Evolution in silico of genes with multiple regulatory modules

on the example of the *Drosophila* segmentation gene *hunchback*

Alexander V. Spirov[1,2]
[1]Computer Science and CEWIT
SUNY Stony Brook
Stony Brook, New York, USA
Alexander.Spirov@cewit.stonybrook.edu
[2]The Sechenov Institute of Evolutionary Physiology & Biochemistry, St.-Petersburg, Russia

David M. Holloway
Mathematics Department
British Columbia Institute of Technology
Burnaby, B.C., Canada
David_Holloway@bcit.ca

***Abstract*— We use in silico evolution to study the generation of gene regulatory structures. A particular area of interest in evolutionary development (evo-devo) is the correspondence between gene regulatory sequences on the DNA (cis-regulatory modules, CRMs) and the spatial expression of the genes. We use computation to investigate the incorporation of new CRMs into the genome. Simulations allow us to characterize different cases of CRM to spatial pattern correspondence. Many of these cases are seen in biological examples; our simulations indicate relative advantages of the different scenarios. We find that, in the absence of specific constraints on the CRM-pattern correspondence, CRMs controlling multiple spatial domains tend to evolve very quickly. Genes constrained to a one-to-one CRM-pattern domain correspondence evolve more slowly. Of these, systems in which pattern domains appear in a particular order in evolution, as in insect segmentation mechanisms, take the longest time in in silico evolutionary searches. For biological cases of this type, it is likely that other selective advantages outweigh the time costs.**

Keywords- computational gene design, gene evolution in silico, genes with multiple regulatory modules, gene expression domains, segmentation patterning, co-linearity principle

I. INTRODUCTION

Spatially patterned gene expression in early embryo development determines cell and tissue types for forming the body plan. Very complex gene regulatory structures have been found as the genes involved in specifying the major axes of the embryo have been increasingly well characterized. In arthropods and vertebrates, the segmentation gene network regulates the formation of early anteroposterior (AP), or head-to-tail, expression patterns [1,2,3]. The fruit fly, *Drosophila*, is the major model organism for studying these genes. Segmentation genes mutually regulate each others` expression, forming the AP patterning of the embryo. This system has uncovered regulatory motifs which are found throughout gene expression networks.

One of the key features found in segmentation is that the genes, in addition to coding regions, have multiple semi-autonomous regulatory elements on the DNA. These cis-regulatory modules (CRMs) have binding sites (BSs) for the regulator molecules (the products of other segmentation genes), and control expression in particular spatial domains. New evidence indicates that the embryonic patterns of all of the *Drosophila* gap genes[1] (the first embryonic segmentation genes expressed) are regulated by multiple CRMs [4]. Large-scale genomic projects, with detailed molecular genetic analysis of the similarities of genes between closely- and distantly-related species, are revealing crucial features of gene functional organization and stimulating new hypotheses on the evolution of body patterning [1-3, 5,6].

CRM-Domain Correspondence: A simple working hypothesis might be that there is a one-to-one correspondence between a CRM and a domain: one CRM, one domain. Some classic molecular genetic studies of pair-rule genes[2] (the segmentation genes expressed after the gap genes) do show such one-to-one correspondence [7-9]. However, this is not universal. One type of departure from one-to-one correspondence is when single CRMs control multiple domains. An extreme example is expression of the *fushi-tarazu* pair rule gene, in which all seven stripe domains are regulated from a single CRM [10]. Other pair-rule genes show a mix, with some stripes one-to-one with a CRM, and other CRMs controlling multiple stripes [7-9]. The other type of departure from one-to-one correspondence is in the actively discussed area of redundancy: many well-known genes for which CRMs have been known for decades are now being found to have 'shadow' elements. These distinct (newly discovered) CRMs functionally duplicate the expression controlled by the well-known (non-shadow) CRMs [11]. Many segmentation gene domains which are not fully redundant still show control from

This work supported by the National Institutes of Health, 2-R01 GM072022.

978-1-4673-1191-5/12/$31.00 ©2012 IEEE

[1] Gap genes are defined by the effect of a mutation in that gene, which causes the loss of contiguous body segments, resembling a gap in the normal body plan.

[2] Pair-rule genes are defined by the effect of a mutation in that gene, which causes the loss of the normal developmental pattern in alternating segments.

two (or even three) CRMs. The role of CRM-domain correspondence in biological development is a very open question.

Order of Domain Appearance: Comparative studies between species show that single genes have been expressed in different ways over evolutionary time, with significant variation in expression domains. For example, the *even-skipped* (*eve*) pair-rule gene is expressed in seven stripes at the blastoderm stage in *Drosophila*, a long-germ band insect[3]. In the long-germ beetle *Callosobruchus*, six pair-rule stripes appear before gastrulation [12]. By contrast, in the short-germ band beetle *Tribolium*, *eve* stripes do not appear simultaneously, and only about two stripes form before gastrulation [13]. In the intermediate-germ beetle *Dermestes*, four *eve* stripes appear before gastrulation. It appears that the evolution of segmentation has added domains to an ancestral simple expression pattern (perhaps several times independently) (Fig. 1). We will refer to a process where stripes (domains) are added sequentially as a *consecutive order rule*.

