe 2 fragments from *D. teissieri, D. orena, D. takahashii* and *D. ananassae* were provided by Misha Ludwig (personal communication). *D. melanogster, D. simulans, D. yakuba, D. erecta* and *D. pseudoobscura* sequences have been reported elsewhere (Ludwig, et al. 1998; Ludwig and Kreitman 1995). Coding sequences for *alcohol dehydrogenase (Adh)* and *yellow (y)* are described in (Jeffs, et al. 1994; Munte, et al. 2001; Takano-Shimizu 1999).

D. erecta, D. pseudoobscura, and *D. virilis* sequences were initially BLASTed against the BDGP database and oriented against the annotated *D. melanogaster* scaffold sequence AE003831. Pairs of sequences from the *Adam-eve* region were then submitted to the VISTA server for alignment and visualization using a window size of 50 bp, identity of 70% and minimal identity plotted: 50% (Dubchak, et al. 2000). DNA and protein sequences homologous to the *eve* coding region were extracted using the *D. melanogaster* transcript structure and aligned using ClustalW (Thompson, et al. 1994). The resulting DNA alignment was improved by comparison to the amino acid alignment using a utility provided by Anton Ne*Kr*etenko (personal communication). Divergence in the *eve* coding region was estimated using K-estimator with Kimura's two parameter model to correct for multiple hits (Comeron 1999). Multiple pairwise analysis of *eve* stripe two sequences was performed simultaneously by concatenating sequences in the 5'-3' direction, and comparing the concatenated sequence to itself using the Filtered DotPlot implementation in the MegAlign Program (DNAStar) (Maizel and Lenk 1981). The parameters used for this analysis were window size 10; identity: 70%; top 429 of 56066 segments. The plot is symmetrical around the top-scoring segment (the main diagonal of identity), so only the top half is shown.

Multiple alignment of the *eve* stripe two enhancer element in the melanogaster species subgroup was performed using default parameters of DiAlign 2.1, with minor manual improvement (Morgenstern 1999). These six species form a monophyletic clade composed of three pairs of sister taxa, and are thought to have the following phylogenetic relationship (tree 1) (((*D. melanogaster*, *D. simulans*), (*D. yakuba*, *D.*

teissieri)), (*D. erecta, D. orena*)) (Cariou 1987; Jeffs, et al. 1994; Lemeunier and Ashburner 1976; Lemeunier and Ashburner 1984). However since previous maximum likelihood analysis of the *Adh* coding sequence revealed that this topology was not significantly better than an alternative which placed the (erecta, orena) lineage closer to the (melanogaster, simulans) lineage, I evaluated evolutionary parameters under alternative topologies as well (Jeffs, et al. 1994). Tree 2 has the topology (((*D. melanogaster, D. simulans*), (*D. erecta, D. orena*)), (*D. yakuba, D. teissieri*)); tree 3 has the topology (((*D. erecta, D. orena*), (*D. yakuba, D. teissieri*)), (*D. melanogaster, D. simulans*)); tree 4 is the unresolved topology ((*D. melanogaster, D. simulans*), (*D. yakuba, D. teissieri*), (*D. erecta, D. orena*)).

I evaluated the nature of variation in rate of substitution across the enhancer using three methods. The first method is to test the fit of the distribution of the number of substitutions per site to the expected distribution (Poisson), under the null hypothesis of uniform distribution of substitutions in the enhancer. Second I used the method of Goss and Lewinton (1996) to the spatial distribution of substitutions across the enhancer under the null hypothesis of uniform distribution. Finally I performed maximum likelihood analysis of *eve* stripe two molecular evolution under various models of rate variation using PAML 3.0d (Yang 2000). For maximum likelihood analyses I also analyzed two coding sequences, *Adh* and *y*, to evaluate assumptions about the topology and variation in rates of evolution among lineage of these species. For these analyses, I assumed the HKY85 substitution model for both the *eve* stripe two enhancer and coding sequence data sets, since both coding and non-coding sequences in Drosophila exhibit

transition bias and have base compositions deviating from equal frequency (Hasegawa, et al. 1985; Moriyama and Hartl 1993; Moriyama and Powell 1997; Shields, et al. 1988) (Chapter I). Gaps in the *eve* stripe two data set were excluded from substitution analyses except where noted.

C. Results

Previous work has revealed the pattern of conservation in portions of the *eve* 5' region for pairs of divergent sequences (Fujioka 1996; Sackerson 1995) and multiple sequences of the minimal stripe 2 enhancer fragment (Kreitman and Ludwig 1996; Ludwig, et al. 1998). My results are the first to show the pattern of conservation for multiple species for the entire *eve* 5' region, allowing inferences about *cis*-regulatory evolution to be made above the level of a single enhancer (Fig. 11). Analysis of the *Adam-eve* region between *D. melanogaster* and *D. erecta* reveals heterogeneity in the percent conservation, with 50 bp windows ranging from nearly 100 percent to less than 50 percent identity (Fig. 11A). The degree of heterogeneity in conservation at this evolutionary distance ($K_s \sim 0.18$) (Munte, et al. 2001), only barely demarcates the boundaries of the three functionally identified enhancers in this region. In contrast, comparison of the same region between *D. melanogaster* and *D. pseudoobscura* reveals a pattern of conservation that identifies three distinct clusters of phylogenetic footprints: bp coordinates 3750-5000, 5500-6500 and 7500-9750, respectively (Fig. 11B). At this level of divergence ($K_s \sim 0.9$), discrete clusters of phylogenetic footprints correspond with, but extend beyond, the sequences shown to be minimally necessary for enhancer function (Fujioka 1996; Goto 1989; Harding 1989; Jiang 1991; Small 1992). Although peaks of conservation decrease, this pattern does not change substantially if more

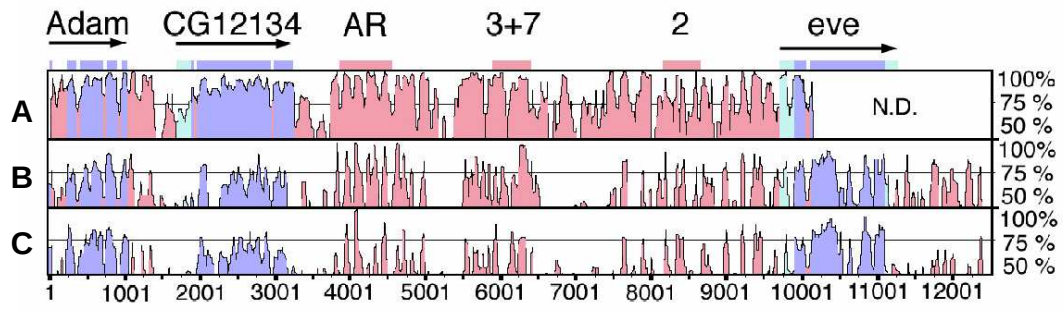


Figure 11. Genomic organization and sequence conservation in the eve 5' region. The location and direction of predicted transcripts are indicated by arrows; blue boxes represent exons and green boxes represent UTRs in the predicted gene structure. Peaks of conservation are colored to correspond to exonic (blue), UTR (green) or non-coding (pink) DNA. Conservation between *D. melanogaster* and (a) *D. erecta*, (b) *D. pseudoobscura*, and (c) *D. virilis* are projected onto bp 121,597 to 133,994 of the *D. melanogaster* scaffold sequence AE003831. Locations of minimal enhancers are indicated by red boxes. Exon 2 of the *D. erecta eve* coding sequence was not determined.

divergent species are compared (*D. melanogaster* vs. *D. virilis*, Fig 11C), suggesting decay to a core set of highly conserved phylogenetic footprints.

Interestingly, the pattern of conservation is similar for all three enhancers, suggesting that the evolutionary process may be similar across complex enhancer sequences in this region, even though they differ in structure and function. This similarity likely results from the fact that divergent sequences (spacer intervals) may have a more uniform rate of evolution, analogous to similarities in Ks across genes (Li and Graur 1991). Contiguous data above the level a single enhancer allows observations not previously possible. For instance, long nonconserved regions separate enhancers from each other and genes, a pattern consistent with the finding that enhancer autonomy in the *eve* 5' region is mediated by spacing (Small 1993). Long unconstrained regions appear to be defined at similar evolutionary distances as short unconstrained regions

within minimally defined enhancers, consistent with the notion that regions unconstrained for primary sequence may evolve at similar rates.

To understand the taxonomic framework which leads to the definition of phylogenetic footprints and spacer intervals, I analyzed minimal *eve* stripe two enhancer sequences for a range of divergence times in the Sophophoran subgenus. For this analysis, I initially attempted to multiply align stripe two sequences for *D. melanogster*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. takahashii*, *D. ananassae* and *D. pseudoobscura*. However, using conventional multiple alignment tools, were unable to find match-mismatch and gap penalties that reliably aligned the set of phylogenetic footprints identified by dot-matrix analysis among individual pairs of sequences. This led us to an alternative representation of the data set in which all pairs of sequences are simultaneously compared by dot-matrix analysis in one graph (Fig. 12). In this plot, each cell represents an individual dot-matrix comparison between homologous *eve* stripe two sequences, with species arranged from left to right and top to bottom in order of increasing divergence from *D. melanogaster*. Each cell can also be thought of as a reconstruction of the stripe two enhancer for the most recent common ancestor of the two species compared or as a reconstruction of an internal node in the phylogeny. This view of the data makes it possible to see why automatic multiple alignment of these sequences is problematic and for which sequences automatic multiple alignment might be possible.

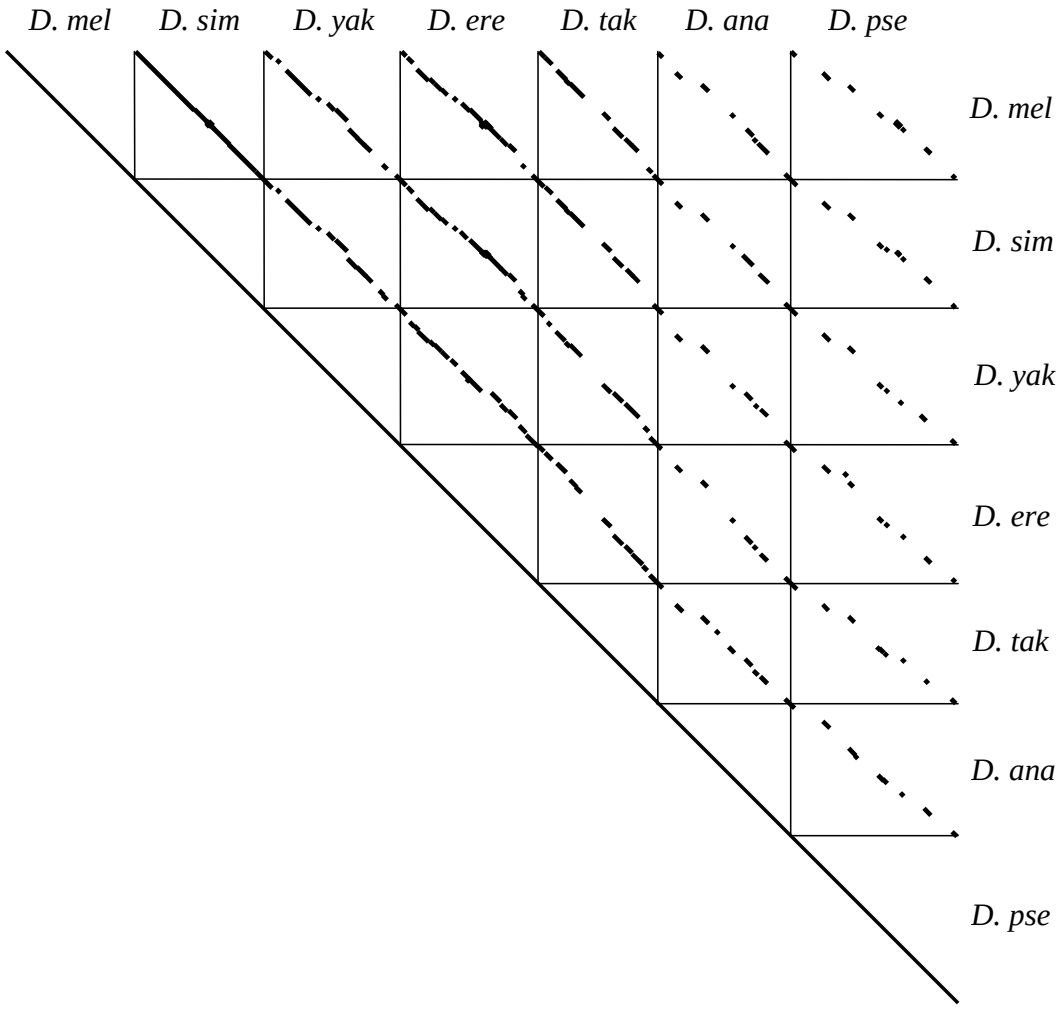


Figure 12. Pairwise comparisons among *eve* stripe two sequences in Sophophora. Homologous fragements were concatenated and analyzed by filtered dotplot (window size 10; identity: 70%; top 429 of 56066 segments). Dotplots for individual interspecific comparisons are within rectangles whose edge lengths correspond to sequence length. Species are arranged along the top of the figure in increasing divergence from *D. melanogaster*.

Focussing on the top row of the figure shows the dynamics of enhancer evolution as a function of increasing divergence time from *D. melanogaster*. This representation shows that the entire *D. melanogaster* stripe two enhancer is alignable with *D. simulans*, with the exception of minor differences due to insertion/deletion substitution (Ludwig

and Kreitman 1995). Likewise, the majority of the *eve* stripe two sequence can be aligned for species in the melanogaster species subgroup, although both point and indel substitution reduce the fraction of alignable sites. Comparisons of the *D. melanogaster* sequence with species outside the melanogaster subgroup but within the melanogaster species group (e.g. *D. ananassae*) shows that a substantial fraction of the stripe two enhancer has diverged beyond the point of reliable alignment. This observation is true for comparisons between the melanogaster and obscura species groups, and for more distant comparisons between the Sophophora and Drosophila subgenera (not shown, but see Figure 11). Interestingly, the pattern of conservation between *D. melanogaster* and *D. ananassae* differs little from that between *D. melanogaster* and *D. pseudoobscura*, except for changes in spacing of alignable sites, suggesting that the decay in the fraction of alignable sites in enhancer sequences may not be a linear function of time.

To address the nature of rate variation among lineages and across sites in a sample of sequences it is necessary to have a reliable multiple alignment. Thus I chose to analyze divergence among six representative species of the melanogaster subgroup which I was able to automatically align (Fig. 14): *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. teissieri*, *D. erecta*, and *D. orena*. My goal is to test predictions of sequence evolution under a model of stabilizing selection with two classes of functional constraint. To begin, I investigated the effects of the underlying topology and variation in the rate of evolution among lineages for the *eve* stripe two enhancer and two reference coding sequences, *Adh* and *y*. Next I tested predictions of the number of substitutions per site

Figure 13. Alignment of *eve* Stripe Two Sequences in the Melanogaster Subgroup. Boxes enclose sequences which have been shown to bind transcription factors *in vitro* (see (Ludwig, et al. 1998) for details). Lines over the *D. melanogaster* sequence represent phylogenetic footprints between these six species and *D. pseudoobscura*.

