ABSTRACT

Title of dissertation: LARGE SYSTEMS OF MANY INTERCONNECTED DYNAMICAL UNITS:
GENE NETWORK INFERENCE,
EPIGENETIC HERITABILITY,
AND EMERGENT BEHAVIOR IN OSCILLATOR SYSTEMS

Wai Lim Ku, Doctor of Philosophy, 2014

Dissertation directed by: Professor Edward Ott
Department of Electrical and Computer Engineering
Professor Michelle Girvan
Department of Physics

In this thesis, which consists of three parts, we investigate problems related to systems biology and collective behavior in complex systems.

The first part studies genetic networks that are inferred using gene expression data. Here we use established transcriptional regulatory interactions (TRIs) in combination with microarray expression data from both *Escherichia coli* (a prokaryote) and *Saccharomyces cerevisiae* (a eukaryote) to assess the accuracy of predictions of coregulated gene pairs and TRIs from observations of coexpressed gene pairs. We find that highly coexpressed gene pairs are more likely to be coregulated than to share a TRI for Saccharomyces cerevisiae, while the incidence of TRIs in highly coexpressed gene pairs is higher for Escherichia coli. The data processing inequality (DPI) of information theory has previously been applied for the inference of TRIs. We consider the case where a transcription factor gene is known to regulate two genes

(one of which is a transcription factor gene) that are known not to regulate one another. According to the DPI if certain conditions hold, the non-interacting gene pairs should have the smallest mutual information among all pairs in the triplets. While we observe that this is sometimes the case for *Escherichia coli*, we find that it is almost always not the case for *Saccharomyces cerevisiae*, thus indicating that the assumed conditions under which the DPI was derived do not hold.

The second part of this dissertation is related to the dynamical process of epigenetic heritability. Epigenetic modifications to histones may promote either activation or repression of the transcription of nearby genes. Recent experimental studies show that the promoters of many lineage-control genes in stem cells have bivalent domains in which the nucleosomes contain both active (H3K4me3) and repressive (H3K27me3) marks. Here we formulate a mathematical model to investigate the dynamic properties of bivalent histone modification patterns, and we predict some interesting and potentially experimental observable features.

The third part of this dissertation studies dynamical systems in which a large number N of identical Landau-Stuart oscillators are globally coupled via a mean-field. Previously, it has been observed that this type of system can exhibit a variety of different dynamical behaviors including clumped states (in which each oscillator is in one of a small number of groups for which all oscillators in each group have the same state which is different from group to group), as well as extensive chaos (a situation in which all oscillators have different states and the macroscopic dynamics of the mean field is chaotic). One of our foci is the transition between clumped states and extensive chaos as the system is subjected to slow adiabatic parameter change.

We observe and analyze explosive discontinuous transitions between the clumped states and the extensively chaotic states. Also, we study the fractal structures of the extensively chaotic attractors.

LARGE SYSTEMS OF MANY INTERCONNECTED DYNAMICAL UNITS: GENE NETWORK INFERENCE, EPIGENETIC HERITABILITY, AND EMERGENT BEHAVIOR IN OSCILLATOR SYSTEMS

by

Wai Lim Ku

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Edward Ott, Chair/Advisor
Professor Michelle Girvan, Co-Advisor
Professor Thomas M. Antonsen Jr.
Professor Thomas E. Murphy
Professor Wolfgang Losert (Dean's Representative)

Dedication

To my grandmother, Dai-Tai Chan, and my parents, Hoi-Tong Lee and Kin-Shun Ku.

Acknowledgments

First, I would like to thank my advisors, Profs. Edward Ott and Michelle Girvan, for giving me the opportunities to work on these interesting research projects. I enjoy a lot working with them. They were always willing to patiently help me and support me on different issues: not only my study, research, but also my career development. I appreciate all the useful and insightful advice they have given me along the way. From them I learnt how I should approach a research problem and what the attitude of doing research should be. Further, I would like to thank Prof. Ott for his many other participants in my work like editing paper and dissertation, and many other helps throughout these years. Thank you also to Prof. Thomas Antonsen, Prof. Thomas Murphy, and Prof. Wolfgang Losert for taking the time to be members of my thesis committee.

I would like to thank my collaborator, Prof. Francesco Sorrentino, for his support and help during my Ph.D. study. In particular, he was always willing to chat with me about my research and study. In addition, I would like thank my collaborator, Prof. Guo-Cheng Yuan, for his time and efforts in our discussion about the research in computational biology.

I also wish to thank Zhi-Xin Lu, Nicholas Mecholsky, Geet Duggal, Kimberly Glass, Matthew Kretschmer, Young-Noh Yoon, Mark Herrera, Waishing Lee, Ming-Jer Lee for providing timely assistance. A special thanks goes to Shane Squires who has helped and encouraged me a lot during my PhD study. He is always willing to discuss with me about my research and teach me English. Thanks to Ching Pui

Hung, kwan-Yuet Ho, Tak Chu Li, Joung-hoon Beh, King-Lam Hui, Kan-Leung Cheng, for their friendships. I also owe my thanks to Jane Hessing for her helps. She is always willing to work with me to solve all those bureaucratic problems.

Lastly, I would like to thank my parents and grandmother for their unconditional love. Without their supports, I would not able to have the opportunities to study and finish my PhD in Maryland. Thank you to my girlfriend, Yuan Li, for her continuous support along the way. She always helps me in many daily issues.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DPI	Data processing inequality
eqf	equal frequency
eqw	equal width
pr	precision
re	recall
TF	Transcription factor
TRI	Transcriptional regulatory Interaction

Chapter 1: Introduction

This thesis is divided into three parts. The first and second parts apply tools from non-linear dynamics and complex networks to questions in systems biology. For the first part, we consider gene networks inferred using experimental gene expression data and an information theoretical approach. We investigate how these expression inferred networks compare to other, much higher quality and more expensive experimentally inferred networks obtained by other means. For the second part, we formulate a model to study pattern formation and heritability in the dynamics of epigenetic processes. In the third part, we study the emergent dynamical behaviors in large interconnected oscillator systems. In particular, we consider a mean-field coupled system of Landau-Stuart oscillators, and study the dynamical states of this system and the transition between these states.