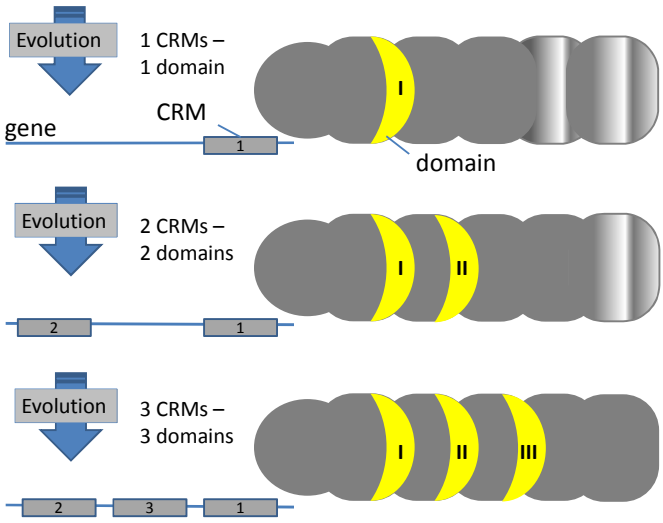

Figure 1. Cartoon illustrating the ordered evolutionary appearance of the expression domains and their CRMs. In particular, the evolutionary transition from short germ to long germ modes of segmentation could proceed in such a way (see text for details).

CRM&Domain Order of Appearance: Part of the domain appearance question is how it correlates with the evolutionary appearance of CRMs in a given gene. As might be expected from the CRM-domain correspondence question, different cases show different degrees of evolutionary correlation between domain and CRM appearance. Many segmentation genes have no evident correlation between domain appearance and CRM appearance. But for some cases, such as the HOM-C and HOX gene clusters [14] the correlation is very striking, suggestive of a *co-linearity principle*, in which the order of expression domains reflects the order of the CRMs on the DNA.

In evolutionary comparisons, related genes (orthologs) tend to maintain what CRM-domain correspondence they do have for closely related species. But the more evolutionary distance between species, the more the CRM-domain correspondence tends to diverge [15-17]. This suggests that there are some general rules for conservation and diversification of the CRM-domain correspondence in biological evolution, with bearing on the order of domain appearance (Cf [17]).

Related to this, segmentation genes within a single species (in single embryos) have a characteristic order of domain appearance. In addition to evolutionary comparisons, the temporal sequence over the course of the minutes to hours of an embryo's development can used to probe the CRM-domain relation.

Comparative genomic and functional studies are not generally sufficient to answer many of these newly formulated evolutionary questions. They need to be complemented by mathematical and computational modeling approaches, in order to quantify the different possible means by which CRMs and domain expression evolve. In this paper, evolutionary computations give insight into how the segmentation network evolved. Evolutionary computations in this area can be very detailed, building on the strong tradition of dynamic modeling of the segmentation gene network [e.g. 18, 19]. Mathematical and computational approaches to gene network evolution (in silico evolution) are becoming increasingly developed within the larger field of systems biology [e.g. 20-22]. In this paper, we develop a new in silico evolution approach for segmentation, and use it to address the issues of *CRM-domain correspondence* and the *CRM&domain order of appearance*.

We find that the in silico evolutionary creation (from a random sequence) of a gene with multiple CRMs, constrained so that each CRM controls its own expression domain, is the most expensive evolutionary mechanism (in terms of time, or number of generations, to solution). On the contrary, if the only constraint is for the computation to match the expression pattern, in silico evolution quickly finds one or several CRMs which control formation of all domains of the gene. Constraints on the CRM-domain correspondence and the domain order of appearance produce time costs in evolution.

II. TEST CASE, THE *HUNCHBACK* GENE

hunchback (*hb*) is a gap gene in the AP segmentation network, one of the first to be expressed in the embryo. It is expressed in an anterior-high pattern which differentiates the head end from the tail end of the embryo. *hb* is a critical component in segmentation, displays the multiple CRMs and domains of interest in this communication, and has been studied extensively with respect to evolution in arthropods. For these reasons, it is an excellent test case for developing an in silico evolution approach for understanding CRM-domain evolution. Figure 2 shows the organization of the *hb* regulatory region into 4 distinct CRMs. Each of these is responsible for different aspects of the *hb* pattern. The

[3] In long germ insects (e.g. *Drosophila*) all segments are specified almost simultaneously within the blastoderm (i.e. prior to gastrulation). In short germ insects (e.g., grasshoppers) only segments of the head are specified in the blastoderm, whereas the remaining segments of the thorax and abdomen form progressively from a posterior growth zone after gastrulation.

oogenesis element (yellow box, Figure 2A) [23] controls the early maternal expression; the proximal element (red) has BSs for Bcd and Hb [23-25] and directs anterior cap expression (Figure 2B); the distal element (green) [26-29] controls the central *hb*-dependent PS4 and posterior stripes (Figure 2A,B). The shadow distal CRM (blue) controls anterior domains of *hb* expression and appears partly redundant with the proximal CRM [4]. (The shadow distal enhancer also contains dominant repression elements that attenuate the activity of the proximal enhancer at the anterior pole [4].) There is a temporal sequence to *hb* expression through these CRMs. At the onset of the 14th nuclear cleavage cycle (cc14A), the stage prior to gastrulation, *hb* is actively expressed in two 'cap' domains, one (extensive) at the anterior and one at the posterior.