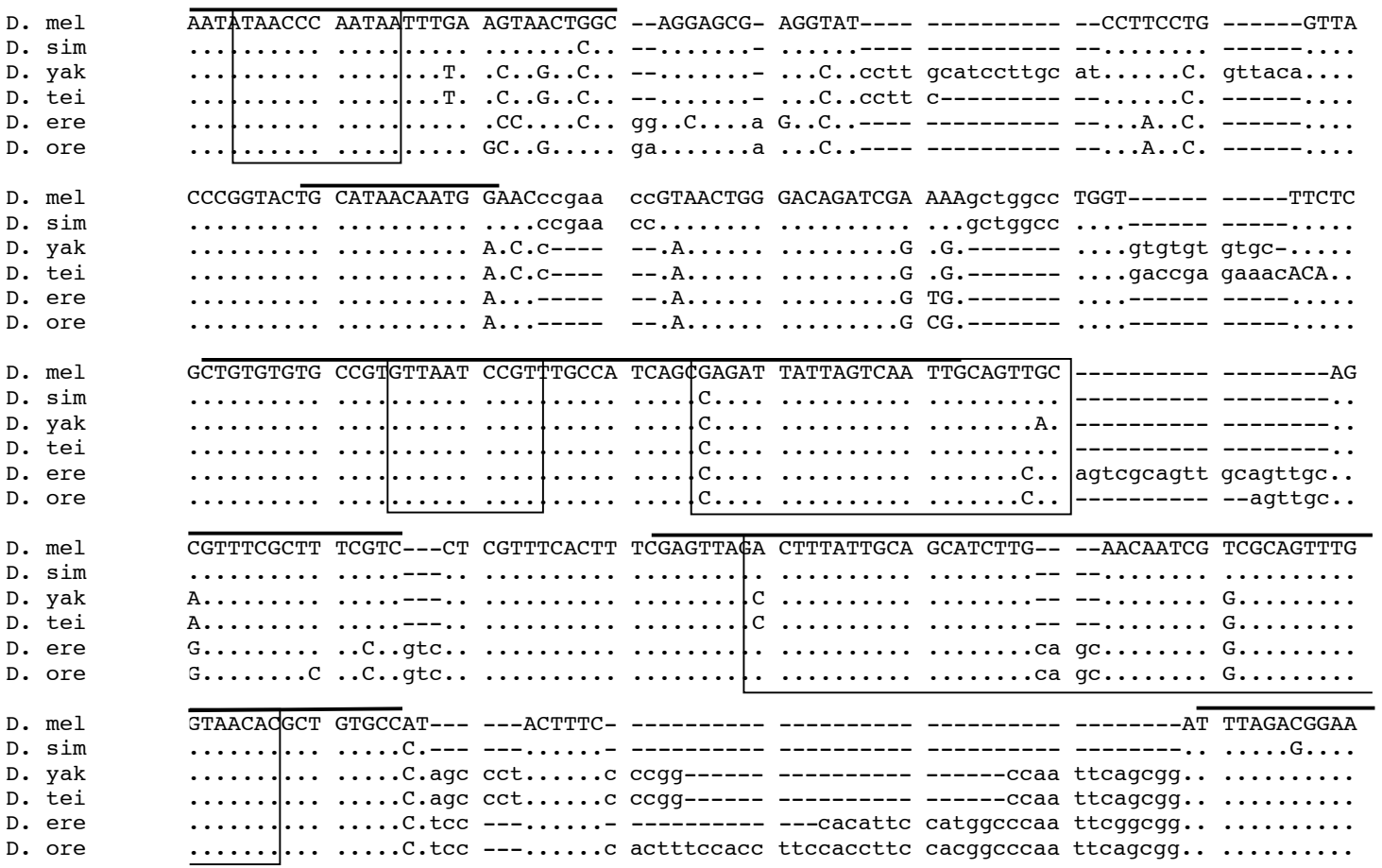


Figure 13. Continued.

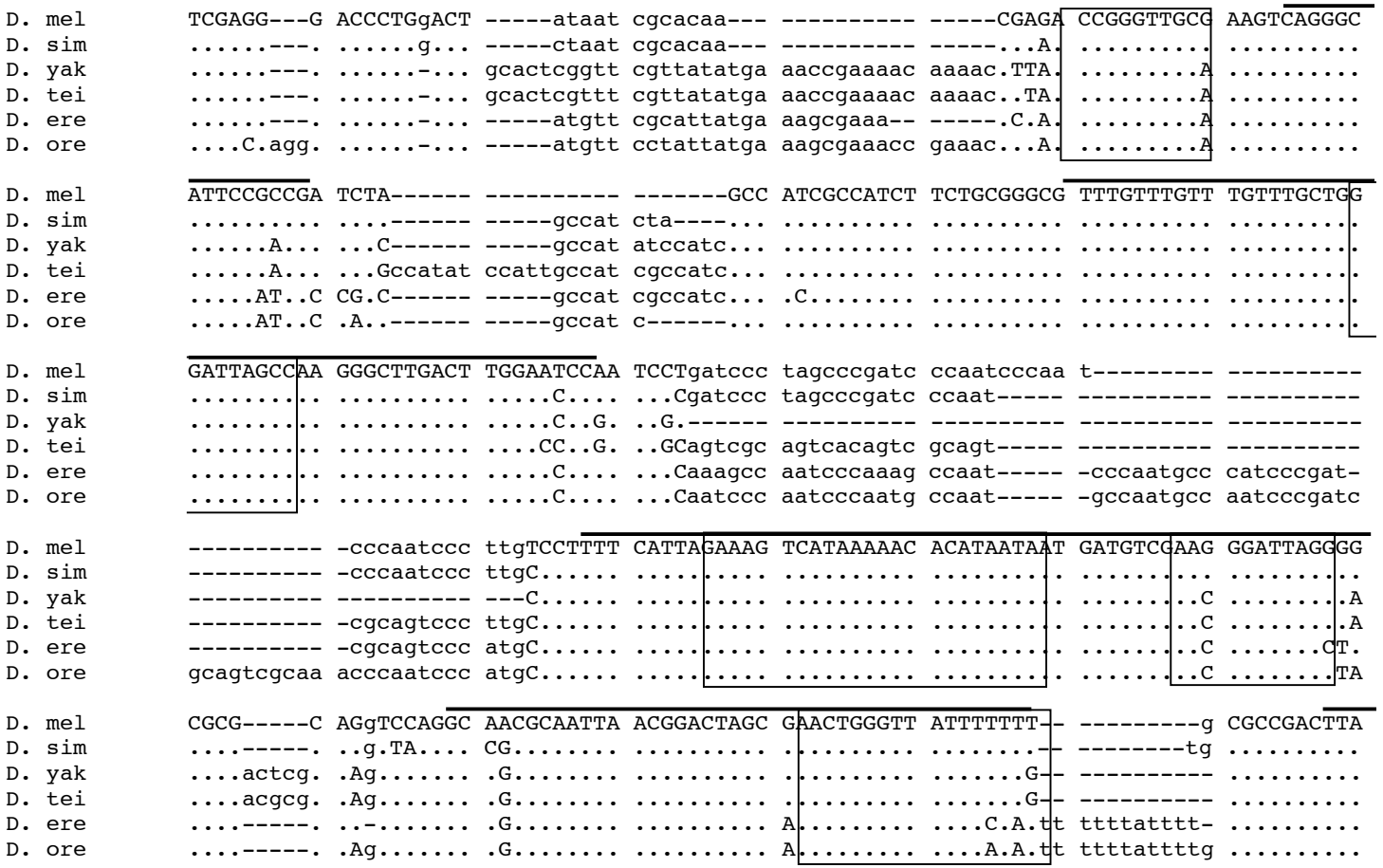
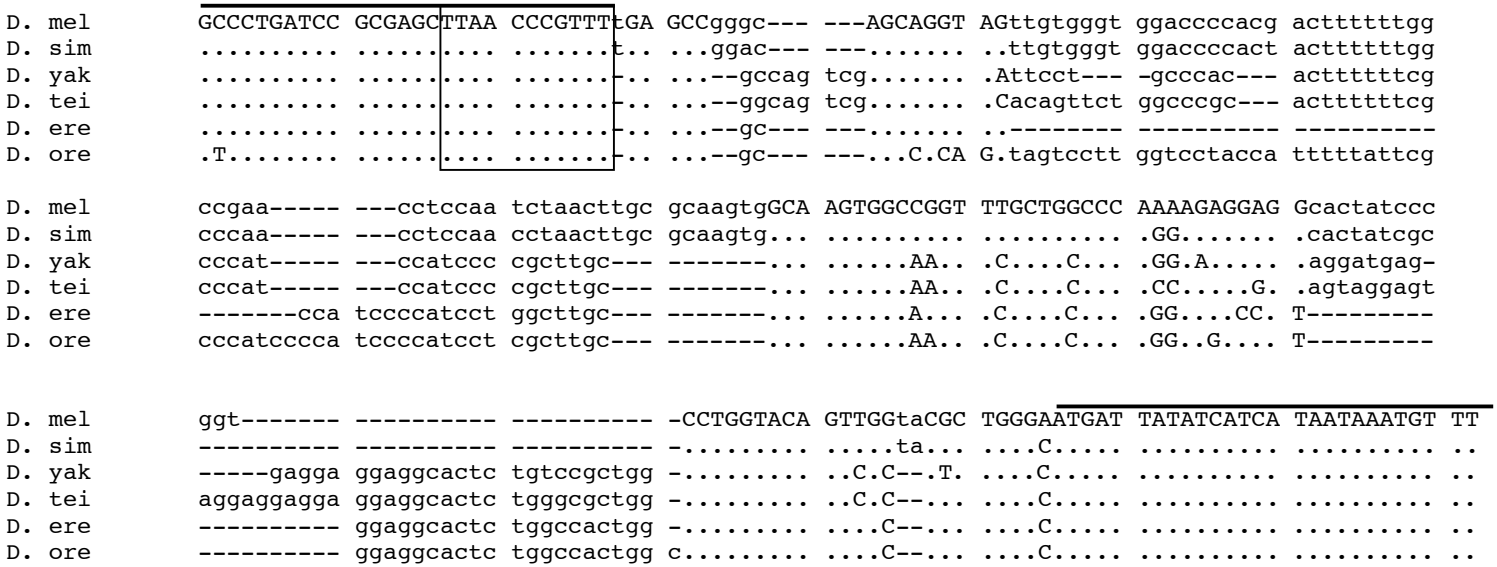


Figure 13. Continued.



and the spatial distribution of substitutions under the null hypothesis of uniform distribution. Then I addressed the specific nature of rate variation across the *eve* stripe two enhancer by contrasting models which allowed rate variation to be distributed across the sequence homogeneously or which partition rate variation *a priori* into functional classes.

It is necessary to establish the phylogenetic hypothesis for the six species analyzed since their relationships are disputed in the literature (see (Jeffs, et al. 1994; Russo, et al. 1995; Shibata 1995) vs. (Inomata, et al. 1997; Nigro, et al. 1991; Pelandakis, et al. 1991) (Munte, et al. 2001)). For all three loci I found slight, but non-significant differences among the likelihood of the four alternative topologies for the six species under the assumption of a molecular clock (Table 4). Relaxing the assumption of the molecular clock leads to an unrooted tree in which the four alternative topologies are identical, and thus the likelihoods are identical to the precision of the analysis. These results indicate that the speciation events of the three sister lineages may in the melanogaster species subgroup have occurred close enough in time for molecular data to give a inconclusive phylogenetic reconstruction. These results are consistent with conflicting reports on the phylogenetic relationships of these six species. Additionally, although maximum likelihood estimates differ among loci, these results demonstrate that the estimation of the transition: transversion rate ratio (κ) and the shape parameter (α) of the gamma distribution are not substantially affected by either topology or assumption of rate variation among lineages (Yang 1996; Yang, et al. 1994).

Table 4. Likelihoods (l) and parameter estimates (κ, α) for four alternative phylogenetic hypotheses within the melanogaster species subgroup. Topologies with the highest likelihood under a given model for each locus are shown in bold. See text for details.

Locus	Topology	clock	# params	l	Δl	pKH	κ	(s.e.)	α	(s.e.)
eve	**1**	**y**	**10**	**-1491.67**	**0**	**N.A.**	**1.871**	**0.401**	**0.266**	**0.101**
	2	y	10	-1491.86	-0.188	0.384	1.870	0.388	0.265	0.099
	3	y	10	-1491.74	-0.073	0.471	1.868	0.400	0.265	0.100
	4	y	9	-1491.86	-0.188	0.384	1.870	0.400	0.265	0.100
eve	1	n	15	-1489.35	0	0.452	1.856	0.398	0.267	0.101
	2	n	15	-1489.35	0	0.43	1.856	0.398	0.267	0.101
	3	n	15	-1489.35	0	0.441	1.856	0.398	0.267	0.101
	4	**n**	**14**	**-1489.35**	**0**	**N.A.**	**1.856**	**0.398**	**0.267**	**0.101**
Adh	1	y	10	-1591.52	-0.156	0.455	3.108	0.699	0.160	0.067
	2	**y**	**10**	**-1591.36**	**0**	**N.A.**	**3.125**	**0.702**	**0.157**	**0.066**
	3	y	10	-1591.72	-0.355	0.339	3.117	0.511	0.159	0.063
	4	y	9	-1591.72	-0.356	0.339	3.117	0.701	0.159	0.066
Adh	1	n	15	-1590.21	0	0.452	3.101	0.697	0.160	0.067
	2	n	15	-1590.21	0	0.48	3.101	0.697	0.160	0.067
	3	**n**	**15**	**-1590.21**	**0**	**N.A.**	**3.101**	**0.697**	**0.160**	**0.067**
	4	n	14	-1590.21	0	0.479	3.101	0.697	0.160	0.067
y	**1**	**y**	**10**	**-2685.48**	**0**	**N.A.**	**4.725**	**0.775**	**0.454**	**0.165**
	2	y	10	-2686.08	-0.591	0.293	4.730	0.776	0.449	0.162
	3	y	10	-2686.08	-0.591	0.293	4.730	0.767	0.449	0.161
	4	y	9	-2686.08	-0.591	0.293	4.730	0.776	0.449	0.162
y	1	n	15	-2674.81	0	0.45	4.707	0.773	0.462	0.169
	2	**n**	**15**	**-2674.81**	**0**	**N.A.**	**4.707**	**0.773**	**0.462**	**0.169**
	3	n	15	-2674.81	0	0.444	4.707	0.773	0.462	0.169
	4	n	14	-2674.81	0	0.493	4.706	0.773	0.462	0.169

Under the hypothesis that the *eve* stripe two enhancer is under stabilizing selection I can make the prediction that the rate of evolution among lineages should be approximately constant. This hypothesis can be tested by evaluating the significance of differences in likelihood (Δl) among a model incorporating rate variation among lineages (relaxing the assumption of a molecular clock) versus one which constrains the rate of evolution to remain constant among lineages (imposing the assumption of a molecular clock) using the fact that $2\Delta l$ is approximately χ^2 distributed with degrees of freedom equal to the difference in the number of free parameters (Yang 1996). Regardless of the phylogenetic hypothesis chosen, for *eve* ($2\Delta l$ = 4.6-5.0, d.f. = 5, $p > 0.1$) and *Adh* ($2\Delta l$ = 2.3-3.0, d.f. = 5, $p > 0.5$) I cannot reject the assumption of a molecular clock for any topology. In contrast, for *y* ($2\Delta l$ = 21.3-22.5, d.f. = 5, $p < 0.001$), I could reject the null hypothesis of the molecular clock since there is a significant improvement when variation in evolutionary rate among lineages is included in the model. Departure from the molecular clock has been reported previously for these species at *y* using a relative rates test, but not at *Adh* (Takano-Shimizu 1999) [see also (Munte, et al. 2001)]. Thus I conclude that the rate of point substitution for the *eve* stripe two enhancer is roughly clock-like, as is expected given previous results which suggest that the stripe two enhancer is under stabilizing selection (Ludwig, et al. 1998; Ludwig, et al. 2000). I note that clock like behavior of *cis*-regulatory sequences has been noted previously in *cis*-regulatory sequences for a primate promoter (Fracasso and Patarnello 1998).

Table 5. Likelihood and parameter estimates for models of rate variation in the *eve* stripe two enhancer. Two class models partition rate variation *a priori* according to either DNAse I or phylogenetic footprints (PF); Γ indicates the inclusion of discrete gamma distributed rate variation in the model.

model	# params	l	κ	α	c
1 class	9	-1502.56	1.81	N.A.	N.A.
1 class + Γ	10	-1491.67	1.87	0.266	N.A.
2 classes (DNAse)	10	-1493.92	1.81	N.A.	3.28
2 classes (DNAse) + Γ	11	-1484.84	1.87	0.333	3.41
2 classes (PF)	10	-1487.85	1.81	N.A.	3.03
2 classes (PF) + Γ	11	-1480.53	1.87	0.425	3.14

Since no single alternative topology explains the data significantly better than the others, I chose to study the nature of rate variation across site in the *eve* stripe two enhancer using the accepted phylogeny (tree 1) which accounts for reproductive isolation, biogeography as well as chromosomal morphology (Ashburner 1989). Using this tree I could map changes on the phylogeny, and count the number of substitutions per nucleotide site in the enhancer. Of the 654 total sites aligned, 572 sites had no substitutions, 66 sites had 1 substitution, and 16 sites had 2 substitutions. If I assume that substitutions can occur uniformly throughout the stripe two enhancer, then the number of substitutions per site should follow a Poisson distribution. Using the average number of hits per site as the estimate of the rate parameter for the Poisson distribution (λ=0.15373), I am able to reject the hypothesis that substitution occurs uniformly throughout the enhancer (χ^2 = 18.83, d.f. = 1). The inability of the Poisson distribution to fit the observed numbers of substitutions per site has classically been taken to indicate that substitution is non-uniform throughout a molecule (Fitch and Markowitz 1970). The uniformity of the spatial distribution of substitution in the enhancer can also be evaluated using a broken stick model (Goss and Lewontin 1996). Using the variance test

of Goss and Lewinton (1996), and assigning both point and indel substitution to the *D. melanogaster* coordinates, I can reject the hypothesis that substitution is uniformly distributed spatially throughout the enhancer ($p<0.001$). These results indicate that the number of substitutions per site and the spatial configuration of substitutions are non-uniform in the *eve* stripe two enhancer.