1.1 Chapter 2: Interpreting Patterns of Gene Expression

In cells, genes can be transcribed to mRNA molecules, which are then translated to proteins. Some of the proteins can, in turn, regulate some of the genes. Thus genes can be regarded as interacting with each other via a network. It is important to recover the structure of gene interaction networks in order to help

elucidate underlying genetic regulatory processes. We refer this as 'reconstruction of gene networks'.

Mutual information and related metrics have been applied to gene expression data to infer previously unknown interactions in genetic networks. In Chapter 2, we investigate the implications of high mutual information between two genes in their occurrence of expression. Such high mutual information may imply that one of the genes directly regulates the other gene, or it is possible that they are both regulated by a third gene, or even something else entirely. To address this issue, we consider *E. coli* and *S. cerevisiae* for which reliable regulatory gene interactions have previously been determined. We demonstrate that a gene pair with high mutual information does not necessarily imply a transcriptional regulatory link between them. We show that these misleading correlations commonly occur due to joint regulation by a third gene. The work in this chapter was published in PLoSOne in 2012 [1].

1.2 Chapter 3: Modeling Dynamics of Histone Modifications

Epigenetic factors are mechanisms which can change gene activity heritably without changing the underlying DNA sequence. In a cell nucleus, DNA is wrapped around proteins called histones, which can undergo several kinds of epigenetic modifications, either activating or repressing the expression of target genes. Modified histones are believed to catalyze similar modifications in nearby histones, leading to complex dynamics and pattern formation. DNA regions wrapped around histones with multiple modifications control certain steps in embryonic development.

However, the mechanisms by which this occurs remain unclear.

In Chapter 3, we proposed and studied a dynamical model of histone modification which, unlike previous models, treats multiple types of modification simultaneously. Our model predicts interesting, potentially experimentally observable features of bivalent domains (such as their slow formation but rapid decay) and suggests that cell differentiation may be due to the existence of several dynamical attractors. The work in this chapter was published in PLoSOne in 2013 [2].

1.3 Chapter 4: A Mean-Field Coupled System of Landau-Stuart Oscillators

Understanding the emergence of macroscopic collective behavior and its nature is important in many scientific disciplines. Here we considered, as a model of such behavior, an interconnected system of oscillators. In this system, all oscillators are assumed to be identical and globally coupled via a mean field. Previously, it has been observed that this type of system can exhibit a variety of different dynamical behaviors, including "clumped states" and extensively chaotic states. For a clumped state attractor, there is a small number of different clumps, and oscillators in each clump behave identically. On the other hand, in extensively chaotic states, each oscillator behaves differently and moves chaotically. In Chapter 4, we investigate the dynamics of mean-field coupled systems of oscillators and how they change with continuous variation of the coupling strength between oscillators. We also study the fractal structures of the extensively chaotic attractors. The qualitative results of

our study should also apply to many types of systems in which a large number of dynamical units are globally coupled.

Chapter 2: Interpreting Patterns of Gene Expression: Signatures of Coregulation, the Data Processing Inequality, and Triplet Motifs

This work in this chapter was published in PLoSOne in 2012 [1].

2.1 Introduction

If two genes share a transcriptional regulatory interaction (TRI), one or both of them must be a transcription factor gene (TF gene) which can produce a protein called a transcription factor (TF) that regulates the mRNA expression of the other gene. The collection of genes and TRIs work as a dynamic network enabling cells to function and cope with changes in their environment [3]. The increased availability of high-throughput gene expression data has lead to a variety of approaches for inferring TRIs [4–8]. A typical assumption of these approaches is that strongly correlated mRNA expression profiles (coexpressed profiles) indicate TRIs between two genes if one or both genes is a TF gene. More sophisticated methods of inferring TRIs integrate gene expression with other information, e.g. position weight matrices from sequence motif analysis, as in [9]. Here, we study the use of gene expression

alone in determining TRIs. In particular, we focus on the z-score metric used in the CLR algorithm (described in section 2.2). This metric has been argued to give good performance in inferring TRIs [4]. On the other hand, it has been shown in the case of *Saccharomyces cerevisiae* that gene pairs with a high degree of positive coexpression according to the Pearson correlation coefficients may indicate coregulation by TFs [10]. This raises the question of how to biologically interpret high levels of coexpression between gene pairs, particularly in the case of non-time-course data. In this study, we use publicly available prokaryotic bacterium *Escherichia coli* (*E.coli*) and eukariotic *Saccharomyces cerevisiae* (yeast) microarray expression data (these data are collected under different experimental conditions) along with established TRIs to evaluate the accuracy of different predicted gene pairs. In particular, we consider gene pairs that are coexpressed above a selected threshold level. By comparing these gene pairs to the TRIs in the established networks, we obtain estimates of the precision and recall for the prediction that these pairs are TRIs and the alternate prediction that these pairs are coregulated. Our goal is to provide researchers with information that will aid them in evaluating the reliability of using coexpression data to predict transcriptional regulatory interactions and/or coregulation.

In addition, we will also study and classify fan-out motifs [3]: subgraphs composed of a TF gene that coregulates two genes that do not interact directly. In some algorithms using coexpressed profile data to infer TRIs, these coregulated gene pairs are identified as TRIs if they have coexpressed profiles and one of the genes is a TF gene. Different approaches have been applied to identify non-interacting gene pairs in triplets of significantly coexpressed genes, where the main motivation

has been to lower the false positive rate of inferring TRIs [5, 11–14]. In this study, we compare the performances of two prominent approaches. One approach is based on application of the data processing inequality (DPI) [5, 15]. The DPI is a general result that can be rigorously derived and states that if, gene X_2 interacts with both genes X_1 and X_3 and $X_1$and X_3 do not interact, then the mutual information between X_1 and X_3 is smaller than the mutual informations of either of the other two gene pairs. (we emphasize that the satisfaction of the technical condition[1] of non-interaction of X_1 and X_3 is not clear for actual gene interactions and we will discuss this subsequently in section 2.3.) In contrast, another approach claims that the non-interacting gene pairs in fan-out motifs have the maximum mutual information of gene pairs in the triplet [14]. Although [16] points out that application of the DPI in the former approach can fail when mRNA and protein levels of the TF are weakly correlated, this does not necessarily imply the failure of that approach, and the DPI continues to be used by some researchers [5, 15]. One purpose of our study is to address the extent to which the DPI is useful in this context by evaluating its performance using both gene expression and established TRI data. Given these data, we extract fan-out motifs in which at least one of the two non-interacting genes is a TF gene (as is the case when the DPI is commonly applied) and coexpression levels of all gene pairs are above certain thresholds. For each such threshold, we calculate the fraction of the non-interacting gene pairs having the largest, intermediate and smallest mutual information of all pairs in the triplet.