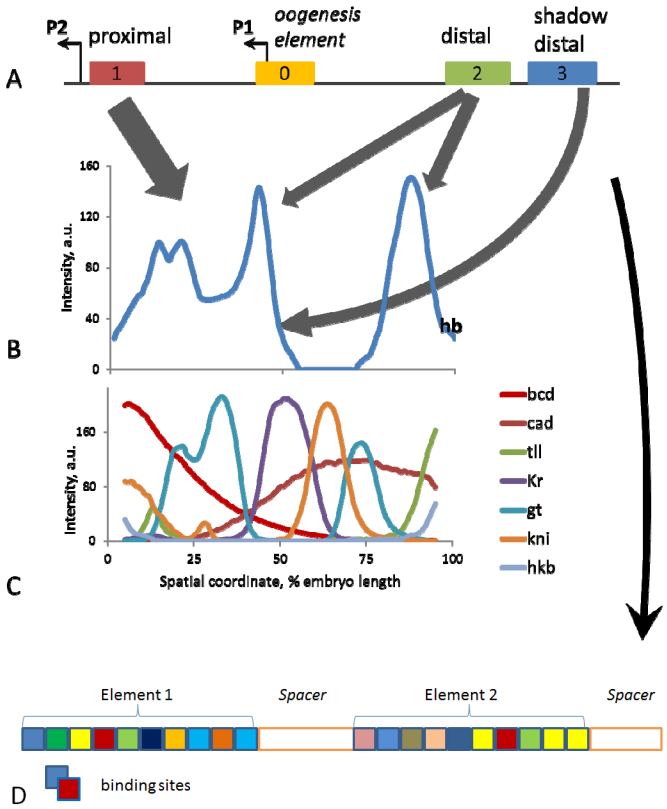

Figure 2. Organization, pattern and regulation of the *Drosophila* segmentation gene *hunchback* (*hb*). A) Organization of the *hb* regulatory region, with four separate autonomous regulatory elements (CRMs). Each regulatory element is a cluster of binding sites (BSs) for the transcription factors directly controlling *hb* expression (depicted as colored bars in D): Bicoid (Bcd), Caudal (Cad), Tailless (Tll), Hunchback (Hb), Giant (Gt), Kruppel (Kr) and Knirps (kni). B) Mature *hb* expression pattern in an early fruit fly embryo (one-dimensional spatial expression profile: fluorescence intensity (proportional to protein concentration) plotted along the main head-to-tail embryo axis; data from BID DB, BDTNP: http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp). C) Spatial profiles of the transcription factors with BSs in the *hb* CRMs. (Data from FlyEx DB: http://urchin.spbcas.ru/flyex). D) represents the BS clusters within CRMs, and the spacer sequences which separate the CRMs.

These domains are controlled by the distal, shadow distal and proximal CRMs (Figure 2A) acting through the P2 promoter [4, 23]. Towards mid cc14A, the distal CRM acts through the P1 promoter to express two new stripes within the earlier domains: the PS4 (parasegment 4) stripe and a posterior stripe. This sequence of events points to highly dynamic regulation, which must be understood in terms of the structure of the *hb* CRMs.

CRM control of expression depends on the binding of at least 8 transcriptional regulators (protein products of other segmentation genes): Bicoid (Bcd), Caudal (Cad), Tailless (Tll), Huckebein (Hkb), Hunchback (Hb; self-activation), Giant (Gt), Kruppel (Kr) and Knirps (kni). The spatial profiles of these factors' expression are shown on Figure 2C. Effects on the *hb* CRMs can be direct, through biding of BSs, and indirect, due to the mutual regulation of these factors. Information on the organization of the *hb* regulatory regions is collected in the HOX pro database [30,31] (http://www.iephb.nw.ru/hoxpro/hb-CRMs.html).

hb, therefore, contains many of the elements found across developmental genes, with complex CRM structure, multiple regulators and multiple expression domains, which change in time. Specifically, at least four distinct and experimentally separable CRMs control the formation of three expression domains, and in a rather redundant fashion (Figure 2A,B), with each of the two anterior domains under the control of two or three CRMs and only the third, posterior domain probably under the control of one CRM. This raises many questions with respect to the evolutionary origin and significance of such a structure. Why does evolution tend to keep such redundant control for *hb* in particular and segmentation genes in general? What control scheme – e.g. redundant vs. unambiguous (one-to-one) – would be faster to evolve from scratch? What is the most probable number of CRMs to evolve for controlling N separate expression domains for a given gene?

III. Our Approach

Our approach to developing dynamic models of gene expression [32,33] can be described as being at a mid-grained level, with respect to biological details. Coarse-grained gene network models treat genes as interacting 'black boxes'. In order to reproduce the CRM structure and dynamics, we must go inside the black boxes and represent the regulatory sequences.

A. Representing hunchback regulation

For in silico evolution, we must find a compact way to represent the regulatory sequences of the gene (Figure 2D), in order to facilitate fast manipulation. We simulate the evolutionary process via genetic algorithms, in analogy to biological DNA evolution (but somewhat abstracted). Figure 3 shows the abstraction of the DNA sequences of the *hb* CRMs to the octal representations used in our computations.

This maintains the element and BS number, but not detailed information on sequence (in bases) and absolute position (i.e. in base pairs). The DNA sequences of the CRMs are translated symbolically into the BSs, which are then represented in octal, with zero being no BS (either within the

CRM, or between, representing spacers), and indices 1 to 7 representing the 7 transcription factors binding the *hb* CRMs.

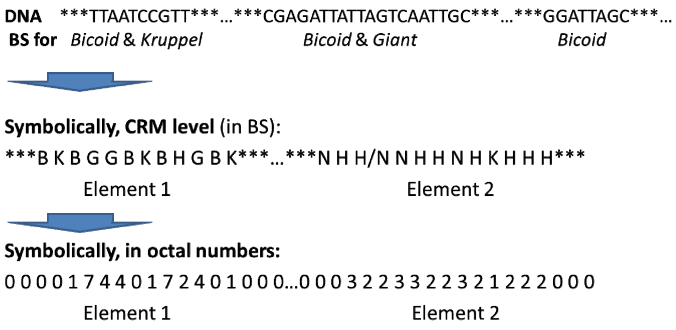

Figure 3. Abstract representation of CRMs as clusters of BSs for transcription factors, delimited by spacers (stretches of zeros=placeholders). Each position on the symbolic string can be occupied either by zero (no BS) or by a number from 1 to 7, representing a BS for one of the seven transcription factors (Bcd, Cad, Tll, Hb, Gt, Kr or kni).