To address the nature of this rate variation across sites in the *eve* stripe two enhancer more closely, I analyzed the data under specific models of rate variation across sites using a maximum likelihood phylogenetic approach. In this analysis I categorized sites as 'sequence-specific' or 'sequence-non-specific' using two alternative *a priori* definitions based on either phylogenetic footprints or *in vitro* footprints. I used data from Stanojevic, et al. (1989) and Small, et al. (1992) to define *in vitro* footprints in the alignment of the six species and the outgroup *D. pseudoobscura* to define phylogenetic footprints (Fig. 14). The results of the PAML analyses are shown in Table 5. It is clear that a model which includes rate variation among sites according to a gamma distribution fits the data significantly better than a model with no rate variation ($2\Delta l = 22$, d.f. = 1, $p < 0.001$). A model which partitions variation in rate according to in vitro footprints ($2\Delta l = 17$, d.f. = 1, $p < 0.001$) or phylogenetic footprints ($2\Delta l = 30$, d.f. = 1, $p < 0.001$) also fit the data better than a model of no rate variation across sites. Using either categorization of the data, the data indicate that footprinted sequences evolve at roughly 1/3 the rate of spacer interval sequences. Although the individual models cannot be directly compared against one another using the χ^2 approximation since they are not nested, it is interesting to note that the order of improvement in fit among models

which include rate variation is phylogenetic footprints, then gamma distribution, then in vitro footprints. These results suggest that simply categorizing the data by *in vitro* binding properties is not the optimal design to model the evolution of *cis*-regulatory sequences.

D. Discussion

Despite recent advances in the experimental analysis and longstanding hypotheses concerning the importance of *cis*-regulatory evolution, relatively little attention has been paid to modeling the divergence of these sequences. The prevailing evolutionary model for *cis*-regulatory sequences with conserved function is one which considers the phenotype of gene expression as a quantitative trait under stabilizing selection (Carroll, et al. 2001; Ludwig, et al. 1998; Ludwig, et al. 2000; Tautz 2000). This model equates binding sites with polygenes, and changes in the sequence or spatial configuration of binding sites with changes in the number of or epistatic interactions among polygenes. This model can explain experimental results based on phenotypic assays, however, since it is a phenotypic model it does not make explicit predictions about the rate or pattern of *cis*-regulatory evolution at the DNA sequence level, other than that binding site turnover should be tolerated. Here I have taken an empirical approach to the genotypic modelling of *cis*-regulatory sequence evolution under stabilizing selection. Based on results from multiple enhancers in the *Adam-eve* and other genomic regions in Drosophila (Chapter II) and other species (Ishihara, et al. 2000), it is clear that the pattern of *cis*-regulatory conservation can be described qualitatively by two classes of functional constraint -- phylogenetic footprints and spacer intervals (Figs 11 and 12). It is my view that this well-documented pattern should form the basis for explicitly modelling and testing *cis*-regulatory sequence evolution.

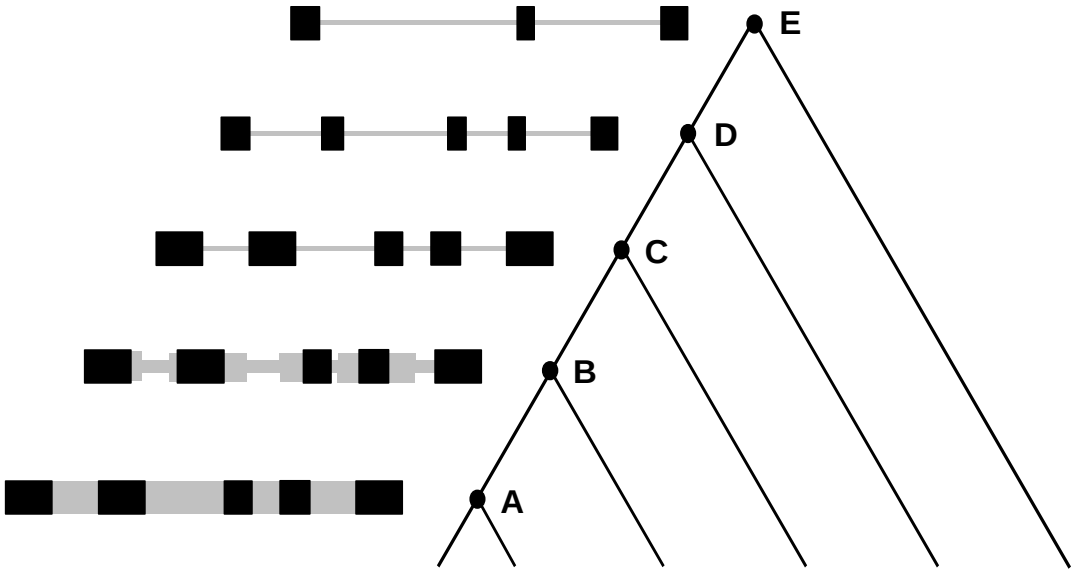


Figure 14. Two-class model of enhancer divergence under stabilizing selection. Black boxes represent core enhancer sequences (footprints) under intense functional constraint for primary sequence; grey boxes represent spacer interval sequences under weak functional constraint for primary sequence. (A) Divergence between closely related taxa (e.g. *D. melanogaster* vs. *D. simulans*) can not discriminate footprints from spacers. (B) Intermediate divergence (e.g. *D. melanogaster* vs. *D. erecta*) reveals differential constraints among aligneable sequences, but does not accurately define footprints sequences. (C) Increasing divergence (e.g. *D. melanogaster* vs. *D. ananassae*) accurately discriminates footprints from spacers but does not allow spacers to be aligned. (D) Divergence past the point of phylogentic footprint definition (e.g. *D. melanogaster* vs. *D. pseudoobscura*) reveals the same footprint architecture as in (C), but minimalizes module sequences (note smaller black boxes). (E) Subsequent divergence in core enhancer architecture leads to loss of footprints otherwise conserved in the ingroup.

Assuming that the pattern of conservation observed in the *Adam-eve* region is representative of complex *cis*-regulatory sequences under stabilizing selection in the Drosophila genome, I can begin to formulate this pattern of into a simple model of enhancer divergence (Fig 14). This model invokes two classes of functional constraint to explain the pattern of *cis*-regulatory conservation: a "sequence-specific" phylogenetic footprint class (black) and a "sequence-non-specific" spacer interval class (gray). Sequence-specific constraint is postulated to occur from protein-DNA interactions at

transcription factor binding sites (Arnone and Davidson 1997); sequence-non-specific DNA is not constrained for primary sequence but is constrained to present for effects on spacing or structure (Ondek, et al. 1988). Under this model, comparisons among very closely related taxa would allow the entire enhancer element to be aligned, since the number of substitutions in either class is too low to differentiate them from one another (14A). Comparisons among distantly related taxa (14C-D) would only allow sites in the highly constrained sequence-specific class to be aligned, since the sequence-non-specific class would have diverged beyond the point of reliable alignment. Comparisons among taxa with intermediate divergence in which the majority of sites are still alignable (14B) provide the opportunity to test patterns of substitution, including variation in rate among lineages and across sites. Quantitatively, this model predicts that *cis*-regulatory sequences under stabilizing selection should show clock-like behavior across lineages, that the pattern of substitution across *cis*-regulatory sequences should be non-uniform, and that rates of point substitution will be lower in the sequence-specific class relative to the sequence-non-specific class. It is worth pointing out that this latter prediction is not a straw man hypothesis, since it is possible that the differential rates of evolution among phylogenetic footprints and spacer intervals may be caused by indel rather than point substitution.

A simple model which makes explicit predictions about *cis*-regulatory molecular evolution is only useful if it can be tested by sequence analysis. Consideration of the dynamics of enhancer divergence suggest that a restricted taxonomic sampling framework (intermediate levels of divergence with reliable multiple alignment) is

necessary to test such a model (Fig 13). Analysis of closely related taxa (or alleles) may not have sufficient levels of variation to detect predictions of the model. Conversely, analysis of extremely divergent taxa precludes effective testing of the model since it is unlikely to observe both functional classes of DNA in multiple alignments which include divergent taxa. Advances in the automatic alignment of divergent non-coding sequences do not alleviate the problem that divergent comparisons tend to bias which sites can be studied in a multiple alignment framework (Jareborg, et al. 1999; Morgenstern 1999; Wasserman, et al. 2000). In particular, increasing the divergence among taxa in a multiple alignment biases analysis towards the most slowly evolving sequences. This bias also applies to the estimation of divergence in sequences under functional constraint (Palumbi, 1989).

Fortuitously, *eve* stripe two enhancer sequences in the melanogaster species subgroup fulfil the criteria for the appropriate taxonomic sampling to test predictions of this model. Using a multiple pairwise alignment algorithm (DiAlign 2.1), I was able to automatically align stripe two sequences for six species in this clade with minimal manual adjustment. Using this alignment I was able to show that point substitution within the *eve* stripe two enhancer is consistent with a molecular clock (table 4), dispelling concern that variation in substitution rate across lineages may complicate interpretation of rate variation across site in the enhancer. I was also able to show that none of the four alternative phylogenetic hypothesis concerning the relationship of lineages in the melanogaster subgroup (see Materials and Methods) explain the data significantly better than the others. This result suggests that the divergence times

between internal branches among these lineages are too short to be reliable reconstructed using such a small amount of data. Accepting the lack of resolution among lineages within the melanogaster subgroup helps explain conflicting reports concerning the phylogenetic relationships among these taxa (Inomata, et al. 1997; Jeffs, et al. 1994; Munte, et al. 2001; Nigro, et al. 1991; Pelandakis, et al. 1991; Russo, et al. 1995; Shibata 1995).

Using the accepted phylogeny of these species, were able to show that the number of substitutions per site in the *eve* stripe two enhancer does not fit a Poisson distribution. Moreover, I was able to show that the location of substituted sites in the enhancer is significantly more variable than expected under a model which distributes events randomly according to a broken stick model. These results together indicate that substitution in the stripe two enhancer is non-uniform, as expected under a model which invokes two class of functional constraint. Further, using maximum likelihood analysis I can show that a model which include rate variation across sites fits that data significantly better than a model which does not include rate variation (table 5). Moreover, a model which constrains rate variation to be partitioned *a priori* according to either *in vitro* or phylogenetic footprints also explain the data better than a model with no rate variation. Interestingly, although I cannot directly test these three models against one another, a model which assigns rate variation *a priori* according to phylogenetic footprints fits the data better than one which assigns variation according to *in vitro* footprints. What is more, a model which assigns rate variation according to a gamma distribution fits the data better than a model which assigns variation *a priori* according to *in vitro* footprints.

I interpret these results as evolutionary proof that the functional characterization of sequence-specific constraint in the *eve* stripe two enhancer is incomplete. These results also suggest that phylogenetic footprints rather than *in vitro* footprints should be used for the *a priori* categorization of sites for statistical tests of *cis*-regulatory evolution.

Since the two class framework is the first attempt to explicitly model *cis*-regulatory molecular evolution at the DNA level, it is no surprise that there are shortcomings which need to be addressed. For instance, both testing fit to the Poisson and evaluating the likelihood of various models of rate variation require an underlying alignment. I have chosen just one alignment to evaluate, and due to ambiguity in alignment arising from indel substitution the results of my analysis may be contingent on this alignment. Multiple alignment of *cis*-regulatory sequences is a recalcitrant problem and is intimately linked to the very nature of *cis*-regulatory divergence. Second, my model does not include indel substitution in the mechanism of *cis*-regulatory divergence. Insertions and deletions are notoriously difficult to model in alignment as well as in dynamical models of substitution. A realistic model of *cis*-regulatory divergence should include the ability to distinguish differential rate and pattern of indel substitutions in the two proposed classes of functional constraint. Finally, the results of Ludwig, et al. (2000) suggest a model of changing functional constraint in enhancer sequences even under stabilizing selection. In contrast, my two class model implicitly assumes that constraints remain constant over time to simplify analysis, although a model of changing functional constraint is clearly more realistic.

CHAPTER IV.

BINDING SITE FLUX DURING ENHANCER DIVERGENCE

A. Introduction

Multicellular eukaryotic tissue- and stage-specific gene expression is often controlled by transcriptional regulatory regions called enhancers. DNA sequences that act as enhancers of transcription are functionally defined by their ability to regulate a reporter gene in transformation or transfection experiments in a manner which recapitulates native gene expression. Generally, enhancer sequences are functional when placed in either the 5'->3' or 3'->5' orientation, either upstream or downstream of the transcriptional start site and are thought to increase the rate of transcription over basal levels by recruiting components the transcriptional complex to the promoter (Lewin 1994; Ptashne 1997). Experimental studies have revealed that the fine structure of enhancers can be typified as a closely linked group of transcription factor binding sites which coordinately bind multiple copies of one or more transcription factors (Arnone and Davidson 1997). Given that the fundamental unit of enhancer structure and function is the transcription factor binding site, it is likely that the transcription factor binding site is also the fundamental unit of evolution for enhancer sequences. On this hypothesis, I elaborate an alignment-independent method for the molecular evolutionary analysis of enhancer sequences with regards to their potential binding site composition.

Blastoderm *Drosophila melanogaster* embryos express the pair-rule gene *even-skipped* (*eve*) in seven transverse stripes (Frasch 1987), each of which has been shown to be governed by unique or shared enhancers (Fujioka 1999; Goto 1989; Harding 1989; Sackerson, et al. 1999). The *cis*-regulatory sequence responsible for the second-most anterior stripe of *eve* expression has been the subject of extensive genetic, biochemical and transgenic studies and represents a model eukaryotic enhancer. *eve* stripe two expression is achieved by the activators *bicoid* (*bcd*) and *hunchback* (*hb*) and spatially limited by the repressors *giant* (*gt*) and *Kruppel* (*Kr*) (Small, et al. 1991; Stanojevic, et al. 1991). Each of these transcription factors have been shown to interact directly with a 670 bp fragment located ~ 1 kb from the transcription start site by DNAse 1 footprinting experiments (Small, et al. 1991; Stanojevic, et al. 1989). These footprints have been demonstrated to be *bona fide* protein binding sites by mutagenesis and transformation (Small 1992 Arnosti, et al 1996; Stanojevic, et al. 1991). However, the precise location and configuration of binding sites in the eve stripe two enhancer is flexible. As shown by *cis*-complementation experiments, suggesting that flux in the binding site composition of the *eve* stripe two enhancer may be possible in evolution (Arnosti, et al. 1996).

In fact, previous work has shown the existence of population genetic variation for point and indel substitutions in both transcription factor binding sites as well as the spacer intervals between binding sites in *eve* stripe two (Ludwig and Kreitman 1995). Similar results have been obtained for the *fushi terazu* 'zebra element' (Jenkins, et al 1995), and the *dpp* intronic enhancer (Richter, et al. 1997). These studies demonstrate that the requisite intraspecific molecular variation necessary for binding site flux exists

in natural populations of Drosophila. It is clear that loss of binding sites occurs during enhancer divergence since functionally characterized known binding sites are not always conserved (Dickinson 1991; Ludwig, et al. 1998). Moreover, experimental dissection of enhancers in other fly lineages besides *D. melanogaster* show that the gain of binding sites likely occurs in other lineages as well. (Bonneton 1997; Moses, et al. 1990). Finally transgenic analysis of interspecific chimeras demonstrates that multiple interacting substitutions must have occurred in *eve* stripe two enhancer evolution (Ludwig, et al. 2000). Thus, just as experimental analysis of the D. melanogaster eve stripe two enhancer would indicate, flexibility in the configuration of binding sites is also tolerated in evolution.

The purpose of this chapter is to develop methods for the analysis of binding site flux in *cis*-regulatory sequences. In particular, I am interested in assessing variation within conserved binding sites, and changes in the overall architecture of the enhancer across taxa. To do this, I take advantage of the current trend in transcription factor binding site prediction which replaces the notion of a 'consensus' binding site sequence and with a probabilistic representation of a binding site referred to as a position weight matrix (PWM) [e.g. (Frech 1997)]. PWMs are used to transform windows of DNA sequence into random variables which measure the likelihood that a sequence binds a given transcription factor, which I term a binding site likelihood scan. The multi-taxa extension of the single taxon-binding prediction is increasingly common way to analyze *cis*-regulatory molecular evolution (Krawczak, et al. 1999; Liu, et al. 2000; Ludwig, et al. 2000). Such methods are able to capture conservation of binding sites, as well as

putative gains/losses of binding sites in any lineage sampled, and provides a convenient way to visualize the potential binding site composition of enhancer sequences. Importantly, these methods are alignment-independent and thus circumvent many of the difficulties associated with alignment of *cis*-regulatory sequences (Chapters I, III).