[1]More formally, if x_1, x_2, x_3 are the expression levels of genes X_1, X_2, X_3, then the probability densities for simultaneously observing expression levels x_1 and x_3 given x_2 satisfy $P(x_1, x_3|x_2) =$

A previous study showed that coregulated gene pairs with a high degree of coexpression tend to be positively correlated [10]. We also explore whether a similar tendency exists in expression correlations between the TF gene and each of the coregulated genes in the datasets we study. In this case, we consider fan-out motifs regardless of whether or not the two coregulated genes interact directly and look for patterns in expression correlations among genes in these three gene subgraphs. To do this, we divide these subgraphs into different types according to the signs of Pearson correlations between gene pairs in the subgraph. There are six such possibilities which we call 'correlation motifs'. Also, we investigate the classification of these motifs in relation to our obtained mutual information and z-score metrics.

In the following, we first describe the data and the z-score similarity measure. Next, we compare the performance of using coexpression to infer TRIs to that of using coexpresssion to infer coregulated gene pairs. We then investigate the DPI in fan-out motifs, and we classify these motifs on the basis of the correlations between pairs of genes in the motifs. Conclusions are drawn in section 2.4.

We emphasize that our study focuses on testing the validity of the DPI method for pruning indirect interactions, and we have not attempted to test other pruning methods, although our testing techniques could possibly be applied them. For example, alternative proposed pruning techniques include MRNET [11], conditional mutual information [12], and conditional independence [13]. Also, see Ref. [17] for a comparison of DPI with some of these method.

$P(x_1|x_2)P(x_3|x_2)$. That is, for fixed x_2, the expression levels x_1 and x_3 are uncorrelated, and the probability of measuring an expression level x_1 (or x_3) depends only on x_2 and not on x_3 (or x_1).

2.2 Methods

2.2.1 Microarray Expression Data

We use gene expression microarray data from the Many Microbe Microarray Database (M^{3D}) [18] to analyze both *E.coli* and yeast. The expression data consist of a compendium of 445 *E.coli* and 247 yeast Affymetrix Antisense2 microarray expression profiles for 4345 and 5520 genes, respectively. These microarray data were collected under different experimental conditions: different genetic backgrounds, media, growth conditions and perturbing chemicals.

2.2.2 Known Transcriptional Regulatory Interactions

We use RegulonDB for the established network for *E. coli* and four databases for yeast. We summarize these databases in Table.2.1.

Table 2.1: The number of TFs, regulated genes and edges in our established TRI data set of known TRIs for *E.coli* and yeast.

Species	Data set of known TRIs	No. of TFs	No. of regulated genes	No. of edges
E.coli	RegulonDB	171	1410	3458
yeast	Lee 02A (Chip-chip)	96	2007	3747
yeast	Harbison 04 (Chip-chip/Sequence motif)	99	1732	3186
yeast	Milo 02 (Compilation)	106	451	801
yeast	Lee 02B (Compilation)	114	536	1017

For *E.coli*, we obtain an established network of TRIs from RegulonDB version 6 [19]. 2% of the genes involving in TRIs from RegulonDB cannot be found in

our microarray data. We remove interactions related to those genes from our TRI established network, as well as self-regulatory TRIs. This results in a TRI established network data set consisting of 3458 interactions between 171 TF genes and 1410 genes.

For yeast, a single, generally accepted standard TRI database (analogous to RegulonDB for *E.coli*) has not been established. Therefore, we use four sources of infered TRIs. As with *E.coli*, we filter out self-regulatory interactions and interactions with genes that are not found in our microarry data.

The **first** database (Lee 02A (Chip-chip)) [20] was obtained using the technology of chromatin immunoprecipitations *in vivo* with microarray (Chip-chip) to identify the binding of TFs to promoter regions in yeast. This database contains 3747 links (bindings) between 96 TFs and 2007 target genes. (Note that the physical bindings of a TF to the promoter regions of a gene does not necessarily imply a regulatory relationship between the TF producing gene and target gene.)

The **second** yeast database (Harbison 04 (Chip-chip/Sequence motif)) [21] was constructed via several steps. First, cis-regulatory sequences, which may act as recognition sites for TFs were identified by combining information from genome-wide location data by Chip-chip, phylogenetically conserved sequences and previously published evidence. Motif discovery methods were applied to these regions in order to discover significant TF-related sequence motifs. Two standards have to be met for these significant motifs in order to conclude the binding of a TF to a promoter region: first, the binding pair is required to have been assigned a high confidence score ($p \leq 0.001$) by Chip-chip; second, the promoter sequences are required to be

conserved among *sensu stricto Sccharomyces* species. The data set thus obtained includes 3186 interactions between 99 TF genes and 1732 genes.

The **third** yeast database (Milo 02 (Compilation)) [22] was extracted from the Yeast Proteome Database (YPD) [23]. This data set, a compilation from various sources in the literature, provides a list of TRIs including 800 interactions between 73 TF genes and 550 genes and is available to download at www.weizmann.ac.il/mcb/UriAlon.

The **forth** yeast database (Lee 02B (Compilation)) [20] is also a compilation of previously discovered TF-gene bindings (proved by *in vivo* binding, *in vitro* binding, indirect binding and sequence analysis). This collection of interactions is used to compare with the TF-gene binding data from Chip-chip experiments. The result yields 1017 TRIs between 87 TF genes and 400 target genes and can be downloaded at web.wi.mit.edu/young/regulator/_network.