These strings can represent random initial sequences, the wild-type *hb* regulatory sequences, and intermediate stages in between. Genetic algorithms are used to perform crossover operations on the strings to evolve them. String fitness depends on how well they reproduce the experimental data (e.g. the profile in Figure 2B).

The strings are formal representations of the real functional connections controlling the *hb* gene via the network of transcription factors (including the Hb factor itself; Figure 2). At each in silico evolution generation, candidate strings are used to solve a reaction-diffusion model of *hb* gene expression. Expression of the gene, C ≡ [Hb], under control of a given CRM is quantitatively described by the following reaction-diffusion equation:

$$\frac{dC}{dt} = D\frac{\partial^2 C}{\partial x^2} + \mathrm{R}\sigma\left(\sum_{i=1}^{n} S_i - \mathrm{h}\right) - \lambda C, \qquad (1)$$

where S_i is the strength of the i-th activator BS, n is the number of the activator BSs in a given CRM, D is a diffusion coefficient, h represents regulatory input from ubiquitous factors, and **λ** is a decay coefficient.

The strength S_i is a sum of three terms: the local concentration A_i, the short-range co-activation term, and the short-range repression (quenching) term:

$$S_i = A_i + \alpha_i\left(\sum_{k=1}^{m} A_k\right) - \sum_{j=1}^{l} R_j,$$

where A_k is local concentration of k-th activator, R_j is local concentration of j-th repressor, α_i is the co-activation coefficient, m is the amount of the neighbor activator BSs and l is the amount of the neighbor repressor BSs. $\sigma(x)$ is a sigmoid regulation-expression function:

$$\sigma(x) = \sqrt{\frac{x^2}{(1+x^2)}}.$$

The model takes into consideration the sum of strengths of all activator BSs in a given CRM. This activation strength is then modified to consider i) repression of BSs by quenching from neighboring repressor BSs, and ii) co-activation by neighboring activator sites. The algorithm is depicted in Figure 4.

Each candidate (evolving) string is put into the reaction-diffusion model. The fitness of the string is determined by how well its model *hb* pattern matches the experimental data (e.g. the spatial profile in Figure 2B).

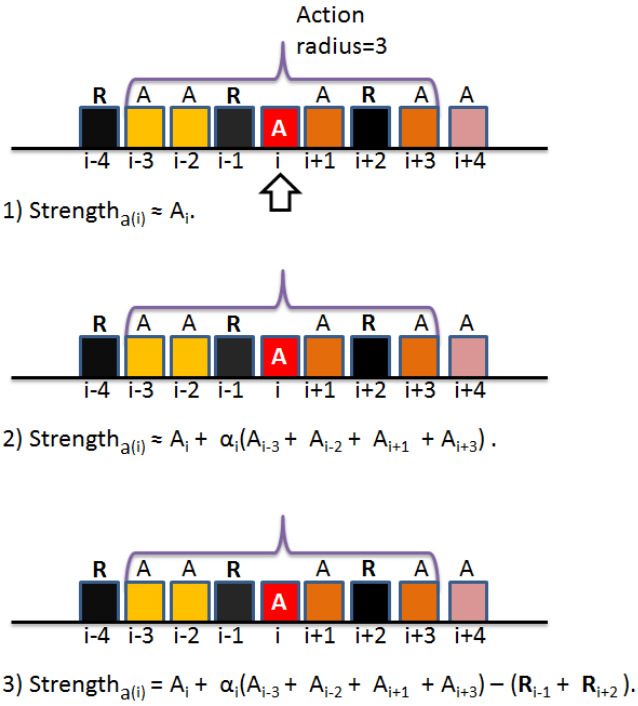

Figure 4. The 3-step algorithm to sum the activation strengths for a given activator BS, taking into account both repression via quenching and co-activation from neighboring BSs. For simplicity, we assume that both repression and co-activation are short-range, limited to three neighboring sites. 1) local BS strengths are tallied; 2) neighboring activation is added (co-activation); 3) neighboring repression is added (quenching).

The set of PDEs (1) was solved numerically by Euler's method [34]. We minimized the sum of squared residuals, using observed values from the expression patterns shown in Figures 2B and 5.

In this report, we use mature (mid cc14A) *hb* mRNA data for fitting the models. We discuss six evolutionary scenarios below. In the first two cases, fits are done to complete, 3-domain *hb* pattern (Figure 2B); in the last 4 cases, the

appearance of separate domains is modeled (Figure 5). We do not fit the *hb* pattern at the ends of embryo (anterior-most 10% and posterior-most 5%); these regions are controlled by the terminal *huckebein* gene, which is not in the model (is not one of the core, trunk gap genes). In each of the six in silico evolution cases, we have performed 100 independent runs, for sound statistics. The following parameters were the same in all computations: 1/3 of the population is replaced via a truncation strategy each round; point mutation rate (P/bit)/generation = 0.01; single-point recombination rate (P/bit)/generation = 0.001. CRM length is 16 BS (=positions) and spacers are 4 positions. We use different population sizes for different experiments, since (as we found in preliminary computations) harder tasks require larger populations for efficient searches. In all computations in this paper, the following four kinetic parameters of the PDE (1) were kept fixed: R = 120, h = 1.1, D = 0.2, λ = 0.5625.

B. *Biologically reasonable constraints on in silico evolution*

In evolving a gene's regulatory sequence, we expect the speed and efficiency of the process to depend substantially on the level of detail to which we match the model results to the biological data. As a first step, we can fit the model to the complete three-domain *hb* profile (as in Figure 2B). A further refinement can be to require the in silico evolution to find each of the *hb* domains sequentially, from first to last (as shown in Figure 5). Such constraints can begin to show the regulatory structure needed to produce such sequences, and give closer insight into the biological problem, in which the gene has evolved with multiple regulatory elements, and the CRM-domain correspondence ranges from one-to-one to uncorrelated.