B. Materials and Methods

My method of binding site prediction is derived from those of (Claverie 1996; Fickett 1996). First, a sample of *n* sequences is obtained, each known to contain the binding site motif but without specifying the exact position of the binding site within the target sequence. Second, a local alignment of width *w* is derived for the sample which contains one instance of the binding site for each sequence. This *n* x *w* block alignment is then collapsed to a 4 x *w* matrix of probabilities, with each entry representing the frequency of a particular nucleotide at each position in the alignment. Observed probabilities (q_{ij}) are corrected for sampling effects, and these estimated frequencies (Q_{ij}) are rescaled relative to expected nucleotide frequencies (p_i), and log transformed to generate a matrix of scores, s_{ij} for each nucleotide-position cell in the matrix, according to $s_{ij} = \log_{10}(Q_{ij}/p_i)$. Finally, the likelihood that a window of sequence is a binding site under a given PWM was then calculated for all overlapping windows in the data set of *Drosophila eve* stripe two enhancer sequences.

I have chosen to use wild type transcription factor binding sites as revealed by DNAse 1 footprinting experiments from multiple *cis*-regulatory regions in the *Drosophila* genome for three of the four transcription factors that act on *eve* stripe 2 -- *bcd*, *hb*, and *Kr*. Only six binding sites have been reported for *gt*, three of which are in the *eve* stripe two enhancer, so I have omitted gt from the PWM-based analyses. In contrast to (Fickett 1996), I do not require the footprinted site to be conserved across

taxa for inclusion into my sample, since comparative data is only available for a small fraction of sites and this criteria may bias the sample towards high affinity sites. For all three factors the I found footprinting data in addition to the sample of sites provided by the TRANSFAC release 3.3 (Heinemeyer 1998). Part of the discrepancy may be due to different criteria for inclusion; part may be due to the differential survey of the literature.

For each site, I extracted the sequences corresponding to the footprinted sequence and approximately 25 bp of upstream and downstream flanking sequence. This was done to circumvent any ascertainment bias of the binding site motif, since footprinting conditions vary among different subsamples for each transcription factor. Where two sites occurred on the same strand for the same factor with less than 25 bp spacing, the flanking sequence was truncated to ensure only one site per sequence in the sample. Preliminary alignment efforts allowed the strand orientation of binding sites to be determined in ambiguous cases. The sample of binding sites was then locally aligned using the predictive update version of the Gibbs sampling algorithm (Lawrence 1993). In brief, this alignment tool operates by maximizing the likelihood of the binding site 'joint distribution,' relative to a null distribution of nucleotide frequencies based on the non-binding site flanking sequences. This implementation of the Gibbs sampler takes the window size, *w*, as a parameter of the search, a value which is unknown *a priori* since the limits of the footprint do not always correspond to the limits of the binding site. Also, the Gibbs sampler is a stochastic algorithm and is not guaranteed to converge to the global optimal solution. For these reasons, the Gibbs sampler was applied to each sample 500 times for each window size, *w*, constrained to the range of 5-15 base pairs.

I adapted a heuristic guideline to choose the optimal window size for subsequent analyses previously set forth for the modelling of protein domains (Lawrence 1993). Lawrence, et al (1993) suggest plotting a measure of alignment non-randomness, called information per parameter (IPP), as a function of window size. IPP is a measure of non-randomness which is proportional to the total log-likelihood or Shannon's information measure of the alignment, scaled to account for differences in window size and sample size. If there is an optimal window size for a particular transcription factor then IPP should increase with *w*, until *w* reaches the optimal window size. For window sizes larger than the optimum, the IPP should decline since additional sequences should only contribute noise to the alignment. I adapt the method of Lawrence, et al (1993) to plot measures of the distribution of IPP since not only average IPP should vary as a function of *w*, but also the variance in IPP. Additionally it is possible that IPP may be a strictly decreasing function of *w* and that the method of would unreasonably suggest a window size of 1 as the optimum.

The real distribution of binding sites used by a transcription factor is unknown, and the sites used for PWM construction are only a sample of this actual distribution. For this reason it is necessary to account for unobserved data due to sampling effects. The problem of estimating residue frequencies has been studied by (Claverie 1994) for the case of amino acids and by (Claverie 1996) for the case of nucleotides. I employ the ‘proportional mode’ of (Claverie 1996) which adds 'pseudo-counts' to observed counts of nucleotides in proportion to their expected background frequencies, according to:

$$Q_{ij} = \frac{q_{ij} + \varepsilon p_i}{n + \varepsilon}$$

where ε is the total number of pseudo-counts added to matrix. Expected frequencies are taken to be p_A=p_T=0.3 and p_G=p_C=0.2 , the base composition typical of non-coding sequences in Drosophila (Moriyama and Hartl 1993). This standard has been used previously in analyses of codon bias (Akashi 1995). I chose the total number of pseudocounts by evaluating the effect of ε on the simulated distribution of each factor.

For a given transcription factor, I can ask the question: how likely is a sequence window to have come from the estimated binding site distribution, than to have come from a distribution of random base usage. For each window I compute the total score of the window

$$L = \sum_{j=1}^{w} \log \left(\frac{Q_{ij}}{p_i} \right)$$

L can be interpreted as the likelihood that the window is derived from the pool of sequences recognized by the transcription factor relative to a pool of random sequences of width *w*. Using the PWMs for three of the four transcription factors which are known to bind to the *eve* stripe two enhancer I scanned all windows in all sequences in my data set for the presence of putative transcription factor binding sites.

C. Results

My literature search for binding site for *bcd*, *hb* and *Kr* based on *in vitro* footprinting evidence revealed 51 sites for *bcd*, 93 sites for *hb*, and 37 sites for *Kr*. The *bcd* sites are taken from 8 independent *cis*-regulatory regions in the *D. melanogaster* genome, the *hb* sites are taken from 13 regions, and the *Kr* sites are taken from 7 regions. The results of this search are shown in Tables 6-8 for *bcd*, *hb* and *Kr*, respectively. Sites are named in accordance with the original reference when possible and/or numbered consecutively from 5'->3'. Site coordinates are given according to the numbering of the *D. melanogaster* scaffold sequence from which the footprinted sequence was extracted. Nucleotides shown in capital letters are protected from DNAse I cleavage as best reported in the original reference.

Local alignment using the Gibbs Sampler was performed 500 independent times for each window size from 5-15 bp to obtain the distribution of IPP scores for each sample of sites. This procedure is necessary to ensure that local optima are not obtained from a single trial of the Gibbs sampler, and to evaluate which window size is the best for subsequent binding site prediction. The results of this analysis are shown in Figures 15-17 for *bcd*, *hb* and *Kr*, respectively. Each curve represents the distribution of IPP scores for a given window size over 500 independent realizations of the Gibbs sampler. As may be expected from the fact that different factors have different specificities, the family of distributions for differing window sizes is unique for each factor.

Table 6. Aligned sample of binding sites for *bcd*.

Site	Sequence	L	Accession	Coordinates	Reference
eve s2e 5+	tgtgccgtGT TAATCCG tttgccatca	2.94	AE003831	129914 - 129962	Small, et al. (1991)
eve s2e 4-	aattgactAA TAATCTC gctgatggca	2.45	"	129983 - 129936	"
eve s2e 3+	ccctggacTA TAATCGC acaacgagac	2.19	"	130086 - 130134	"
eve s2e 2-	agcccttgGC TAATCCC agcaaacaaa	2.83	"	130227 - 130180	"
eve s2e 1-	tgcgcgcccC TAATCCC TTCgacatca	2.83	"	130348 - 130300	"
hb A1+	ctgaccaaCG TAATCCC catagaaaac	2.83	AE003680	34016 - 33968	Driever, et al. (1989)
hb A2+	gtttctGCTC TAATCCA GAatggatca	2.32	"	33906 - 33858	"
hb A3+	tctgcccaTC TAATCCC TTGacgcgtg	2.83	"	33801 - 33753	"
hb B1+	taattcatgC TAATCTG ATGActgata	2.56	"	36067 - 36019	"
hb B2+	TAATCACCTT TAATCCC AAGtactcaa	2.83	"	36017 - 35969	"
hb X1+	agctcgcTGC TAAGCTg gccatccg	2.36	"	33980 - 33938	"
hb X2+	ggccatcCGC TAAGCTc ccggatca	2.25	"	33949 - 33922	"
hb X3+	cgGATCATCC Aaatcca agtgcgcata	1.83	"	33932 - 33892	"
hb x+	gcgcaatCCT CAATCCG CGATccgtga	2.27	"	33880 - 33819	Ma, et al. (1996)
Kr 730 1-	tgaaaaaatT TAATCCG Tttctgaagg	2.94	AE003466	247497 - 247450	Hoch, et al. (1991)
Kr 730 2+	ttcagacaAA TAATCCA gccttaagca	2.32	"	247470 - 247513	"
Kr 730 3-	ggatcaagcT TAATCAC catgcttaag	2.07	"	247539 - 247492	"
Kr 730 4+	gattttccTT AAATCCG tctgt	2.44	"	247541 - 247574	"
Kr 730 5+	ccgtctGT TAATCTC cggcttagag	2.45	"	247570 - 247601	"
Kr 730 6+	taactgaACT AAATCCG gcttaggatt	2.44	"	247872 - 247932	"
kni 64 1+	gtggtaCC TAAGCca gcgatttcgt	2.13	AE003592	68569 - 68539	Rivera-Pomar, et al. (1995)
kni 64 2-	ttaggtaacG AAATCGc tggcttaggt	1.70	"	68534 - 68566	"
kni 64 3+	tttcgttaCC TAATCgc gggatcagct	2.19	"	68555 - 68521	"
kni 64 4-	cagcttagg taaGCTG ATCCcgcgat	2.36	"	68513 - 68545	"
kni 64 5+	tcagcttaCC TAAGCTg cagattatcc	2.36	"	68535 - 68500	"
kni 64 6-	ctagGA TAATCTg cagcttaggt	2.56	"	68500 - 68525	"

Table 6. Continued

tll 1+	TTTTTTattt taatcag tgcatat	2.18	AE003775	180837 - 180879	Liaw, et al. (1993)
tll 2a+	aAACG CAATCTG AGCtccgcaa	1.89	"	180763 - 180787	"
tll 3+	TTTTATTtat taagcat ttataa	1.46	"	180704 - 180749	"
tll 4+	aggcaACGCC TAATCTG GCTcagccgc	2.56	"	180603 - 180655	"
tll 5+	ctgtccactt gAATCCT AAAGGCTCtc	1.40	"	180566 - 180620	"
tll 6-	acagtGCCTC TAATCTC GCTTggtcct	2.45	"	180604 - 180542	"
tll 7+	atctgct taagcgg CGAGCTTAAG	2.11	"	180502 - 180560	"
tll 8+	ctaaAATCCG TAATCTG CTtaagcggc	2.56	"	180477 - 180522	"
h s7e 1+	cgaaaccaGT CAATCTG gactaggtag	1.89	AE003554	48653 - 48711	LaRosee, et al. (1997)
h s7e 2-	gaaaatctTT TAATCTT gatagaatca	2.03	"	48815 - 48756	"
h s7e 3-	cgaaaaggta aaaggca aact	0.46	"	48919 - 48861	"
h s7e 4+	TAACGTTtgc taatgac aactggccag	0.89	"	48870 - 48928	"
h s7e 5+	tagccttgAC AAATCGC cgtggcttgg	1.70	"	48929 - 48979	"
h s7e 6+	ttgctctaCT TAAGCCG aaaaatgcgt	2.75	"	49168 - 49224	"
h s7e 7-	tgttaaaaAT TAGTCTT gctttgcttt	0.88	"	49284 - 49226	"
h s6e 1+	gtccgttttT TAAGCCt ttctgctctg	2.22	"	50348 - 50404	Hader, et al. (1998)
h s6e 2-	ggctgtttTT TAAGCCT Tttgcctaat	2.22	"	50671 - 50612	"
h s6e 3-	tttaagtcga aaagcca aaAAGTCAAa	1.64	"	50713 - 50765	"
spalt 1-	gaggtcattg cAAGCCG TTTTTCAGGg	2.08	AE003632	142053 - 142104	Kuhnlein, et al. (1997)
spalt 2+	cgatcttCGA TAAGCCG GAggaaaatg	2.75	"	141836 - 141785	"
spalt 3+	gcccacGGAC AAATCCT TTggccaaca	1.92	"	141806 - 141754	"
spalt 4+	taatggcTGC AAATCCg acgcgccata	2.44	"	141772 - 141724	"
spalt 6-	gcgtactAAT TAAGCAT GGCtcaagag	1.46	"	141668 - 141720	"
spalt 7+	gctggaatTA TAATCCC TTCGatcg	2.83	"	141626 - 141590	"

Table 7. Aligned sample of binding sites for *hb*.

Site	Sequence	L	Accession	Coordinates	Reference
en 1-	gacatttaat TTATTTATT TGcccatcga	1.85	AE003825	59502 - 59552	Zuo, et al. (1991)
en 1+	tcgtcagctg tTTTTCAAG Gcacatttaa	1.91	"	59502 - 59454	"
abd-b iab-2 1-	ttCCGCTGCA TTTTTTATG AGactgaaca	3.59	AE003715	11564 - 11610	Shimell, et al. (2000)
abd-b iab-2 2+	aaattgTCCC TTTTTTATT TTGTCtgctt	3.19	"	11507 - 11450	"
abd-b iab-2 3-	tcaaCGGCAA TTTTTTATG GTTtgcaagt	3.59	"	11399 - 11449	"
abd-b iab-2 4-	cctgtgcgaa TTTTTTGCG cgacggtgtg	3.32	"	11173 - 11226	"
abd-b iab-2 5-	tcttgtctag ttttttgTT TCTTATTTTT	3.15	"	11142 - 11192	"
ubx BRE A	tagggaaccg TTTTTTATG Tgtgcggctg	3.59	AE003714	133459 - 133417	Qian, et al. (1991)
ubx BRE B	ctggaacgaT TTTTTAATG tttctcatgt	2.48	"	133376 - 133438	"
ubx BRE C1	taacattgta CTTTTTATG Acctcgtaaa	3.01	"	133260 - 133289	"
ubx BRE C2	tgtgtgcagT TTTTTTACG aggtcataaa	3.36	"	133314 - 133260	"
ubx PBX 1-	tgctAAATCA TTTTTAAGG Gaaaaatcct	2.01	"	205260 - 205315	Zhang, et al. (1991)
ubx PBX 2+	acaatcATAA TTTTTTGCC ATggctaata	2.88	"	205254 - 205201	"
ubx PBX 3-	ttggACGCAG TTTTTTATT Agccatggca	3.19	"	205181 - 205235	"
ubx PBX 4+	cgaaGGCACC TTTTTTAAT GGCcgacggc	2.69	"	205102 - 205045	"
ubx PBX 5+	cttggtgcGA ATTTTTAAG CGgaaaaagg	2.30	"	205060 - 205008	"
ubx PBX 6+	gcaacacACA TTTTTTATG GCGcattccc	3.59	"	205034 - 204979	"
ubx PBX 7+	tcggatttGG TTTTTTACC AACAgccttt	2.91	"	204985 - 204931	"
eve s3e 16-	ccggccca ATTTTTAGT Ggaaattcga	1.93	AE003831	127465 - 127426	Stanojevic, et al. (1989)
eve s3e 15+	cgggacgcgc CTTTTTATT Ggtgcacctt	2.61	"	127493 - 127553	"
eve s3e 14a+	actagatcag TTTTTTGTT Ttggccgacc	3.15	"	127610 - 127655	"
eve s3e 14b+	ggccgaccg ATTTTTGTG Cccggtgctc	2.76	"	127646 - 127676	"
eve s3e 14c+	gcccggtgct CTCTTTACG Gtttatggcc	1.86	"	127661 - 127707	"