Among our four TRI yeast databases, we believe that the first two (Chip-chip and Chip-chip/Sequence motif) are of generally better quality. We also note that these first two database (in contrast to the other two) cover almost the whole genome. However, since the four yeast databases may reflect different aspects of the true TRIs, we will give results of analyses using all four.

2.2.3 Quantifying the Similarity of Expression Profiles

For each pair of genes, we characterize the similarity between their mRNA expression profiles by three metrics: Pearson correlation (r), mutual information (MI), and z-score (z). The z-score is used by the CLR algorithm and is related to

the empirical distribution of MI values. We here provide a brief review of these metrics.

The Pearson correlation r. Given m genes (including all TF genes), we compute an estimate of the $m(m-1)/2$ Pearson correlations between gene X_i and X_j, $r(X_i, X_j)$, using

$$r(X_i, X_j) = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})}{(n-1)s_i s_j},$$

where $x_{ik}(x_{jk})$ is the gene expression level of gene $X_i(X_j)$ in the kth experimental condition, and n denotes the number of conditions. $\bar{x_i}(\bar{x_j})$ and $s_i(s_j)$ are the mean and stardard deviation of the gene expression level of gene $X_i(X_j)$.

The mutual information, MI. We compute an estimate of the mutual information between genes X_i and X_j based on the formula,

$$MI(X_i; X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p_1(x_i) p_2(x_j)}, \tag{2.1}$$

where $x_i(x_j)$ is the variable denoting the expression level of gene $X_i(X_j)$. Also, $p(x_i, x_j)$ is the joint probability distribution, and $p_1(x_i)$ and $p_2(x_j)$ are the marginal probability distribution function for each gene. The expression levels from our databases are continuous variables. To compute the mutual information between continuous random variables, we use a B-spline mutual information estimation code from M^{3D} website [18], where this code used a B-spline smoothing and discretization method with 10 bins and third order B-spline to estimate the probabilities in (2.1) [18, 24].

The z-score. The CLR algorithm is an extension of the Relevance network method based on mutual information [5] and uses the z-score between two genes to

infer TRIs. The **z-score**, $Z(X_i; X_j)$, is defined as

$$Z(X_i; X_j) = \sqrt{Z_i^2 + Z_j^2},$$

where

$$Z_i = \frac{MI(X_i; X_j) - \overline{MI_i}}{\sigma_i};$$

$\overline{MI_i}$ and σ_i are the mean and standard deviation of the set of values of $\mathrm{MI}(X_i; X_k)$, $k = 1, ..., m$.

2.2.4 Error Bars on A Fraction

For a sample population of size N, and $\widetilde{N} < N$ of these measured to have some specific property, the standard error of $\widetilde{N}/N$ is estimated to be

$$[\widetilde{N}(N - \widetilde{N})]^{1/2}/N^{3/2}. \tag{2.2}$$

2.3 Results

As detailed in section 2.2, we obtain microarray expression data for *E.coli* and yeast from M^{3d} [18], and established transcriptional regulatory interaction data sets from RegulonDB [19] for *E.coli* and from four data sets [20–22] for yeast. We use these data in two different types of analyses. In the first type of analysis, we use the z-score metric (described in section 2.2) to determine strongly coexpressed gene pairs, and we compare these with gene pairs in our established TRI data sets. In the second type of analysis, we use the established TRI data together with expression

correlation values (using different metrics) to obtain different types of three-gene interaction motifs.

2.3.1 Signatures of Coregulation

There is a question as to whether the degree of coexpression is a predictor of a transcriptional regulatory interaction (TRI), a coregulated gene pair, or both. A high degree of coexpression, as measured by Pearson correlation, has been claimed to indicate coregulated gene pairs [10]. We also note that, a high degree of coexpression between expression profiles of TF-gene pairs, as measured by a high z-score, has been argued to represent TRIs between TF genes and target genes [4]. A benefit of using the z-score to measure the degree of coexpression is that it takes into account the noise in gene expression levels and is therefore considered to be a better measure of coexpression than raw MI. In what follows, we use the z-score to investigate the above question. We find that a high degree of coexpression is more likely to predict coregulated gene pairs for yeast, while it is more likely to predict TRIs for *E.coli*.

When using coexpression to infer TRIs, a TRI is predicted when a gene pair has at least one TF gene and a z-score above a chosen cutoff. When using coexpression to infer coregulation, a gene pair is predicted to be coregulated if its z-score is above a chosen cutoff. To evaluate the quality of these predictions, we use several quantitative measures, namely, the precision (pr), the recall (re), and the F-score. For coregulated gene pairs/TRIs, the precision (pr) is defined as the ratio of the number of correctly predicted coregulated gene pairs/TRIs to the total

number of predicted coregulated gene pairs/TRIs. The recall (re) is defined as the ratio of the number of correctly predicted coregulated gene pairs/TRIs to the total number of coregulated gene pairs/TRIs. Then F-score defined as $2Pr \times Re/(Pr + Re)$, is a measure of the quality of the prediction that reflects the tradeoff between precision and recall. Figure 2.1 shows plots of F-score versus z-score cutoff for *E.coli* (Fig. 2.1A) and for yeast (Figs. 2.1B-E) for three different predictions (the red, green and blue curves). For *E.coli* (Fig. 2.1A), the F-score for the prediction of coregulated gene pairs (blue curve) is larger than that for TRIs (red curve) when the z-score cutoff is smaller than 3. However, when the z-score cutoff is greater than 3, prediction of TRIs performs better. For the four established TRI data sets of yeast (Figs. 2.1B-E), F-score values for the prediction of coregulated gene pairs (blue curves) are significantly larger than those for the prediction of TRIs (red curves) for all z-score cutoff, so indicating that the performance of using z-score to predict coregulated gene pairs is better than that of using z-score to predict TRIs. Also, for both predictions of coregulated gene pairs and TRIs (Figs. 2.1D-E), the plots corresponding to the Milo02 and Lee02B TRI data sets have F-score peaks around z-score cutoffs of 3-4 while the other two plots have their maximum F-score at z-score cutoffs of 1. This is an indication for the differences among the TRIs in the four established TRI data sets.