Considering that the level of fitting can affect the solutions that are evolved and to model different evolutionary possibilities for the *hb* gene, we ran a series of computational experiments with different levels of constraint. We will report on results from six different scenarios, starting from the most loosely defined search, and adding constraints:

Case 1) In the simplest case, the complete *hb* pattern is used for fitting. CRMs are free to evolve, but the number of CRMs is constrained (one, two, or three depending on the computation). There is no requirement for CRM-domain correspondence: solutions are allowed in which only one (of the three) CRMs controls formation of all domains; in which CRMs are one-to-one with domains; or in which the CRMs share control of domains.

Case 2) In this case, CRMs are still free to evolve, but we constrain the order of finding expression domains. In this way, the in silico evolution can model the order of domain appearance in biological evolution. Here, we follow the sequential order rule (discussed in the Introduction): i.e. the 1st domain must be found first, then the 2nd, then the 3rd. But we do not control exactly how the CRMs govern domain formation.

Only in the first two cases do we allow control of multiple domains by single CRMS, redundancy (multiple CRM control of single domains), or even non-functional CRMs. In the last four cases we set constraints on CRM-domain correspondence.

Case 3) In this case, in addition to sequential appearance of the domains, we map these to particular CRMs one-to-one: the 1st domain is controlled by any of the three CRMs; the 2nd domain is controlled by one of remaining two CRMs, and the 3rd domain is controlled by the final CRM. This is a parallel search of the expression domains and CRMs.

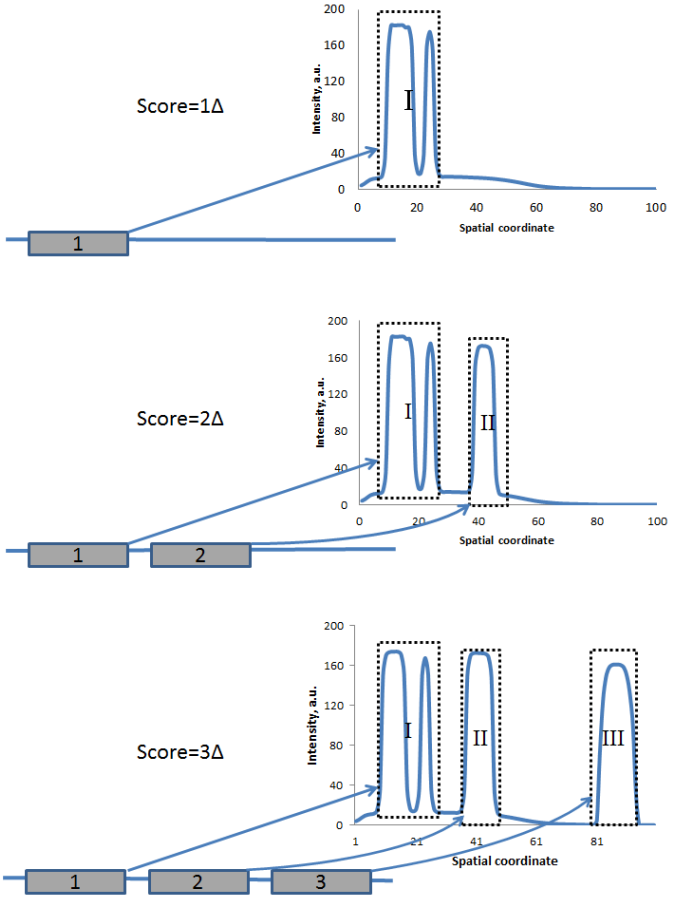

Figure 5. One possible simplified scenario for the evolutionary origin of *hb* gene organization: sequential appearance of domains; one-to-one CRM–domain correspondence. A single CRM could insert into an ancestral gene (with no prior elements). Due to increased fitness for each such insertion, the gene would build up CRMs (maximum three elements here). Gene organization and the corresponding patterns of gene expression are shown schematically. Starting from a single CRM with fitness score = Δ, finding of the 2nd CRM by the evolutionary search would double the score (2Δ), and so on sequentially to completion.

Case 4) Here we constrain the order of finding the CRMs. The order of domain appearance is free, but bound one-to-one with the CRMs (it is the converse of case 3 above). First, the most 3' CRM must be found, then its nearest neighbor, and so on. The domain appearance order is not constrained, but is searched in parallel with the CRMs. In this scenario, the 1st CRM can control any one of the three domains; the 2nd CRM

controls one of the remaining two domains; the final CRM controls the final domain. We do not know of biological examples of this type of evolution, but test it here for comparison with Case 3.

Case 5) This case is an evolutionary search according to the *co-linearity principle*. This has a strictly ordered one-to-one correspondence, in which the 1st CRM must control the 1st domain, the 2nd CRM the 2nd domain, and so on. This is the scenario in Figure 5, and is biologically seen with the very important HOX and HOM-C gene clusters. It is both a consecutive search of CRMs and a consecutive search of the expression domains.

Case 6) For comparison, in this case we have a parallel search of CRMs and parallel search of the expression domains: any one of the three domains can be controlled by one of the CRMs, arbitrarily chosen. This is not likely close to biological reality, but we study it here for completeness.

IV. Results and Discussion

Here we present computational results from the different cases introduced above. The evolutionary task is to find the three domain *hb* pattern; we are interested in characterizing the degree to which the different constraint scenarios affect the speed of the evolutionary search.

A. *single CRMs tend to control multiple domains when there are no constraints on CRM-domain correspondence*

Case 1 (same numbering as in section III.B): We begin with the simplest case, in which the complete three-domain pattern is fit without any constraints on CRM-domain correspondence. The starting point is random sequences in the areas of prospective/future CRMs; spacer regions (zeros) between CRMs are kept intact. We performed runs with three, two and one CRM.