Table 7. Continued

eve s3e 13+	atttcccagc TTCTTTGTT Ccgggctcag	2.23	"	127685 - 127744	"
eve s3e 12+	gtatatgcag aTTTTTATG Ggtcccggcg	2.80	"	127734 - 127792	"
eve s3e 11a+	gtagatcacg TTTTTTGTT Cccattgtgc	3.15	"	127836 - 127882	"
eve s3e 11b+	ccattgtgcg cTTTTTTCG Ctgcgctagt	2.18	"	127868 - 127899	"
eve s3e 11c+	ctgcgctagT TTTTTTCCC Cgaacccagc	2.15	"	127886 - 127933	"
eve s3e 10+	actgctctaa TTTTTTAAT TCttcacggc	2.69	"	127906 - 127969	"
eve s3e 9+	TAAGAtccgt ttgtttgtg tttgtttgtc	2.37	"	128010 - 128068	"
eve s3e 8+	tggcattcac GTTTTTACG Agctc	2.78	"	128058 - 128096	"
eve s3e 6-	ttgcctcgcg TTTTTAATG Cttacacaaa	2.48	"	128295 - 128242	"
eve s2e 3-	attattatgT GTTTTTATG actttctaat	3.01	"	130322 - 130273	"
eve s2e 2+	actgggTTAT TTTTTTgcg ccgacttagc	3.32	"	130354 - 130403	"
eve s2e 1+	accccaCGAT TTTTTTggc caaacctc	2.63	"	130437 - 130478	"
hb p1 1+	gtgcgcATAA TTTTTTGTT TCTgctctaa	3.15	AE003680	33924 - 33869	Treisman, et al. (1989)
hb p1 2+	CCCGTTTTGC GTTTttaat aatatttact	2.11	"	34408 - 34353	"
hb p1 3+	cgatatATAG TTGTTTAAT TATaattcat	1.52	"	36182 - 36127	"
hb p2 4-	ctctttGCCG TTTTTTGGC ATCtccgctt	2.63	"	37338 - 37393	"
hb p2 5-	tcgccaATTG TTTTTTGGG CAActttaag	3.08	"	37431 - 37487	"
hb p2 6-	tgtattCCAC CTTTTTAAG CTAatttcgg	2.51	"	37643 - 37698	"
hb p2 7-	tttagaCCAA TTTTTTTCC CAAgcggaat	2.31	"	37744 - 37799	"
hb p2 8+	aataatagTT TTTTTTTTT AGTCCaaaat	2.58	"	37885 - 37829	"
Kr 730 1-	tagcaaattG CTTTTTATG ATCatgcatg	3.01	AE003466	247394 - 247344	Hoch, et al. (1991)
Kr 730 2+	caatttATAT TTTTTTGCt tttccttctt	2.93	"	247370 - 247421	"
Kr 730 3-	tacacTTTTT CTTTTTCTg atccagatcc	2.24	"	247467 - 247417	"
Kr 730 4+	acggattaaA TTTTTTCAG Acaaataatc	2.33	"	247457 - 247507	"

Table 7. Continued

Kr 730 5+	agcgcgacGC GTTTTTTCG Cgactccgcc	2.18	"	247578 - 247625	"
Kr 730 6+	cctgcaTTGT TTTTTTTTC AGtttcttca	2.54	"	247607 - 247655	"
Kr 730 7-	gtt GTTTTTAAG agaaatttga	2.51	"	247852 - 247822	"
Kr 730 8-	actttgaCA TTTTTTGTt gtt	3.15	"	247867 - 247847	"
Kr 730 9-	ttgcatT GTTTTTACT Ttgacatt	2.38	"	247880 - 247857	"
Kr 730 10-	gttacacatC TTTTTTGCA Ttgt	2.14	"	247906 - 247872	"
h s3/4e 1-	acatgtaaca ttTTTTTTT CTTCcttcgc	2.58	AE003554	47917 - 47857	Hartmann, et al. (1994)
h s3/4e 2+	atgcgacgag ATTTTTGCG Taatttctca	2.54	"	47926 - 47985	"
h s3/4e 3-	gaaattcgtC CTTTTTATG Ttgtccctgt	3.01	"	48022 - 47962	"
h s3/4e 4+	accgaaataT GTTTTTATG Agaggatcat	3.01	"	48001 - 48048	"
h s3/4e 5+	atcataaATT TTTTTTCCc ctaagaatgg	2.15	"	48040 - 48083	"
h s3/4e 6+	atctcgttcC ATTTTTAGC Ggaactgtcc	1.88	"	48088 - 48148	"
h s3/4e 7-	gttcatC TTTTTTGTC Acgtgctggt	3.10	"	48260 - 48220	"
h s3/4e 8-	ggtcgctctC TTTTTTGTT Tgttcatc	3.15	"	48295 - 48254	"
h s3/4e 9+	tgctcgcaGA TTTTTAGTG cgaattccgc	2.45	"	48264 - 48324	"
h s3/4e 10+	catgttcccA TTTTTTGTT Ctaatta	3.15	"	48371 - 48412	"
h s3/4e 11+	gttctaatTA TTTTTTATG Agttcagtgc	3.59	"	48403 - 48447	"
h s3/4e 12+	attgtcgcaC TTCTTTGTG Gccagccgca	2.63	"	48426 - 48486	"
h s3/4e 13+	ctattcatgA TTTTTTATT Tgccccgagc	3.19	"	48502 - 48556	"
h s3/4e 14+	agccctgctg TTCTTTTTG GCcctgtttt	2.06	"	48537 - 48572	"
h s3/4e 15+	ccctgttttC TTTTTTGTG Gttagaagtg	3.55	"	48562 - 48599	"
h s3/4e 16+	agtggacccA ATTTTTAGC Taataattgt	1.88	"	48583 - 48622	"
h s3/4e 17+	aattgttgcA ATTTTTGTG Gttttgggcc	2.76	"	48605 - 48656	"
h s7e 3+	atcgataatC TTTTTTATT aaatcttt	3.19	"	48721 - 48762	LaRosee, et al. (1997)

Table 7. Continued

h s7e 4+	taaatctttt TTTTTTTTT Gattctatca	2.58	"	48750 - 48798	"
h s7e 5-	tttgaggcaA CTTTTTGGG aatttgtttt	2.50	"	49259 - 49204	"
h s7e 6-	ctgactcaga gtttttgtG TTTTTTCTGc	2.97	"	49305 - 49246	"
h s7e 7-	ac ttttttatt ggccggcatt	3.19	"	49429 - 49370	"
h s7e 8-	tgcgctctaC TTTTTTATT ggccggcatt	3.19	"	49453 - 49404	"
h s6e 1+	gggatCGCAG TTTTTTACG ATCctcaacg	3.36	"	50313 - 50361	Hader, et al. (1998)
h s6e 2+	cctcCGTCCG TTTTTTAAG CCTTtctgct	3.09	"	50341 - 50401	"
h s6e 3+	gaacaATAAA GTTTTTGAC CAGATTccgg	2.03	"	50453 - 50513	"
h s6e 4+	AAATGTTTTT TTTTTTTTG TTTTTTTTAG	2.98	"	50497 - 50572	"
h s6e 5-	ttgcgggcTG TTTTTTAag ccttttgcct	3.09	"	50676 - 50618	"
h s6e 6-	taaggacaTC TTCTTTATc tatccatctc	2.22	"	50719 - 50660	"
h s6e 7+	tcaaATGCGA TTTTTTATG ggaacaagac	3.59	"	50716 - 50780	"
h s6e 8-	cgcaAAACCC TTTTTTGTC ttgttcccat	3.10	"	50799 - 50740	"
kni 223 1+	gcaggctGAG TTTTTTAGG CCAAttcttg	3.11	AE003592	68616 - 68551	Rivera-Pomar, et al. (1995)
kni 223 2-	ccatccgCAA TTTTTTAAG CGgaaaaggg	3.09	"	68721 - 68780	"
kni PstNru 1+	cggctcggtt tTTTTTATT atttttgaaa	3.19	"	68485 - 68424	Pankratz, et al. (1992)
kni PstNru 2-	ccggcttaaG TTTTTTGcc gcccagcaat	2.88	"	68383 - 68444	"
kni NruEco 3+	ctgaactgcG TTTTTTGta acgaattttc	2.37	"	68304 - 68249	"
kni NruEco 4+	acaacgagtc TTTTTTATg gtgtgagaaa	3.59	"	68273 - 68233	"
kni NruEco 5+	tgcttttga cTTTTTTAC aagcc	1.46	"	68137 - 68100	"
kni NruEco 6+	gTTTTTTCCc gtttttacg cggaattcct	2.78	"	68099 - 68057	"
spalt 1-	ttgcAAGCCG TTTTTCAGG ggcgttaccg	1.93	AE003632	142066 - 142105	Kuhnlein, et al. (1997)

Table 8. Aligned sample of binding sites for *Kr*.

Site	Sequence	L	Accession	Coordinates	Reference
en 1+	aaggcacatt tAACTGGTTA Attgaaggcc	2.18	AE003825	59486 - 59437	Zuo, et al. (1991)
eve s2e 6-	tacttcaaat tATTGGGTTA Tattgcgccc	1.99	AE003831	129829 - 129784	Stanojevic, et al. (1989)
eve s2e 5-	cgctgatggc aAACGGATTA Acacggcaca	3.11	"	129965 - 129916	"
eve s2e 4+	cgcacaacga gACCGGGTTG Cgaagtcagg	2.81	"	130103 - 130152	"
eve s2e 3+	taatgatgtc GAAGGGATTA Ggggcgcgca	2.67	"	130297 - 130347	"
eve s2e 2+	cggactagcg aACTGGGTTA Ttttttgcg	2.94	"	130348 - 130398	"
eve s2e 1-	gcccggctca AAACGGGTTA Agctcgcgga	3.86	"	130440 - 130392	"
hb p2 1-	ataaaaagta aaaaggatTG CGGGACTTAA	2.98	AE003680	34052 - 34104	Treisman, et al. (1989)
hb p2 2-	caaaacggga aaAAGGGGCA TTTAcggaat	2.52	"	34369 - 34420	"
kni 3+	gttGACTTTT AAAAGGGTTA Caattaaatt	4.08	AE003592	67725 - 67659	Rivera-Pomar, et al. (1995)
kni 2+	AATATTCATA AAAAGAGTTA AGtgccgcca	3.30	"	67794 - 67728	"
kni K1+	ttagcGGCAT AAAAGGGTTA AACAGGtagc	4.08	"	67849 - 67798	Pankratz, et al. (1989)
kni K2+	gcgcGCCCAT AAAAGGGTTA AGcacatcgg	4.08	"	67879 - 67828	"
kni 1-	gcATTGTACC AAAAGGGTTG TTCagaccca	3.74	"	67845 - 67914	Rivera-Pomar, et al. (1995)
abd-b iab-2 1-	cgcagtgcgt GAAAGGGTGA Agctaccaat	3.48	AE003715	11011 - 11060	Shimell, et al. (1994)
ubx PBX 1-	gaccgcatgc aAAAGGGTCA Cggatgtggg	3.40	AE003714	204859 - 204906	Zhang, et al. (1991)
h s6e 1+	ttacgatcct cAACGGGTTT Tacgacctcc	3.13	AE003554	50320 - 50379	Langeland, et al. (1994)
h s6e 2-	tggcagagca gAAAGGCTTA Aaaaacggac	3.02	"	50397 - 50349	"
h s6e 3-	gatttgaact gAACGGGTCA Gaggatggca	3.09	"	50433 - 50386	"
h s6e 4a+	CGGGTTCTca tagcgggttg taaaat	1.90	"	50411 - 50463	"
h s6e 4b+	gcgatttttt AAAATGTTTt tttagaagtg	1.46	"	50495 - 50567	Hader, et al. (1998)

Table 8. Continued.

h s6e 5+	ttgattaggc aAAAGGCTTA Aaaaacagcc	3.11	"	50611 - 50669	Langeland, et al. (1994)
h s6e 6+	acagcccgca accTTGGTGT TAAttgagat	0.13	"	50637 - 50678	"
h s6e 7+	aacaagacaa aAAAGGGTTT Tgcggagata	3.69	"	50744 - 50803	"
h s5e 1-	gcctGATTCG CAAAgagttt ttacgacacg	2.56	"	53640 - 53584	"
h s5e 2-	tagagttt CCACGGATTA gacctctgtg	1.88	"	53742 - 53701	"
h s5e 3-	ttctcTCTGG GATAGagttt agagtttcca	1.89	"	53761 - 53727	"
h s5e 4-	gtctCGCGCC GAAAgaggtt ctctct	1.93	"	53783 - 53754	"
h s7e 1-	agttaaagta gAAGGAGTGA Cactctttcg	2.14	"	48866 - 48806	La Rosee-Borggreve, et al. (1999)
h s7e 2-	ttcggtttca aGCTGAGTTA Aaactcctca	1.13	"	49015 - 48956	"
h s7e 3-	ggttGGCAGG GAAAgtgttg acatgatggt	2.43	"	49071 - 49012	"
h s7e 4-	ttaattgtgg cAAAGGGTTT Cgcccggctc	3.34	"	49137 - 49078	"
h s7e 5+	caagggtcac cAAAGGGTTC Gtgaggatct	2.88	"	49314 - 49373	"
spalt 1-	AGTGTCATAG AAATGGGTGA AATTCTgttc	3.05	AE003632	141958 - 142024	Kuhnlein, et al. (1997)
spalt 2+	gatcccacgc AAAAGGATGG CACAAATTTC	2.48	"	141981 - 141919	"
spalt 3+	TTTGGCCAAC AAAAGGGTAA TGGCTGcaaa	3.06	"	141802 - 141729	"
spalt 4-	AaatccgaTC GAAGGGATTA Taattccagc	2.67	"	141579 - 141630	"

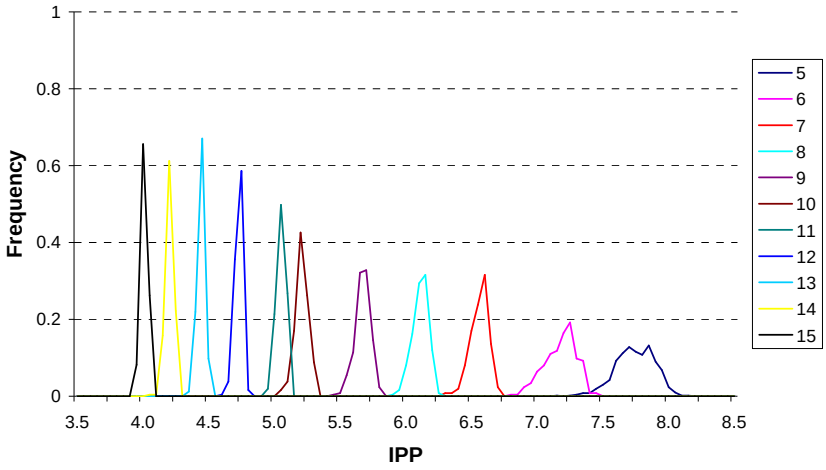


Figure 15. Distribution of IPP scores for alignments of *bcd* binding sites. The frequency of a given IPP value for 500 realizations of the Gibbs sampler are shown for fixed window sizes of 5-15 bp.

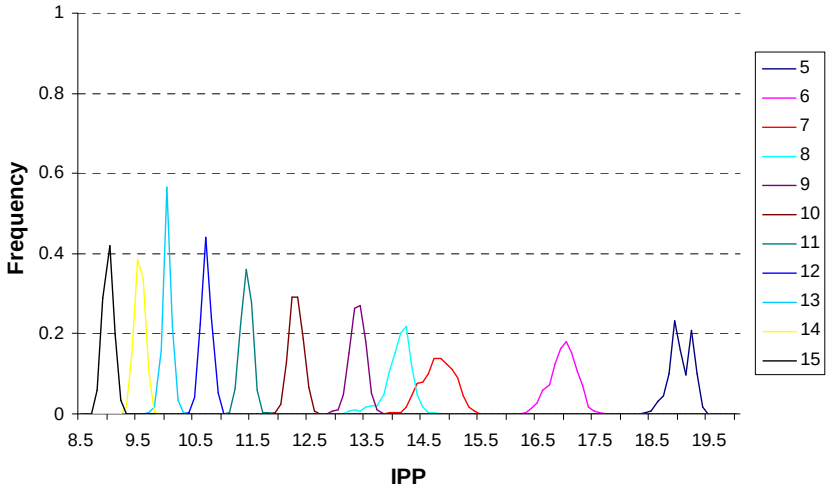


Figure 16. Distribution of IPP scores for alignments of *hb* binding sites. The frequency of a given IPP value for 500 realizations of the Gibbs sampler are shown for fixed window sizes of 5-15 bp.

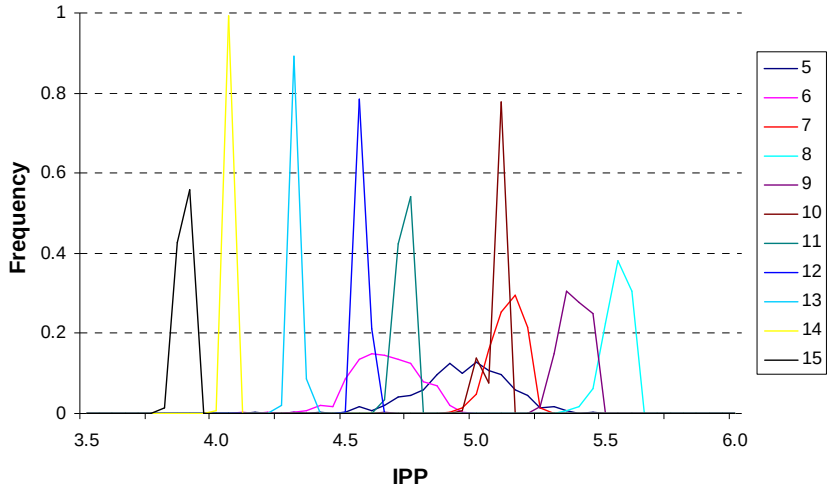


Figure 17. Distribution of IPP scores for alignments of *Kr* binding sites. The frequency of a given IPP value for 500 realizations of the Gibbs sampler are shown for fixed window sizes of 5-15 bp.