In addition to exploring the incidence of coregulation in all gene pairs with z-score above a certain value, we separately consider only the set of gene pairs with at least one TF gene and z-score above said value. The corresponding F-score curves are plotted in green in Fig. 2.1 for both *E.coli* and yeast. For *E.coli*, this

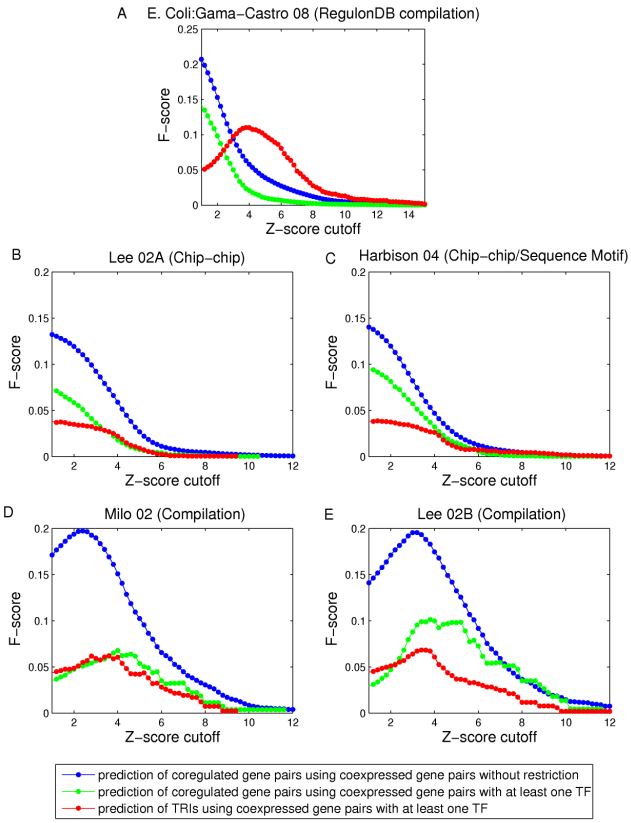

Figure 2.1: **F-score vs. z-score cutoff** F-score versus z-score cutoff for prediction of coregulated gene pairs and TRIs are plotted in blue and red resepectively. Also, the F-score curves for the prediction of coregulated gene pairs in coexpression gene pairs with at least one TF gene is plotted in green. The five subplots correspond to the five established TRI data sets for *E.coli* and *yeast*(Table.2.1), A) RegulonDB, B) Lee et al. 2002(Chip-chip), C) Harbison et al. 2004 (Chip-chip/sequence motif), D) Milo et al. 2002 (Compilation) and E) Lee et al. 2002 (Compilation).

green F-score curve is always below that of prediction of coregulated gene pairs from non-restricted coexpressed gene pairs (blue curve). Also, it is below the red F-score curve for prediction of TRIs when z-score cutoff is greater than 2. For yeast, considering Figs. 2.1B and 2.1C, we see that the F-score curve for prediction of coregulated gene pairs from restricted coexpressed gene pairs is below that of prediction of coregulated gene pairs from non-restricted coexpressed gene pairs, but above the F-score curve for prediction of TRIs. This indicates that, for both *E.coli* and yeast, coregulated gene pairs with at least one TF are likely to have smaller z-score compared to the unrestricted coregulated gene pairs. We have also studied the precision-recall graphs for all the prediction for both *E.coli* and yeast and the same results are obtained (Fig. 2.2). Our studies reveal that when we go from *E.coli* to yeast, the performance of predicting TRIs using z-score degrades. However, the performance of using z-score to predict coregulated gene pairs from coexpressed gene pairs without restriction is reasonable for both *E.coli* and yeast.

Because the microarray sample size for *E. coli* is much larger than that for yeast, we also employed a sampling approach to demonstrate that the difference in sample sizes does not bias the above conclusions. Specifically, we have recomputed Fig. 2.1A using randomly selected sets of *E. coli* samples comparable in size to that for our yeast results (Figs. 2.1B-E). This result, given in Fig. 2.3B, shows that the *E.coli* patterns using the smaller sample size are virtually identical to that in Fig. 2.1A.

Also, TRIs are relatively easier to justify for *E. coli* than for yeast since *E. coli* is a much simpler organism than yeast. This might suggest that the yeast TRI