In this case with free evolution of CRMs, good solutions are found within a few tens of thousands of evaluations (population size, 6000). This is a very fast evolutionary search for the complexity of the pattern. In comparison, a search on a less detailed (coarse-grained) model of four segmentation genes (including *hb*) took hundreds of millions of evaluations to converge on a good solution, using Simulated Annealing [35]. (But Genetic Algorithms are more efficient for the problem [36].) In terms of speed, the one and two CRM computations are comparable to each other (mean±std.dev. number of evaluations to convergence 38,679±12,363 and 37,728±10,573, respectively) and slightly faster than with three CRMs (43,322±10,258 evaluations).

Case 2: Here, finding CRMs is still free, but the domains must be found in sequential order. This, in broad terms, may be the way in which the *hb* pattern evolved in nature. This constraint on order of domain appearance makes the task about five times harder for the in silico evolution process, compared to Case 1: number of evaluations = 211,674±36,771 (population = 6000).

Case 1 and 2 computations indicate that if speed of evolution is an important factor, gene regulatory structures may tend to favor single CRMs controlling multiple genes. For instance, it is possible that cases like the *fushi-tarazu* gene, with 7 stripes controlled by one CRM, evolved quicker (and with less constraints) than other pair-rule genes (in which CRMs typically control one or two stripes).

Careful analysis of the good solutions shows some interesting trends. Typical solutions have only one CRM (sometimes two) controlling the expression pattern. In only a few percent of the solutions (2 out of 100) are all three CRMs involved in the patterning - an example is shown in Figure 6. Such multiple control is one aspect of real *hb* control, e.g. the action of the distal element in Figure 2A,B. It is intriguing how rarely solutions with all three functional CRMs appear. Such functionally redundant organization may provide selective advantages and may provide robustness to *hb* patterning.

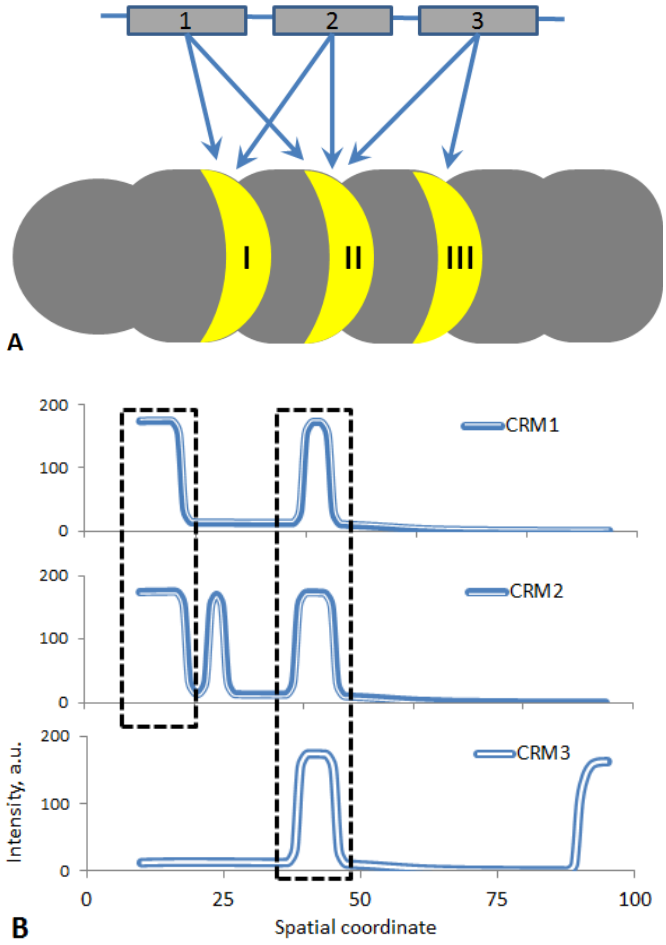

Figure 6. A solution of the *hb* gene problem with all three CRMs participating in patterning the anterior domains (Case 2 scenario). A) CRM-domain diagram. B) Solution of the hb gene problem for each of three CRMs, with redundancies outlined by the dashed boxes.

B. *It takes substantially more time to evolve a CRM for each domain*

Cases 3 – 6 have one-to-one constraints on CRM-domain correspondence.

Case 3: In addition to sequential appearance of domains (Case 2), we constrain one-to-one CRM control of the three domains. In this scenario, any of the potential CRMs can control a particular domain; in Case 5 we explore the constraint that the CRM order on the DNA must match the spatial order of the domains.

To our surprise, this scenario produced the most time-consuming in silico evolution computations. It was also the only case with success rate <100%: only 54% of the runs achieved the desired solution (*hb* fit) within 6 million evaluations. Number of evaluations was 4,366,084±1,138,597 (mean±std.dev.); population, 18,000. The evolutionary constraints in this case make it about a one hundred fold harder problem than Case 1.

This scenario, with single CRMs co-opted into the genome corresponding one-to-one with newly appearing domains, has likely occurred several times in arthropod evolution, and is reminiscent of the short- to long-germ band transition in segmentation mechanisms, via intermediate forms [17]. That this case appears to be one of the most time-expensive scenarios suggests that such transitions have real evolutionary (fitness) importance.

Case 4: This is the converse of Case 3. CRMs must appear sequentially, then domains are bound to them one-to-one. We do not know of biological examples at this point, but compute this scenario for comparison. This evolutionary scenario is relatively fast, on the order of the Case 2 speeds. Number of evaluations: 229,912±76,245 (population, 2500). We believe the efficiency of this case is due to domains being searched in parallel with the CRMs.