To evaluate which summary properties of the family of IPP distributions can identify the optimal window size for binding site prediction, I plotted the mean, standard deviation and coefficient of variation for the IPP of different window sizes for the *bcd*, *hb* and *Kr* samples in Figures 18-20. For both *bcd* and *hb* the mean IPP is a decreasing function of length, whereas for *Kr*, there is a peak in mean IPP at 8 bp, after which IPP decreases with length. This is reflected in the distribution plots as well with the 8 bp *Kr* distribution being centered at the highest IPP value. Thus for *Kr*, it is possible to use the

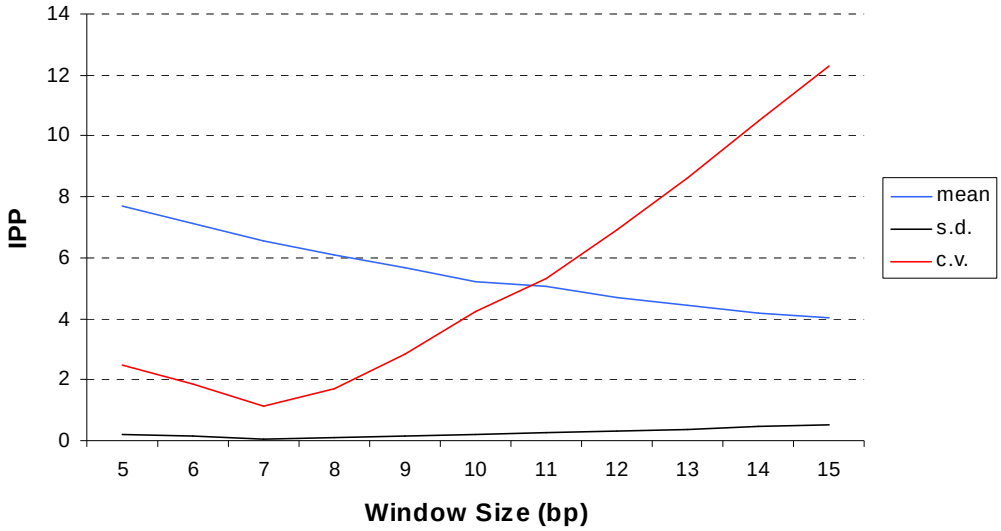


Figure 18. Summary statistics for *bcd* IPP scores. Plotted are the mean, standard deviation (s.d) and coefficient of variation (c.v. = s.d.x100/mean) for IPPs of 500 independent realizations of the Gibbs sampler for fixed window sizes 5-15 bp.

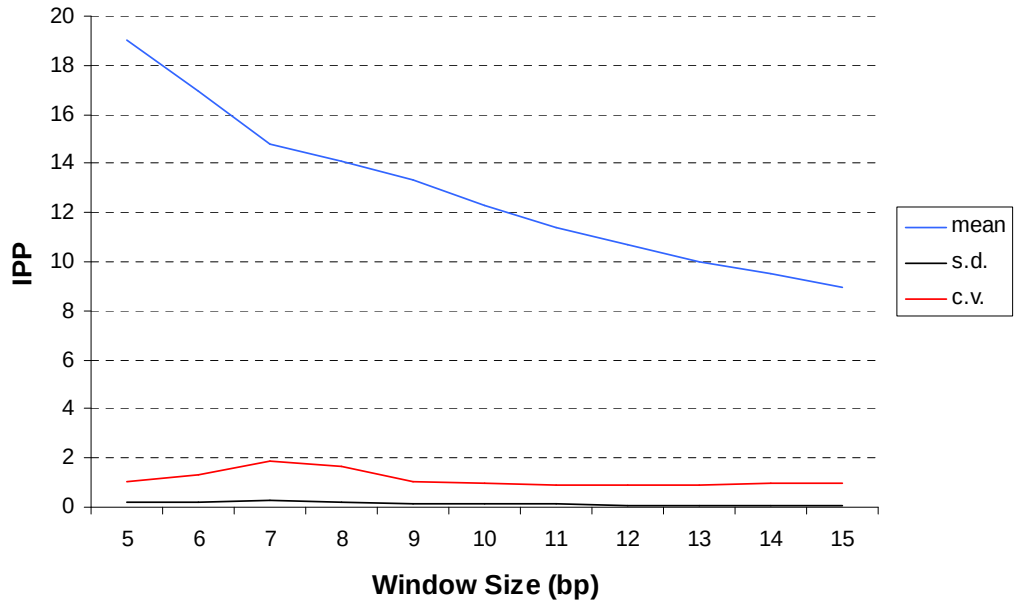


Figure 19. Summary statistics for *hb* IPP scores. Plotted are the mean, standard deviation (s.d) and coefficient of variation (c.v. = s.d.x100/mean) for IPPs of 500 independent realizations of the Gibbs sampler for fixed window sizes 5-15 bp.

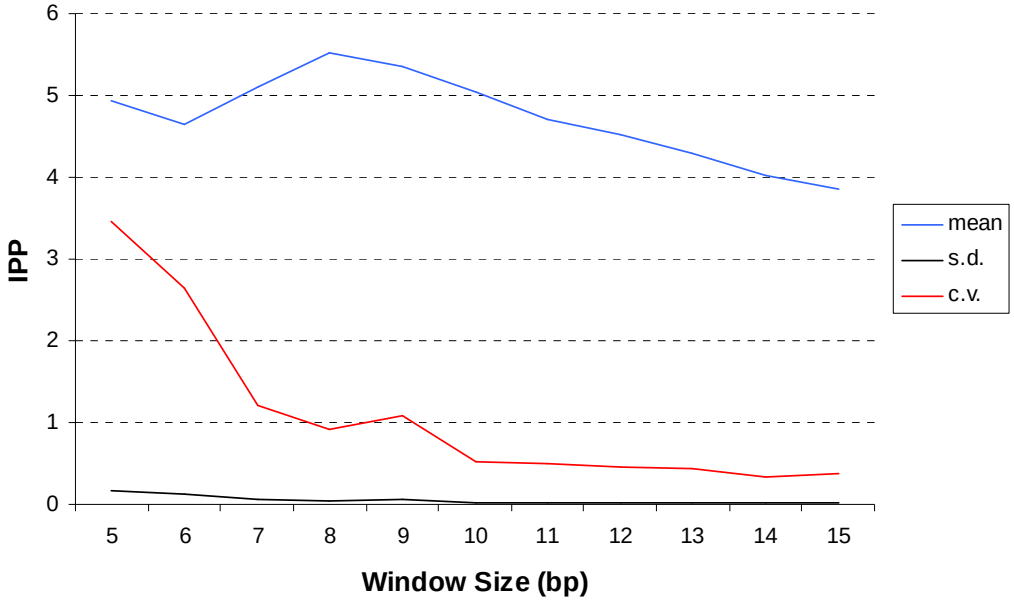


Figure 20. Summary statistics for *Kr* IPP scores. Plotted are the mean, standard deviation (s.d) and coefficient of variation (c.v. = s.d.x100/mean) for IPPs of 500 independent realizations of the Gibbs sampler for fixed window sizes 5-15 bp.

heuristic of Lawrence, et. al. (1993) to choose an optimal window size, but to identify the optimal window size for binding site prediction for *bcd* and *hb*, I plotted the mean, standard deviation and coefficient of variation for the IPP of different window sizes for the *bcd*, *hb* and *Kr* samples in Figures 18-20. For both *bcd* and *hb* the mean IPP is a decreasing function of length, whereas for *Kr*, there is a peak in mean IPP at 8 bp, after which IPP decreases with length. This is reflected in the distribution plots as well with the 8 bp *Kr* distribution being centered at the highest IPP value. Thus for *Kr*, it is possible to use the heuristic of Lawrence, et. al. (1993) to choose an optimal window size, but for *bcd* and *hb* I must use alternative criteria. The standard deviation for each window size is typically small but varies as a function of window size. Since the standard deviation is typically related to the mean (Sokal and Rohlf 1995), I analyzed the

coefficient of variation for each window size to see if properties of the variation, rather than the mean, IPP value are informative indicators of optimal window size

For all three factors, the coefficient of variation has a local minimum or a range of window sizes in which the coefficient of variation plateaus, suggesting that there is an optimal window size that minimizes variation in IPP while maximizing the overall information content of the sample. For *bcd* this local minimum is at 7 bp which corresponds exactly to length of the established consensus for this factor (TAATCCC) (Driever and Nusslein-Volhard 1989). The optimal window size based on coefficient of variation is 9 bp for *hb* and 10 bp for *Kr*. The similar estimates of optimal window size for these two factors is perhaps not surprising since they both are multiple zinc finger containing proteins. Based on these results, I chose one representative alignment from the optimal window size to estimate nucleotide frequencies for *bcd*, *hb* and *Kr* PWMs (Tables 6-8).

An additional consideration in the construction of a PWM is how to deal with estimating the actual nucleotide usage of a factor from observed frequencies in a sample of sites. Small or biased samples of sites will miss rare or low affinity sites, and PWMs that do not account for this will have characteristic biases in their predictive utility. Claverie and Audic (1996) discuss several methods for the estimation of PWMs from data and I employ their proportional mode of adding pseudo-counts of imputed data to account for sites unobserved in the sample (see Materials and Methods). Although this estimation technique has desirable properties, the effect of the number of pseudocounts on the distribution of PWM scores is unknown. Lawrence et al (1993) use pseudocounts

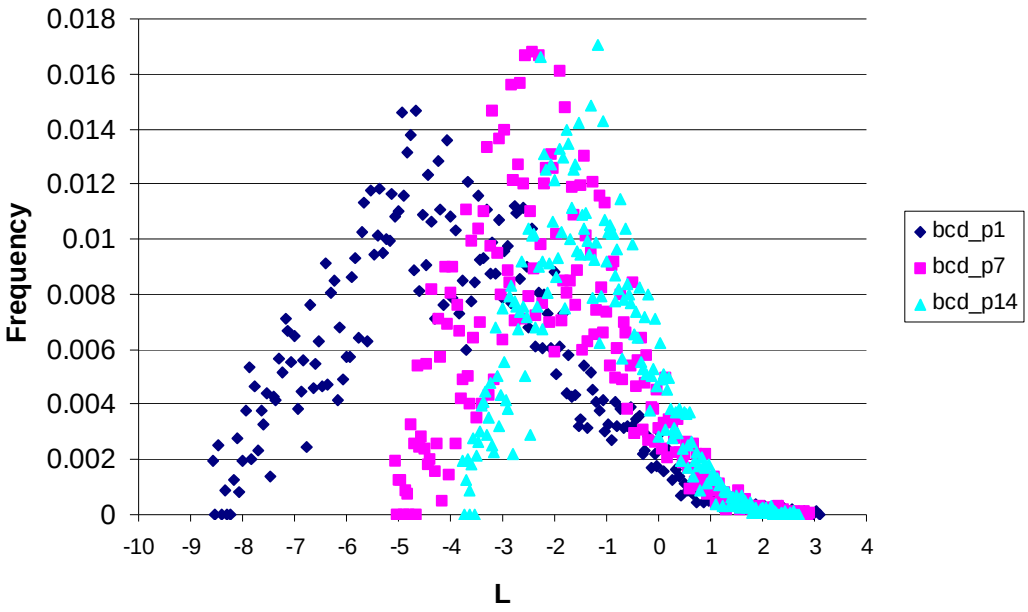


Figure 21. Simulated distribution of scores under the *bcd* PWM. Each point represents one of 10^6 simulated 7-mers scored with PWMs corrected with 1, 7, or 14 pseudocounts for a sample of size 51 sites. Data were binned into 200 categories.

to score sites during their local alignment procedure and suggest a heuristic of $n^{1/2}$ pseudocounts, where n is the sample size of real sites, although no argument or evaluation of this claim is given.

The distribution of PWM scores can be simulated easily, given an assumption of background base composition (Claverie 1996). To evaluate the claim of Lawrence et al (1993) and gain a better sense of the effect of the total number of pseudocounts on the distribution of *bcd*, *hb* and *Kr* PWMs I simulated 10^6 sequences of length equal to each PWM. The simulated sequences were scored using the PWM and then binned into 200 categories ranging from the minimum to the maximum score under the PWM. The results of these analyses are shown for 1, $n^{1/2}$ and $2*n^{1/2}$ pseudocounts for the integer

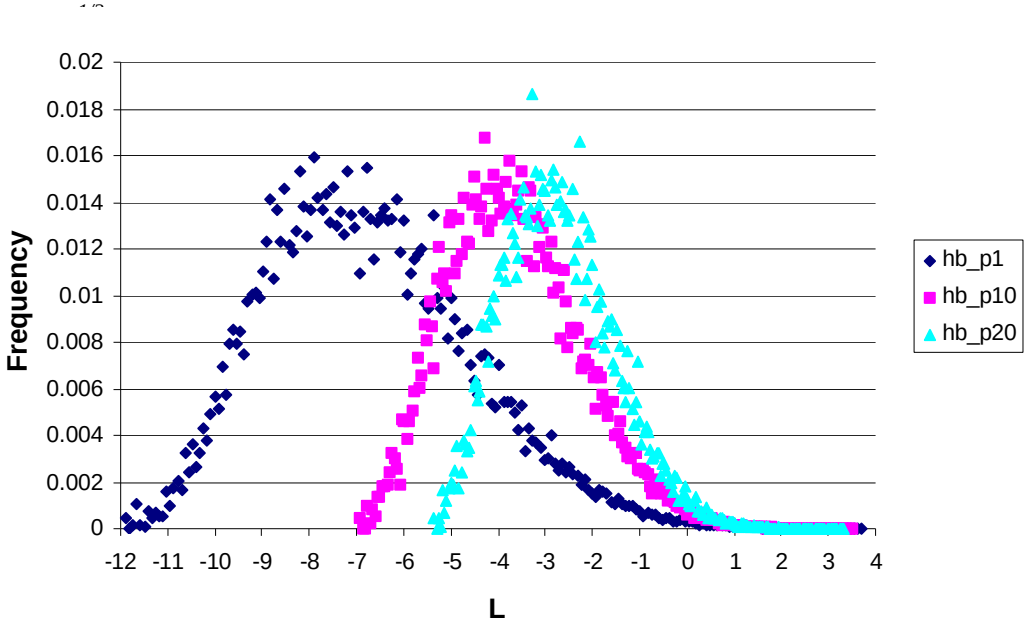


Figure 22. Simulated distribution of scores under the *hb* PWM. Each point represents one of 10^6 simulated 9-mers scored with PWMs corrected with 1, 10, or 20 pseudocounts for a sample of size 93 sites. Data were binned into 200 categories.

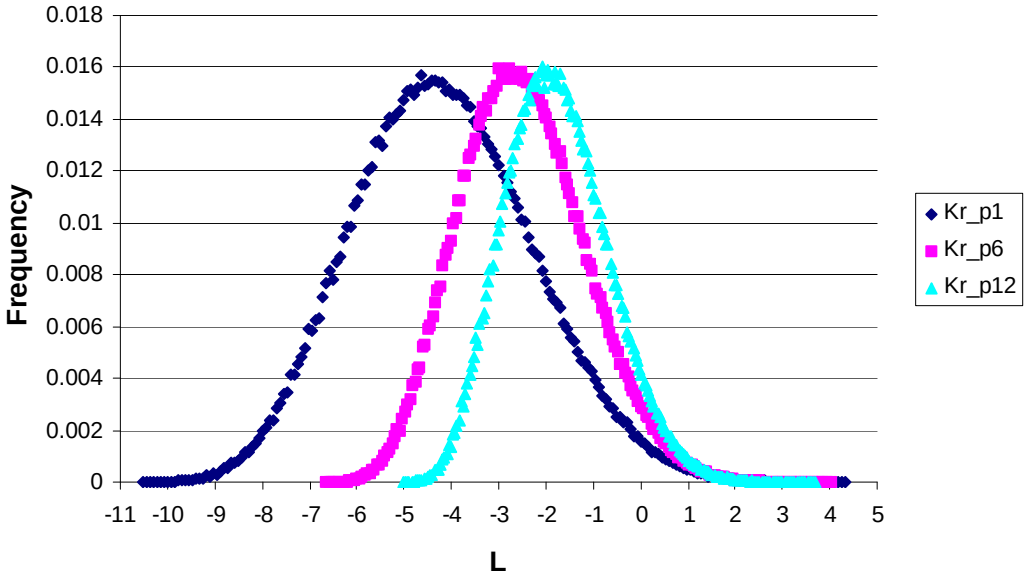


Figure 23. Simulated distribution of scores under the *Kr* PWM. Each point represents one of 10^6 simulated 10-mers scored with PWMs corrected with 1, 6, or 12 pseudocounts for a sample of size 37 sites. Data were binned into 200 categories.

decreasing the range of scores. Also, increasing the number o pseudocounts tends to collapse the distribution closer to the case of no specificity under the PWM, where all sequences have a score of 0. Because the bulk of the distribution for all three factors is below zero, this tendency affects the lower tail of the distribution more than the upper tail. Thus the number of pseudocounts does not substantially affect the region of the distribution of scores which is of interest biologically. Thus I choose the number of pseudocounts according to the heuristic of Lawrence, et al. (1993) so that both alignment and scoring use the same estimation procedure. I note that the raggedness of the distributions varies inversely as a function of the length of the PWM across factors.