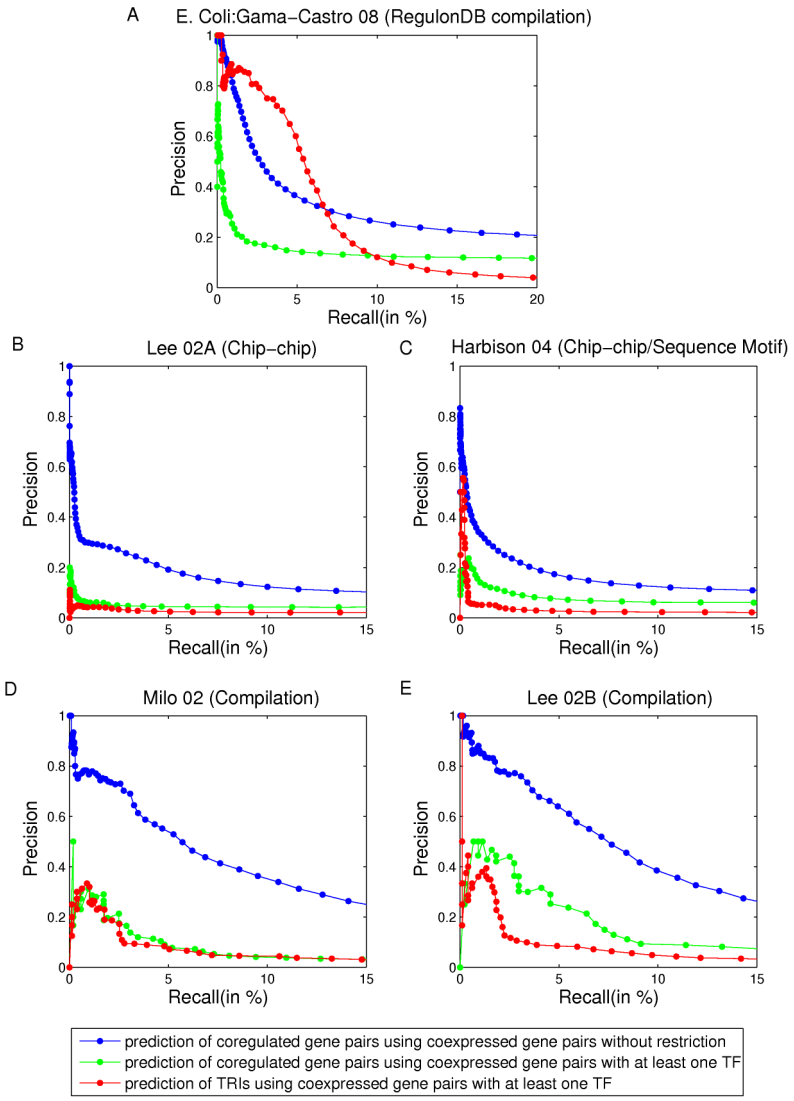

Figure 2.2: **Precision vs. recall**. A-E) Precision versus recall for prediction of coregulated gene pairsand TRIs are plotted in blue and red, resepectively. Also, the precision-recall curve for the prediction of coregulated gene pairs in coexpression gene pairs with at least one TF gene is plotted in green. The five subplots correspond to the five established TRI data sets for *E.coli* and *yeast*(Table. 2.1), A)RegulonDB, B) Lee et al. 2002(Chip-chip), C) Harbison et al. 2004 (Chip-chip/sequence motif), D) Milo et al. 2002(Compilation) and E) Lee et al. 2002 (Compilation).

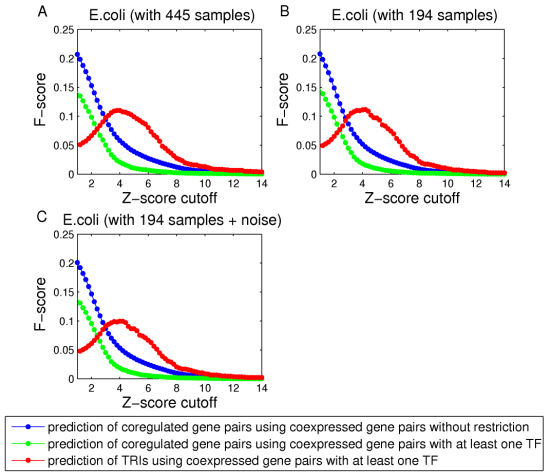

Figure 2.3: F-score vs. z-score cutoff for *E. coli*. F-score versus z-score cutoff for prediction of coregulated gene pairs and TRIs are plotted in blue and red, resepectively. Also, the F-score curves for the prediction of coregulated gene pairs in coexpression gene pairs with at least one TF gene is plotted in green. A B-spline estimator is used to calculate the mutual information. The three subplots, A, B and C, correspond to different number of samples, A) uses 445 samples (this figure is the same as Fig. 2.1A), B) uses 194 samples, and C) uses 194 samples and adds noise. The number 194 is derived from 247 (samples for yeast in the data used to derive Figs. 2.1B-E) 4345 (E.coli genes) 5520 (yeast genes)=194. For B), the smaller number of samples was obtained by random selecting from the 445 E. coli RegulonDB samples used in A). For C), the number of sample is the same as B), and 10% of the links in RegulonDB are deleted and each deleted link is replaced by a link from a randomly selected TF gene to a randomly selected gene. The fact that these figures are virtually identical confirms that any difference between our result in A) with the corresponding yeast results (Figs. 2.1B-E) is not due to the larger sample size of the *E. coli* microarray database or to lower noise in the RegulonDB database relative to our *yeast* databases.

databases are more noisy than the RegulonDB database. In order to demonstrate that noise in yeast TRI databases does not bias our conclusions, we recompute the *E. coli* result (Fig. 2.3B) by randomly deleting 10% of the links in RegulonDB and then replacing each deleted link by a link from a randomly selected TF gene to a randomly selected gene. This result, given in Fig. 2.3C, shows that the *E. coli* patterns in Fig. 2.1A are robust to adding noise to the TRI database.

The above tests (decrease of the *E. coli* sample size and addition of noise to RegulonDB) confirm the robustness of our conclusion (based on Fig. 2.1) that when we go from *E.coli* to yeast, the performance of predicting TRIs using z-score degrades while the performance of predicting coregulated gene pairs from coexpressed gene pairs without restriction is reasonable for both *E.coli* and yeast.

2.3.2 MI-motifs

Given an established TRI data set, we can identify all fan-out motifs, where a fan-out motif is defined as a subgraph formed by two non-interacting genes and a TF gene that coregulates them. Here we only consider fan-out motifs in which one of the two coregulated genes is itself a TF gene. The three gene pairs in each fan-out motif are assigned values according to their respective mutual information values. Then we define the three types of MI-motifs shown in Fig. 2.4A, MI_1, MI_2 and MI_3, which refer to the case that the value of MI of the non-interacting gene pair is the largest, intermediate and smallest as compared to that of the two TF-gene pairs respectively. If more fan-out motifs are identified as MI_3-motifs, the

data processing inequality(DPI) is a good tool for inferring the non-interacting gene pairs in fan-out motifs. Conversely, if MI_1-motifs dominate, the non-interacting gene pairs predominantly have the largest MI values within their fan-out motifs, and one might predict that the largest MI indicates coregulation in such a situation, we call this the 'max MI approach'. [14].