Case 5: This is the colinear case, where the one-to-one CRM-domain correspondence includes both order of domain appearance and CRM order on the DNA. Biological examples of this are not ubiquitous, but the cases which do display this, such as the HOX cluster, are quite important and famous. The evolutionary computations with this scenario are quite slow, but are also quite reproducible (with a small standard deviation). Number of evaluations: 1,373,246±198,698 (population, 18,000). Biological examples such as the HOX cluster are extremely well conserved through evolution, comparing between species. It is possible that the reproducibility of this search is associated with the stability and conservation of these regulatory structures.

Case 6: For completeness, we computed the scenario where CRMs and domains were searched in parallel. Order does not matter, but there is a one-to-one CRM-domain correspondence. In terms of the evolutionary computations, this scenario shows comparable efficiency to Case 4.

C. Comparison with the known evolutionary biology of hb

hb and other key early segmentation genes were first discovered in *D. melanogaster* and have since been studied in many other species (see recent review [37]). Current information supports the hypothesis that the *hb* gene has been independently recruited (co-opted) into the segmentation ensemble several times in arthropod evolution, and that it has been recruited from other gene networks (controlling neurogenesis, mesoderm specification, etc.). Even when *hb* is functioning as a gap gene, its segmentation patterning can differ dramatically between species.

For instance, the moth midge *Clogmia albipunctata* displays *Drosophila*-like *hb* (*Calb-hb*) patterning in the anterior (see Fig. 5C, left panel, from [38]). But the posterior domain forms substantially later, after gastrulation, and is shifted to the posterior [38,39]. As a result, the embryos have only six-stripe (instead of seven) pair-rule patterns. As a lower dipteran, it is likely that *Calb-hb* patterning is more primitive and resembles the ancestral Drosophilid patterning. Unfortunately nothing is known about the regulatory organization of the moth midge *hb* gene. We believe that more detailed computational evolutionary experiments could stimulate further, more precise molecular studies in the moth midge on the evolution of segmentation. This approach is especially promising since the moth midge segmentation gene network has recently come under systematic experimental and theoretical analysis [37-39].

V. CONCLUSIONS

Our evolutionary computations indicate that it can be roughly a hundred times easier to find one CRM governing formation of all three domains of the *hb* pattern, than to find three separate CRMs independently controlling separate *hb* domains (one CRM – one domain). This suggests that genes which show multiple domain control by single CRMs may have evolved quite quickly. These computational results produce features of real biological *hb* patterning ([4]; Figure 2A), including redundant control of expression domains, which may confer robustness to external variability and internal noise [4].

In general, there is abundant evidence that evolution of autonomous CRMs is responsible for many cases of morphological evolution [17]. The computational approach outlined here will help to understand the correspondence between CRM evolution and domain appearance (morphological effect). As shown here, different cases of the CRM-domain dependence lead to different evolutionary costs, and help to understand how a number of regulatory motifs have arisen in evolution, and what their particular advantages might be. This approach presents a new method for quantifying these evolutionary processes for the modern study of evolutionary development (evo-devo).

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the paper.

REFERENCES

[1] M. D. Schroeder, C. Greer, and U. Gaul, „How to make stripes: deciphering the transition from non-periodic to periodic patterns in Drosophila segmentation,“ Development, vol. 138, pp. 3067–3078, 2011.