Given a better understanding of the appropriate parameters to constrain PWMs for binding site prediction, it is now possible to determine a PWM for each factor and score the likelihood of target sequences for potential binding sites. The final PWMs chosen for *bcd, hb* and *Kr* are shown in Tables 9-11. As a preliminary step in using PWMs to analyze evolutionary changes in *cis*-regulatory sequences, I determined a threshold score for representing putative binding sites based on the sample of known binding sites. Determination of the threshold for scoring sites is not a simple matter of finding a cutoff under a null hypothesis since the distribution of scores under the PWM is under the hypothesis that a sequence *is* a binding site, not the null hypothesis that it is *not* a binding site (Manolis Demertzakis, personal communication). Furthermore, it is unclear how many tests are performed when scanning a sequence larger than the width of the PWM as a result of overlapping windows and the complementarity of DNA. Thus I have chosen

Table 9. Position weight matrix for *bcd*. Each cell represents the likelihood ratio of each nucleotide at various positions based on the alignment in Table 6 adjusted using 7 pseudocounts realtive to background frequencies of A=T=0.3 and G=C=0.2.

Position	A	C	G	T
1	0.70	0.47	0.21	2.19
2	3.05	0.12	0.12	0.12
3	2.99	0.12	0.21	0.12
4	0.12	0.12	1.41	2.19
5	0.12	4.34	0.29	0.12
6	0.41	2.36	0.55	0.98
7	0.47	1.50	1.93	0.58

Table 10. Position weight matrix for *hb*. Each cell represents the likelihood ratio of each nucleotide at various positions based on the alignment in Table 7 adjusted using 10 pseudocounts realtive to background frequencies of A=T=0.3 and G=C=0.2.

Position	A	C	G	T
1	0.36	0.58	0.58	2.20
2	0.10	0.10	0.10	3.11
3	0.13	0.34	0.19	2.85
4	0.10	0.10	0.10	3.11
5	0.10	0.10	0.10	3.11
6	0.23	0.19	0.10	2.91
7	1.68	0.29	1.55	0.42
8	0.55	1.02	0.58	1.72
9	0.16	0.87	2.43	0.97

Table 11. Position weight matrix for *Kr*. Each cell represents the likelihood ratio of each nucleotide at various positions based on the alignment in Table 8 adjusted using 6 pseudocounts realtive to background frequencies of A=T=0.3 and G=C=0.2.

Position	A	C	G	T
1	1.61	0.72	1.30	0.37
2	2.78	0.37	0.26	0.14
3	2.47	0.60	0.26	0.29
4	1.77	1.07	0.49	0.53
5	0.14	0.14	4.09	0.37
6	0.60	0.14	3.63	0.22
7	0.60	0.37	3.40	0.22
8	0.14	0.14	0.37	2.85
9	0.22	0.49	0.72	2.31
10	1.84	0.26	0.84	0.76

to score each site in the sample and determine an empirical cutoff that discriminates sites which are known to bind a factor (Schneider 1997). I determined this cutoff by removing the bottom 10% of the scores in the sample of known binding sites for each factor. These cutoffs are 1.63, 2.06, and 1.90 for *bcd*, *hb* and *Kr* respectively.

Given fixed parameters and cutoffs for the three PWMs, I can know apply these models of binding site specificity to analyze divergence in binding site composition for homologous eve stripe two fragments in Drosophila. For all overlapping windows in each eve stripe two sequence, I calculated the likelihood that each window is a putative binding site relative to the likelihood of the sequence under background base usage. Positive L values indicate putative binding sites on the plus strand relative to the eve transcript and negative values represent putative sites on the minus strand. The x-axis represents the position of the first nucleotide of the window under consideration. The methods successfully identifies all functionally verified binding sites in the eve stripe two enhancer (Ludwig, et al. 2000)

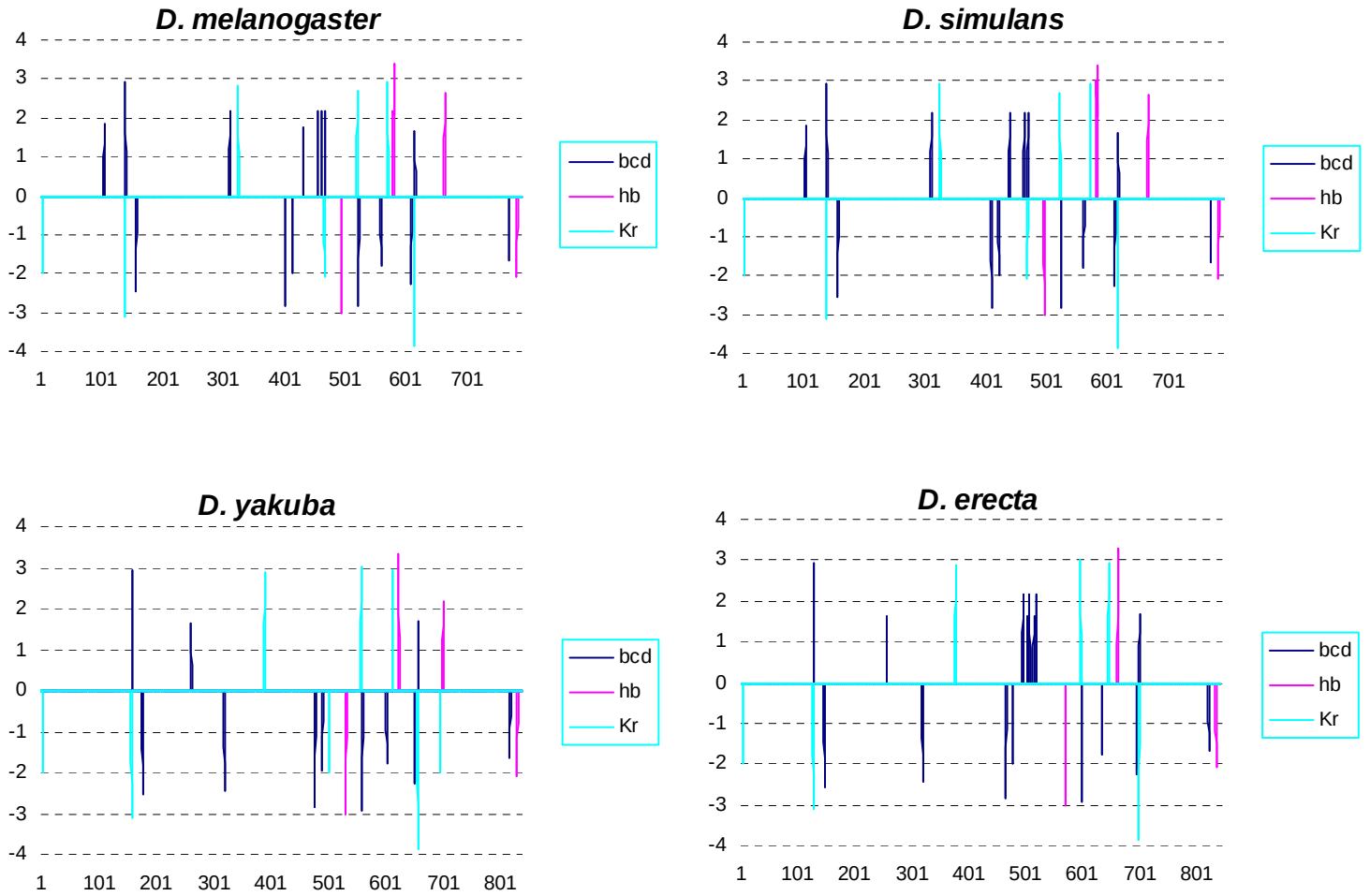


Figure 24. Binding site likelihood scans for the melanogaster species subgroup. See text for details.

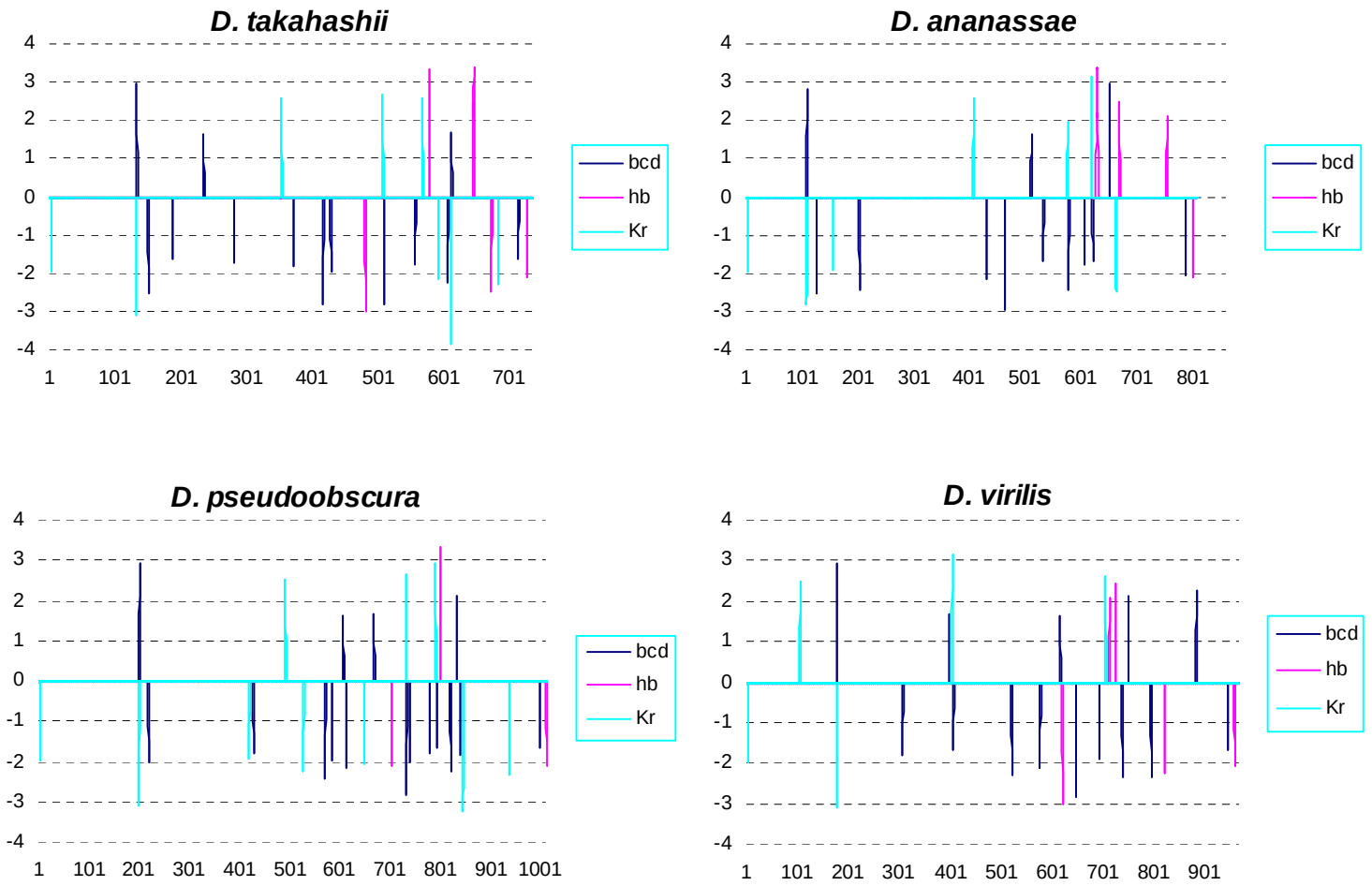


Figure 25. Binding site likelihood scans for species in the genus Drosophila. See text for details.

D. Discussion

I have presented a framework for the quantitative *in silico* evaluation of the binding site composition of uncharacterized target sequences based on PWM based binding site prediction. This approach is used often in genomic and transcriptional analyses but has only been recently proposed for use as a tool in molecular evolution. The approach relies on large samples of high quality binding site information, and this a priori requirement is the limiting factor in any analysis of this kind. For three factors known to bind to the eve stripe two enhancer high quality samples were available in the literature (Tables 6-8), and thus this target sequence represents an excellent system to evaluate multi-species binding site prediction as an approach to analyze the evolution of *cis*-regulatory sequences.

A number of technical consideration must be addressed when constructing PWMs, among them the optimal window size, the estimation of nucleotide frequencies from samples and the cutoffs for accurately scoring potential target sequences. Here I have presented a novel method to evaluate the optimal window size based on variability in stochastic local alignment outcomes via the Gibbs sampler (Figs 15-20). Although there is apparently no best method for determining window size for all three factors studied here, it is clear that the heuristic of Lawrence, et al. (1993) is insufficient for determining the optimal window size of binding sites, since IPP is generally a decreasing function of window size (Figs 15 & 16, but see Fig. 17). One characteristic common to all three factor is that at the smallest window sizes the distribution of scores is more

variable than at larger window sizes. This pattern obtains since at window sizes smaller than the true width of the binding site, several "optimal", offset, sub-alignment can found by the Gibbs sampler. At window sizes equal to or larger than the true width of the binding site there is only one "optimal" alignment state to which the sampler can converge. I propose that evaluating the variance or the coefficient of variation as a function of window size may be more appropriate for determining optimal binding site PWM lengths (Figs 18-20).

I have also studied the effects of adding pseudocounts of imputed data in the estimation of nucleotide frequencies. Although necessary to correct for rare or unobserved binding sites, my results suggest that the total number of pseudocounts does not substantially affect the distribution of scores under PWMs for any of the three factors in the regime of biological interest. High scoring putative sites are high scoring sites, regardless if 1, $n^{1/2}$ or $2*n^{1/2}$ pseudocounts are added to a sample of n real sites. These results also suggest that the fraction of high scoring sites under each of the three PWMs is low, as expected since these factors exhibit specificity. The proper statistical evaluation of this claim is difficult considering the distribution of sites which are *not* bound is unknown, and thus we cannot describe the distribution of the null hypothesis. We can use functional criteria to define cutoffs for which the majority of putative sites above this cutoff are likely to be recognized by these factors.

A number of interesting features of *cis*-regulatory structure can be revealed using a binding site prediction approach. First, it is possible to recapitulate the known binding sites in the eve stripe two enhancer using in silico methods alone (Ludwig, et al. 2000).

Second, many putative sites are observed for all three factors in addition to known sites. This may represent false positive predictions or a too lenient threshold of the current method. Alternatively these sites could be weak binding sites or sites not observed in DNAse cleavage studies. The latter alternative is supported by the argument that the functional characterization of the *eve* stripe two enhancer is a non-optimal description of constraint operating on this sequence (Chapter III). Finally, there appears to be no strand specificity in the orientation of binding sites for these three factors as would be expected if these factor do not work stereospecifically with the transcription initiation site.

Insight into *cis*-regulatory evolution as well as structure can be revealed by this approach. For example, changes in the likelihood of conserved peaks may reflect quantitative shifts in affinity, assuming *L* and affinity are correlated (Berg 1987). Also, overlapping or clustered site appear to be more conserved than individual sites, as is expected if the module is the fundamental unit of conservation (Chapter II). Each species has its own binding site profile, suggesting that the putative gain and loss of binding sites is a frequent occurrence in each lineage. Moreover, it does not appear that gains or losses affect activator (*bcd*, *hb*) or repressors (*Kr*) sites preferentially. Finally, loss and gain of sites for the same factor can occur in the same position of the enhancer suggesting that compensatory changes may occur preferentially in localized regions of the enhancer. These results support the idea that stabilizing selection on the phenotype of gene expression permits flux in the binding site composition of *cis*-regulatory sequences over evolutionary time (Carroll, et al. 2001; Ludwig, et al. 2000; Tautz 2000).