In order to address the utility of the DPI in this context, we compare the relative abundances of the three MI-motifs in the set of fan-out motifs described above, and we assess how the coexpression levels of gene pairs in fan-out motifs is related to these relative abundances. To do this, we generate different groups of fan-out motifs as we vary the z-score cutoff. For each z-score cutoff, we include only those fan-out motifs in which all gene pairs have a z-score above the cutoff. For each group of fan-out motifs, we compare the relative abundance of the three MI-motifs. We plot the fractions of the three MI-motifs found as a function of the z-score cutoff on all gene pairs. Figs. 2.4B-F show results for both *E.coli* and yeast. For *E.coli* (Fig. 2.4B), the relative abundance of MI_3-motif is always higher than 40% while that of MI_1-motif is always lower than 25%. When the z-score cutoff is larger than 2, the relative abundances of MI_1, MI_2 and MI_3-motifs have no distinguishable differences. For the analyses of the Lee02A, Harbison04 and Lee02B data sets of yeast (Figs. 2.4C,D and F), the relative abundances of MI_3-motif are always lower than 30% while those of MI_1-motif are always higher than 40%. Especially, for the analyses of the Lee02A and Harbison 04 data sets, the relative abundances of MI_1-motif are always around 50%. However, for the analysis of the Milo02 data set (Fig. 2.4C), the relative abundances of the three MI-motifs

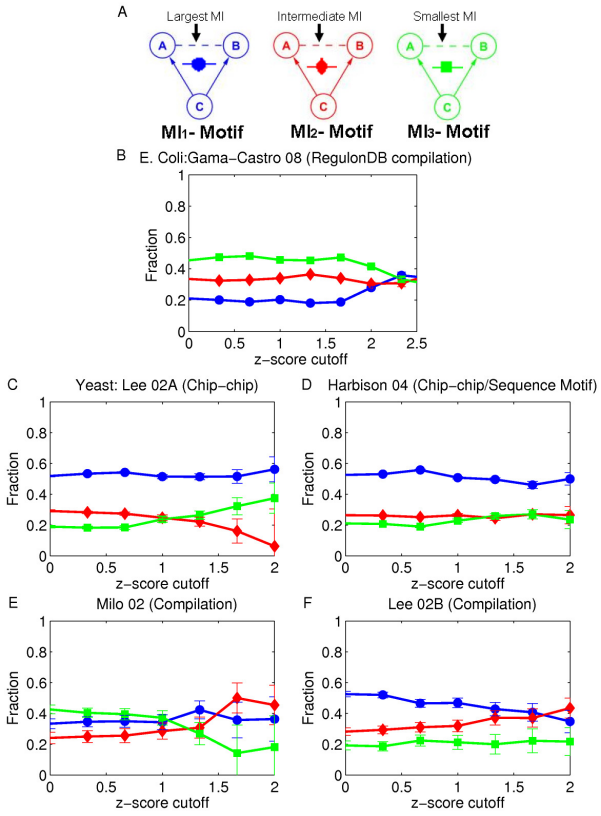

Figure 2.4: **Fractions of MI-motifs vs. the z-score cutoff of non-interacting gene pairs.** Non-interacting gene pairs in fan-out motifs are restricted to gene pairs with at least one TF gene. A)MI-motifs in which the non-interacting gene pair has the largest, intermediate and smallest MI. Fractions of MI_1, MI_2 and MI_3-motifs are plotted in blue, red and green respectively for B)*E.coli* and C-F)*yeast*. The five subplots correspond to the five established TRI data sets for E.coli and *yeast*(Table.1), B) RegulonDB, C) Lee et al. 2002(Chip-chip), D) Harbison et al. 2004 (Chip-chip/sequence motif), E) Milo et al. 2002 (Compilation) and F) Lee et al. 2002 (Compilation).

are similar and cannot be distinguished. For all four yeast databases, there is no obvious increasing/decreasing trend for these relative abundances with increasing z-score cutoff. This implies that the DPI in the case of *E.coli* works better than the max MI approach and the random prediction for inferring non-interacting gene pairs in fan-out motifs (relative abundance of each MI-motif is equal to one-third in random prediction). However, the performances of the DPI and the max MI approaches are the opposite for yeast. The max MI approach works better than the random case while the DPI fails in inferring non-interacting gene pairs in fan-out motifs (i.e., the DPI prediction is more often false than a random unweighted guess of the non-interacting links).

Similar to Fig. 2.3, we show in Fig. 2.5 that the main important features of Fig. 2.4B are robust to decrease of the *E.coli* sample size to be comparable to the yeast sample size, and also robust to add noise to the *E. coli* TRI database.

In order to demonstrate that our results are not snesitive to the method used for mutual information estimation (a B-spline estimator), we have recomputed Fig. 2.4B for *E.coli* and Figs. 2.4C-F for yeast using both empirical [11] and Miller-Madow [25] estimators with both equal-width and equal-frequency binning (10 bins for both). We choose these two estimators because it has been shown that the ARACNE inference method (a method based on DPI) gives the better performance when using these two estimators with equal-frequency binning [17]. The results are given (Figs. 2.6-2.10), and show that both the *E.coli* and yeast results recomputed using the empirical and Miller-Madow mutual information estimators with both equal-width and equal-frequency are similar to those in Fig. 2.4B and Figs. 2.4C-F.

In particular as before, for *E. coli* the DPI approach for pruning the non-interacting links in fan-out motifs works better than random and the max MI approach, but it works worse than random in yeast in general.

Regarding the strikingly poor performance in yeast, we note that the DPI, while a rigorous result, only applies when the hypothesis under which it was derived applies (see section 2.1), and it is unclear to what expect this is the case for gene expression data. One mechanism violating the necessary hypothesis is the possible imperfect correlation between a TF's mRNA level and the production rate of its protein (see Ref. [16]). Another mechanism that would have an equivalent effect is that it can take considerable time for mRNA to be translated into its protein, and thus there can be a significant time lag between the expression levels of a TF and that of its target genes. Still another mechanism that might be relevant is that the expression of target genes may be dependent, not only on the presence of the TF protein involved in the fan-out motif considered, but may also be strongly influenced by other fluctuating factors. Our results suggest that at least one mechanism like those above is most often operative in yeast, but not in *E.coli*. Therefore, the applicability of the data processing inequality may be organism-dependent.