[2] A. D. Peel, A. D. Chipman, and M. Akam, „Arthropod segmentation: beyond the Drosophila paradigm," Nat. Rev. Genet., vol. 6, pp. 905–916, 2005.
[3] W. G. M. Damen, „Evolutionary conservation and divergence of the segmentation process in arthropods," Dev. Dynam., vol. 236, pp. 1379–1391, 2007.
[4] M. W. Perry, A. N. Boettiger, and M. Levine, „Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo," Proc. Nat. Acad. Sci. USA, vol. 108, pp. 13570–13575, 2011.
[5] B. Z. He, A. K. Holloway, S. J. Maerkl, and M. Kreitman, „Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules," PLoS Genet., vol. 7, e1002053, 2011.
[6] R. K. Bradley, X-Y Li, C. Trapnell, S. Davidson, L. Pachter, et al., „Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species," PLoS Biol., vol. 8, e1000343, 2010.
[7] K. Harding, T. Hoey, R. Warrior, and M. Levine, „Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila," EMBO J., vol. 8, pp. 1205–1212, 1989.
[8] G. Riddihough, and D. Ish-Horowicz, „Individual stripe regulatory elements in the Drosophilahairy promoter respond to maternal, gap, and pair-rule genes," Genes. Dev., vol. 5, pp. 840–854, 1991.
[9] M. Klingler, J. Soong, B. Butler, and J. P. Gergen, „Disperse versus compact elements for the regulation of runt stripes in Drosophila," Dev. Biol., vol. 177, pp. 73–84, 1996.
[10] C. Tsai and P. Gergen, „Pair-rule expression of the Drosophila fushi tarazu gene: a nuclear receptor response element mediates the opposing regulatory effects of runt and hairy," Development, vol. 121, pp. 453-462, 1995.
[11] S. Barolo, „Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy," BioEssays, vol. 34, pp. 135-141, 2011.
[12] N. H. Patel, B. G. Condron, and K. Zinn, Pair-rule expression patterns of even-skipped are found in both short and long germ beetles. Nature, vol. 367, pp. 429-434, 1994.
[13] C. Eckert, M. Aranda, C. Wolff, and D. Tautz, „Separable stripe enhancer elements for the pair-rule gene hairy in the beetle Tribolium," EMBO Rep., vol. 5, pp. 638–642, 2004.
[14] C. Kenyon, "If birds can fly, why can't we?: homeotic genes and evolution," Cell, vol. 78, pp. 175-180, 1994.
[15] M. Z. Ludwig, A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan, et al., „Functional Evolution of a cis-Regulatory Module," PLoS Biol., vol. 3, e93, 2005.
[16] Y. Goltsev, W. Hsiong, G. Lanzaro, and M. Levine, „Different combinations of gap repressors for common stripes in Anopheles and Drosophila embryos," Dev. Biol., vol. 275, pp. 435-446, 2004.
[17] S. B. Carroll, "Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution," Cell, vol. 134, pp. 25–36, 2008.
[18] J. Jaeger, "Modelling the Drosophila embryo," Mol. BioSyst., vol. 5, pp. 1549–1568, 2009.
[19] J. B. Hengenius, M. Gribskov, A. E. Rundell, C. C. Fowlkes, and D. M. Umulis, "Analysis of Gap Gene Regulation in a 3D Organism-Scale Model of the Drosophila melanogaster Embryo," PLoS ONE, vol. 6, e26797, 2011.
[20] P. Francois, and V. Hakim, "Design of genetic networks with specific functions by evolution in silico," Proc. Nat. Acad. Sci. USA, vol. 101, pp. 580-585, 2004.
[21] P. François, V. Hakim, and E. Siggia, "Deriving structure from evolution: metazoan segmentation," Nature/EMBO J. Mol. Syst. Biol., vol. 3, 154, 2007.
[22] M. B. Cooper, M. Loose, and J. F. Y. Brookfield, "The Evolutionary Influence of Binding Site Organisation on Gene Regulatory Networks," Biosystems, vol. 96, pp. 185-193, 2009.
[23] C. Schroder, D. Tautz, E. Seifert, and H. Jackle, „Differential regulation of the 2 transcripts from the Drosophila gap segmentation gene Hunchback," EMBO J., vol. 7, pp. 2881–2887, 1988.
[24] W. Lukowitz, C. Schroder, G. Glaser, M. Hulskamp, and D. Tautz, „Regulatory and coding regions of the segmentation gene Hunchback are functionally conserved between Drosophila-virilis and Drosophila-melanogaster," Mech. Dev., vol. 45, pp. 105–115, 1994.
[25] J. S. Margolis, M. L. Borowsky, E. Steingrimsson, G. W. Shim, J. A. Lengyel, et al., „Posterior stripe expression of Hunchback is driven from 2 promoters by a common enhancer element," Development, vol. 121, pp. 3067-3077, 1995.
[26] W. Driever, and C. Nusslein-Volhard, „The Bicoid protein is a positive regulator of Hunchback transcription in the early Drosophila embryo," Nature, vol. 337, pp. 138-143, 1989.
[27] W. Driever, G. Thoma, and C. Nusslein-Volhard, „Determination of spatial domains of zygotic gene-expression in the Drosophila embryo by the affinity of binding-sites for the Bicoid morphogen," Nature, vol. 340, pp. 363–367, 1989.
[28] G. Struhl, K. Struhl, and P. M. Macdonald, "The gradient morphogen Bicoid is a concentration-dependent transcriptional activator," Cell, vol. 57, pp. 1259-1273, 1989.
[29] Q. Gao and R. Finkelstein, "Targeting gene expression to the head: the Drosophila orthodenticle gene is a direct target of the Bicoid morphogen," Development, vol. 125, pp. 4185-4193, 1998.
[30] A.V. Spirov, T. Bowler. and J. Reinitz, „HOX-Pro: A Specialized Database for Clusters and Networks of Homeobox Genes," Nucleic Acids Res., vol. 28, pp. 337-340, 2000.
[31] A. V. Spirov, M. Borovsky, and O. A. Spirova, „HOX Pro DB: The functional genomics of hox ensembles," Nucleic Acids Res., vol. 30, pp. 351-353, 2002.
[32] A. V.Spirov and D. M. Holloway, „Design of a dynamic model of genes with multiple autonomous regulatory modules by evolutionary computations," Procedia Comp. Sci., vol. 1, pp. 1005-1014, 2010.
[33] D. M. Holloway and A. V. Spirov, „Genetic Algorithm inspired by the mechanisms of retroviral recombination and its application to the design of genes by evolution in silico," in Genetic Algorithm / Book 2, Olympia Roeva Ed. 26 P., InTech Open Access Publisher, 2012.
[34] A. V. Spirov and D. M. Holloway, „The effects of gene recruitment on the evolvability and robustness of gene networks," in Advances in Computational Algorithms and Data Analysis, S-I Ao, B Rieger, and S-S Chen, Eds. Springer 2008, pp. 29-50.
[35] W. H. Press, B. P. Flannery, S.A. Teukolsky, and W.T. Vetterling, Numerical Recipes, Cambridge University Press, Cambridge, 1988.
[36] Manu, S. Surkova, A. V. Spirov, V. V. Gursky, H. Janssens, A.-R. Kim, O. Radulescu, C. E. Vanario-Alonso, D. H. Sharp, M Samsonova, and J Reinitz, „Canalization of Gene Expression and Domain Shifts in the Drosophila Blastoderm by Dynamical Attractors," PLoS Comput. Biol., vol. 5, e1000303, 2009.
[37] J. Jaeger, „The gap gene network," Cell. Mol. Life Sci., vol. 68, pp. 243-274, 2011.
[38] K. B. Rohr, D. Tautz, and K. Sander, „Segmentation gene expression in the mothmidge Clogmia albipunctata (Diptera, psychodidae) and other primitive dipterans," Dev. Genes Evol., vol. 209, 145-154, 1999.
[39] M. García-Solache, J. Jaeger, and M. Akam, „A systematic analysis of the gap gene system in the moth midge Clogmia albipunctata," Dev. Biol, vol. 344, pp. 306-318, 2010.