REFERENCES

Adams, MD *et al.* (2000) The genome sequence of Drosophila melanogaster. Science 287:2185-95

Akashi, H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139:1067-1076

Andolfatto, P, Wall, JD & Kreitman, M (1999) Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of Drosophila melanogaster. Genetics 153:1297-311

Arnone, MI & Davidson, EH (1997) The hardwiring of development: organization and function of genomic regulatory systems. Development 124:1851-1864

Arnosti, DN, Barolo, S, Levine, M & Small, S (1996) The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development 122:205-214

Ashburner, M (1989) Drosophila: A laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 434

Badger, JH & Olsen, GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16:512-24

Batzoglou, S, Pachter, L, Mesirov, JP, Berger, B & Lander, ES (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res 10:950-8

Berg, OG, von Hippel, P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol 193:723-750

Blackman, RK & Meselson, M (1986) Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the Drosophila hsp82 gene. J Mol Biol 188:499-515

Blackman, RK, Sanicola, M, Raftery, LA, Gillevet, T & Gelbart, WM (1991) An extensive 3' cis-regulatory region directs the imaginal disk expression of decapentaplegic, a member of the TGF-beta family in Drosophila. Development 111:657-66

Bonneton, F, Shaw, P.J., Fazakerley, C., Shi, M., Dover, G.A. (1997) Comparison of bicoid-dependent regulation of hunc*hb*ack between Musca domestica and Drosophila melanogaster. Mech Dev 66:143-156

Brady, JP, Richmond, R.C. (1990) Molecular analysis of evolutionary changes in the expression of Drosophila esterases. Proc Natl Acad Sci U S A 87:8217-8221

Bray, SJ & Hirsh, J (1986) The Drosophila virilis dopa decarboxylase gene is developmentally regulated when integrated into Drosophila melanogaster. EMBO J 5:2305-2311

Bucher, P (1999) Regulatory elements and expression profiles. Curr Opin Struct Biol 9:400-7

Cariou, ML (1987) Biochemical phylogeny of the eight species in the Drosophila melanogaster subgroup, including D. sechellia and D. orena. Genet Res 50:181-5

Carroll, SB, Grenier, JK & Weatherbee, SD (2001) From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Science, Inc., Malden, Mass., 192

Claverie, JM (1994) Some useful statistical properties of position-weight matrices. Comput Chem 18:287-294

Claverie, JM, Audic, S. (1996) The statistical significance of nucleotide position-weight matrix matches. Comput Appl Biosci 12:431-439

Claverie, JM (1997) Computational methods for the identification of genes in vertebrate genomic sequences. Hum Mol Genet 6:1735-44

Comeron, JM (1999) K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics 15:763-4

Comeron, JM & Kreitman, M (2000) The correlation between intron length and recombination in Drosophila. dynamic equilibrium between mutational and selective forces. Genetics 156:1175-1190

Crowley, EM, Roeder, K & Bina, M (1997) A statistical model for locating regulatory regions in genomic DNA. J Mol Biol 268:8-14

Deutsch, M & Long, M (1999) Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res 27:3219-28

Driever, W & Nusslein-Volhard, C (1989) The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo. Nature 337:138-143

Dubchak, I *et al.* (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res 10:1304-6

Duret, L & Bucher, P (1997) Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 7:399-406

Feller, W (1966) An introduction to probability theory and its applications. John Wiley and Sons, Inc., New York,

Fickett, JW (1996) Quantitative discrimination of MEF2 sites. Mol Cell Biol 16:437-441

Fickett, JW & Wasserman, WW (2000) Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol 11:19-24

Fitch, WM & Markowitz, E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-93

Fracasso, C & Patarnello, T (1998) Evolution of the dystrophin muscular promoter and 5' flanking region in primates. J Mol Evol 46:168-79

Frasch, M, Levine, M. (1987) Complementary patterns of even-skipped and fushi tarazu expression involve their differential regulation by a common set of segmentation genes in Drosophila. Genes Dev 1:981-995

Frech, K, Quandt, K., Werner, T. (1997) Finding protein-binding sites in DNA sequences: the next generation. Trends Biochem Sci 1997 Mar;22(3):103-104 22:103-104

Fromental, C, Kanno, M, Nomiyama, H & Chambon, P (1988) Cooperativity and hierarchical levels of functional organization in the SV40 enhancer. Cell 54:943-53

Fujioka, M, Miskiewicz, P., Raj, L., Gulledge, A., Weir, M., Goto, T. (1996) Drosophila Paired regulates late even-skipped expression through a composite binding site for the paired domain and the homeodomain. Development 122:2697-2707

Fujioka, M, Emi-Sarker, Y., Yusibova, G.L., Goto, T., Jaynes, J.B. (1999) Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. Development 126:2527-2538

Galtier, N & Gouy, M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15:871-9

Gillespie, JH (1991) The causes of molecular evolution. Oxford University Press, New York,

Goss, PJ & Lewontin, RC (1996) Detecting heterogeneity of substitution along DNA and protein sequences. Genetics 143:589-602

Goto, T, Macdonald, P., Maniatis, T. (1989) Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell 57:413-422

Gottgens, B *et al.* (2001) Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. Genome Res 11:87-97

Gray, S, Szymanski, P., Levine. M. (1994) Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes Dev 8:1829-1838

Gu, X & Li, WH (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J Mol Evol 40:464-73

Harding, K, Hoey, T., Warrior, R., Levine, M. (1989) Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. EMBO J 8:1205-1212

Hardison, RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16:369-72

Hartl, DL & Lozovskaya, ER (1994) Genome evolution: between the nucleosome and the chromosome. EXS 69:579-592

Hasegawa, M, Kishino, H & Yano, T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-74

Heinemeyer, T, Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F.A., Podkolodny, N.L., Kolchanov, N.A. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Res 26:364-370

Hepker, J, Blackman, RK & Holmgren, R (1999) Cubitus interruptus is necessary but not sufficient for direct activation of a wing-specific decapentaplegic enhancer. Development 126:3669-77

Hewitt, GF *et al.* (1999) Transcriptional repression by the Drosophila Giant protein: cis element positioning provides an alternative means of interpreting an effector gradient. Development 126:1201-1210

Horn, C & Wimmer, EA (2000) A versatile vector set for animal transgenesis. Dev Genes Evol 210:630-7

Hoskins, RA *et al.* (2000) A BAC-based physical map of the major autosomes of Drosophila melanogaster. Science 287:2271-4

Huelsenbeck, JP & Nielsen, R (1999) Variation in the pattern of nucleotide substitution across sites. J Mol Evol 86-93

Inomata, N, Tachida, H & Yamazaki, T (1997) Molecular evolution of the Amy multigenes in the subgenus Sophophora of Drosophila. Mol Biol Evol 14:942-50

Ishihara, K *et al.* (2000) Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in Igf2/H19 imprinting. Genome Res 10:664-71

Jareborg, N, Birney, E & Durbin, R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res 9:815-24

Jeffs, PS, Holmes, EC & As*hb*urner, M (1994) The molecular evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in the Drosophila melanogaster species subgroup. Mol Biol Evol 11:287-304

Jiang, J, Hoey, T., Levine, M. (1991) Autoregulation of a segmentation gene in Drosophila: combinatorial interaction of the even-skipped homeo box protein with a distal enhancer element. Genes Dev 5:265-277

Johnson, NL & Kotz, S (1969) Distributions in Statistics: Discrete Distributions. Houghton Mifflin Company, Boston,

Karlin, S & Altschul, SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A 87:2264-2268

Kassis, JA, Wong, M.L., O'Farrell, P.H. (1985) Electron microscopic heteroduplex mapping identifies regions of the engrailed locus that are conserved between Drosophila melanogaster and Drosophila virilis. Mol Cell Biol 5:3600-3609

Khoury, G, Gruss, P. (1983) Enhancer elements. Cell 33:313-4

Kimura, M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, 367

Kopp, A, Blackman, RK & Duncan, I (1999) Wingless, decapentaplegic and EGF receptor signaling pathways interact to specify dorso-ventral pattern in the adult abdomen of Drosophila. Development 126:3495-507

Korochkin, LI, Panin, V.M., Pavlova, G.V., Kopantseva, M.R., Shostak, N.G., Bashkirov, V.N., Gabitova, L.B., Sergeev, P.V. (1995) A relatively small 5' regulatory region of esterase S gene of Drosophila virilis determines the specific expression as revealed in transgenic experiments. Biochem Biophys Res Commun 213:302-310

Krasney, PA, Carr, C., Cavener, D.R. (1990) Evolution of the glucose dehydrogenase gene in Drosophila. Mol Biol Evol 7:155-177

Krawczak, M, Chuzhanova, NA & Cooper, DN (1999) Evolution of hte proximal promoter region of the mamalian growth hormone gene. gene 237:143-51

Kreitman, M & Ludwig, M (1996) Tempo and mode of even-skipped stripe 2 enhancer evolution in Drosophila. Seminars in Cell and Developmental Biology 7:583-592

Kwiatowski, J, Skarecky, D, Bailey, K & Ayala, FJ (1994) Phylogeny of Drosophila and related genera inferred from the nucleotide sequence of the Cu,Zn Sod gene. J Mol Evol 38:443-54

Langeland, JA & Carroll, SB (1993) Conservation of regulatory elements controlling hairy pair-rule stripe formation. Development 117:585-596

Lawrence, CE, Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208-214

Lawrence, P (1992) The making of a fly: the genetics of animal design. Blackwell Scientific Publications, Oxford ; Boston, 228

Lemeunier, F & As*hb*urner, MA (1976) Relationships within the melanogaster species subgroup of the genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc R Soc Lond B Biol Sci 193:275-94

Lemeunier, F & As*hb*urner, MA (1984) Studies on the evolution of the melanogaster species subgroup of the genus Drosophila (Sophophora). IV. The chromosomes of two new species. Chromosoma 89:343-351

Lewin, B (1994) Genes V. Oxford University Press, Oxford, 1272

Li, W-H (1997) Molecular Evolution. Sinauer Associates, Inc., Sunderland, MA, 487

Li, WH & Graur, D (1991) Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland, MA, 284

Liu, H, Ma, C & Moses, K (1996) Identification and functional characterization of conserved promoter elements from glass: a retinal development gene of Drosophila. Mech Dev 56:73-82

Liu, T, Wu, J & He, F (2000) Evolution of cis-acting elements in 5' flanking regions of vertebrate actin genes. J Mol Evol 50:22-30

Loots, GG *et al.* (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288:136-40

Ludwig, M, Patel, N & Kreitman, M (1998) Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development 125:949-958

Ludwig, MZ, Bergman, C, Patel, N & Kreitman, M (2000) Evidence for stabilizing selection in a eukaryotic cis-regulatory element. Nature 403:564-567

Ludwig, MZ & Kreitman, M (1995) Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. Mol Biol Evol 12:1002-1011

Maizel, JV & Lenk, RP (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc Natl Acad Sci U S A 78:7665-9

McVean, GA & Vieira, J (2001) Inferring Parameters of Mutation, Selection and Demography From Patterns of Synonymous Site Evolution in Drosophila. Genetics 157:245-257

Morgenstern, B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211-8

Morgenstern, B (2000) A space-efficient algorithm for aligning large genomic sequences. Bioinformatics 16:948-9

Moriyama, EN & Hartl, DL (1993) Codon usage bias and base composition of nuclear genes in Drosophila. Genetics 134:847-858

Moriyama, EN, Petrov, DA & Hartl, DL (1998) Genome size and intron size in Drosophila. Mol Biol Evol 15:770-3

Moriyama, EN & Powell, JR (1996) Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol 13:261-277

Moriyama, EN & Powell, JR (1997) Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J Mol Evol 45:378-391

Moses, K, Heberlein, U & Ashburner, M (1990) The Adh gene promoters of Drosophila melanogaster and Drosophila orena are functionally conserved and share features of sequence structure and nuclease-protected sites. Mol Cell Biol 10:539-548

Muller, B & Basler, K (2000) The repressor and activator forms of cubitus interruptus control hedgehog target genes through common generic gli-binding sites. Development 127:2999-3007

Munte, A, Agude, M & Segarra, C (2001) Changes in the recombinational environment affect divergence in the yellow gene of Drosophila. Mol. Biol. Evol. 18:1045-1056

Nei, M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York, 513

Nigro, L, Solignac, M & Sharp, PM (1991) Mitochondrial DNA sequence divergence in the melanogaster and oriental species subgroups of Drosophila. J Mol Evol 33:156-62

Ondek, B, Gloss, L & Herr, W (1988) The SV40 enhancer contains two distinct levels of organization. Nature 333:40-45

Palumbi, S (1989) Rates of molecular evolution and the fraction of nucleotide positions free to vary. J. Mol. Evol. 29:180-7.

Pelandakis, M, Higgins, DG & Solignac, M (1991) Molecular phylogeny of the subgenus Sophophora of Drosophila derived from large subunit of ribosomal RNA sequences. Genetica 84:87-94

Petrov, DA & Hartl, DL (1998) High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol Biol Evol 15:293-302

Petrov, DA & Hartl, DL (1999) Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proc Natl Acad Sci U S A 96:1475-9

Petrov, DA, Lozovskaya, ER & Hartl, DL (1996) High intrinsic rate of DNA loss in Drosophila. Nature 384:346-349

Powell, JR (1997) Progress and prospects in evolutionary biology: the Drosophila model. Oxford University Press, Oxford, 576

Ptashne, M, Gann, A. (1997) Transcriptional activation by recruitment. Nature 386:569-577

Richter, B, Long, M, Lewontin, RC & Nitasaka, E (1997) Nucleotide variation and conservation at the dpp locus, a gene controlling early development in Drosophila. Genetics 145:311-323

Rodríguez-Trelles, F, Tarrío, R & Ayala, FJ (2000) Evidence for a High Ancestral GC Content in Drosophila. Mol Biol Evol 17:1710-1717

Rong, YS, Golic, K.G. (2000) Gene targeting by homologous recombination in Drosophila. Science 288:2013-8

Rorth, P *et al.* (1998) Systematic gain-of-function genetics in Drosophila. Development 125:1049-57

Russo, CA, Takezaki, N & Nei, M (1995) Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol 12:391-404

Sackerson, C (1995) Patterns of conservation and divergence at the even-skipped locus of Drosophila. Mech Dev 51:199-215

Sackerson, C, Fujioka, M & Goto, T (1999) The even-skipped locus is contained in a 16-kb chromatin domain. Dev Biol 211:39-52

Saitou, N & Ueda, S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol Biol Evol 11:504-512

Schneider, TD (1997) Information content of individual genetic sequences. J Theor Biol 189:427-41

Schwartz, S *et al.* (2000) PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10:577-86

Shabalina, SA & Kondrashov, AS (1999) Pattern of selective constraint in C. elegans and C. briggsae genomes. Genet Res 74:23-30

Shabalina, SA, Ogurtsov, AY, Kondrashov, VA & Kondrashov, AS (2001) Selective constraints in intergenic regions of human and mouse genomes. Trends in Genetics 17:373-376

Shibata, H, Yamazaki, T. (1995) Molecular evolution of the duplicated Amy locus in the Drosophila melanogaster species subgroup: concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation. Genetics 141:223-236

Shields, DC, Sharp, PM, Higgins, DG & Wright, F (1988) "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704-716

Small, S, Blair, A., Levine, M. (1992) Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J 11:4047-4057

Small, S, Arnosti, D.N., Levine, M. (1993) Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. Development 119:762-772

Small, S, Kraut, R, Hoey, T, Warrior, R & Levine, M (1991) Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev 5:827-839

Sokal, RR & Rohlf, FJ (1995) Biometry. W. H. Freeman and Co., New York, 850

Stanojevic, D, Hoey, T & Levine, M (1989) Sequence-specific DNA-binding activities of the gap proteins encoded by hunc*hb*ack and *Kr*uppel in Drosophila. Nature 341:331-335

Stanojevic, D, Small, S & Levine, M (1991) Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. Science 254:1385-1387

Stern, DL (2000) Perspective: evolutionary developmental biology and the problem of variation. Evolution 54:1079-1091

Tagle, DA, Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., Jones, R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203:439-455

Takano-Shimizu, T (1999) Local recombination and mutation effects on molecular evolution in Drosophila. Genetics 153:1285-96

Tautz, D (2000) Evolution of transcriptional regulation. Curr Opin Genet Dev 10:575-9

Thompson, JD, Higgins, DG & Gibson, TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-80

Wagner, A (1997) A computational genomics approach to the identification of gene networks. Nucleic Acids Research 18:3594-3604

Wagner, A (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics 15:776-84

Wakeley, J (1994) Substitution-rate variation among sites and the estimation of transition bias. Mol Biol Evol 11:436-42

Wasserman, WW & Fickett, JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278:167-81

Wasserman, WW, Palumbo, M, Thompson, W, Fickett, JW & Lawrence, CE (2000) Human-mouse genome comparisons to locate regulatory sites. Nat Genet 26:225-8

Yang, Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution 11:367-372

Yang, Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587-96

Yang, Z, Goldman, N & Friday, A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11:316-324

Yi, S & Charlesworth, B (2000) Contrasting patterns of molecular evolution of the genes on the new and old sex chromosomes of Drosophila miranda. Mol Biol Evol 17:703-17

Zhu, J, Liu, JS & Lawrence, CE (1998) Bayesian adaptive sequence alignment algorithms. Bioinformatics 14:25-39