2.3.3 Correlation-motifs

A previous study showed that coregulated gene pairs with a large magnitude of Pearson correlation coefficient between their expression profiles tend to be positively correlated [10, 26]. In our study, instead of using Pearson correlation, we will

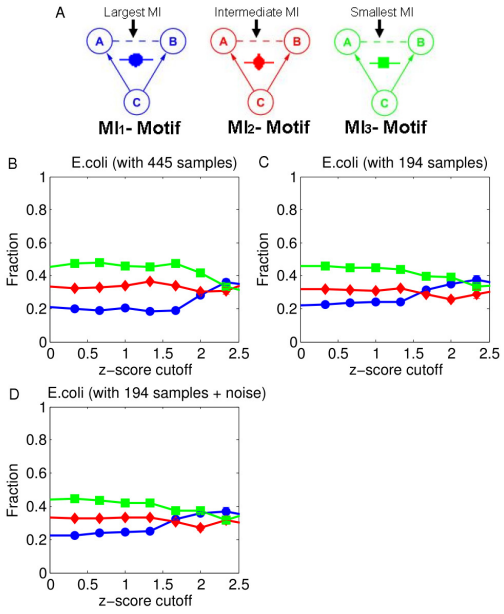

Figure 2.5: **Fractions of MI-motifs vs. the z-score cutoff of non-interacting gene pairs for *E. Coli*.** Non-interacting gene pairs in fan-out motifs are restricted to gene pairs with at least one TF gene. A) MI-motifs in which the non-interacting gene pair has the largest (MI1 schematic), intermediate (MI2 schematic) and smallest (MI3 schematic) MI. Fractions of MI_1, MI_2 and MI_3 motifs are plotted in blue, red, and green, respectively. A B-spline estimator is used to calculate the mutual information. As in Fig. 2.3, the three subplots, B, C and D, correspond to B) 445 samples (this is the same as Fig. 2.4B), C) 194 samples, and D) 194 samples plus noise.

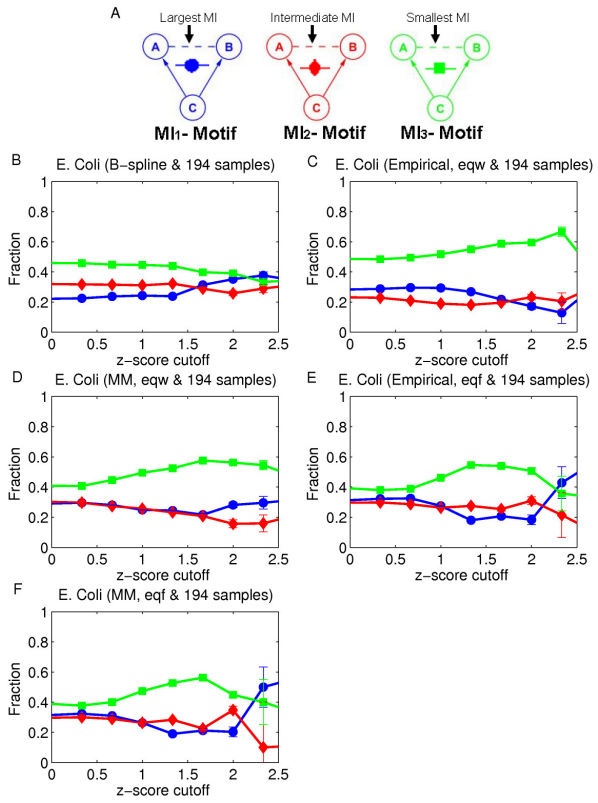

Figure 2.6: **Fractions of MI-motifs vs. the z-score cutoff of non-interacting gene pairs for *E. Coli* with using different MI estimators as in Fig. 2.4B.** A) MI-motifs in which the non-interacting gene pair has the largest, intermediate and smallest MI. Fractions of MI_1, MI_2 and MI_3 - motifs are plotted in blue, red and green respectively. The five subplots correspond to the use of different MI estimators and discretization method, B) B-spline (the same figure as in Fig.2.5C), C) Empirical [9] and equal width (eqw), D) Miller- Madow (MM) [24] and equal width (eqw), E) Empirical and equal frequency (eqf) and F) Miller- Madow (MM) and equal frequency (eqf). These plots show that the conclusion that the green plot is generally above the red and blue plots is independent of the MI estimator that is employed.

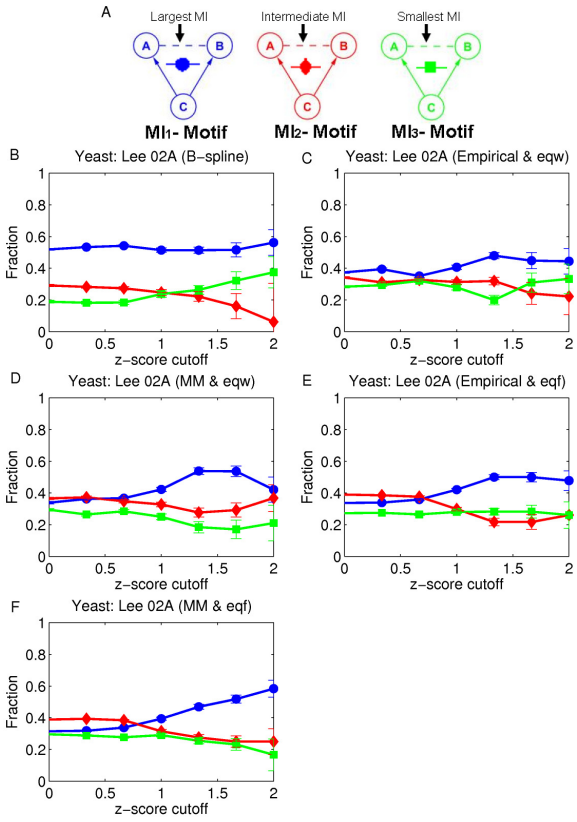

Figure 2.7: **Fractions of MI-motifs vs. the z-score cutoff of non-interacting gene pairs for Lee02A (Chip-chip) of *yeast* as in Fig. 2.4C.** A) MI-motifs in which the non-interacting gene pair has the largest, intermediate and smallest MI. Fractions of MI_1, MI_2 and MI_3 - motifs are plotted in blue, red and green respectively. The five subplots correspond to the use of different MI estimators and discretization method, B) B-spline (the same figure as in Fig. 2.4C), C) Empirical[9] and equal width (eqw), D) Miller- Madow (MM)[24] and equal width (eqw), E) Empirical and equal frequency (eqf) and F) Miller- Madow (MM) and equal frequency (eqf). These plots show that (in contrast to Fig. 2.6) the green plot is consistently below the blue plot independent of the MI estimator that is employed.

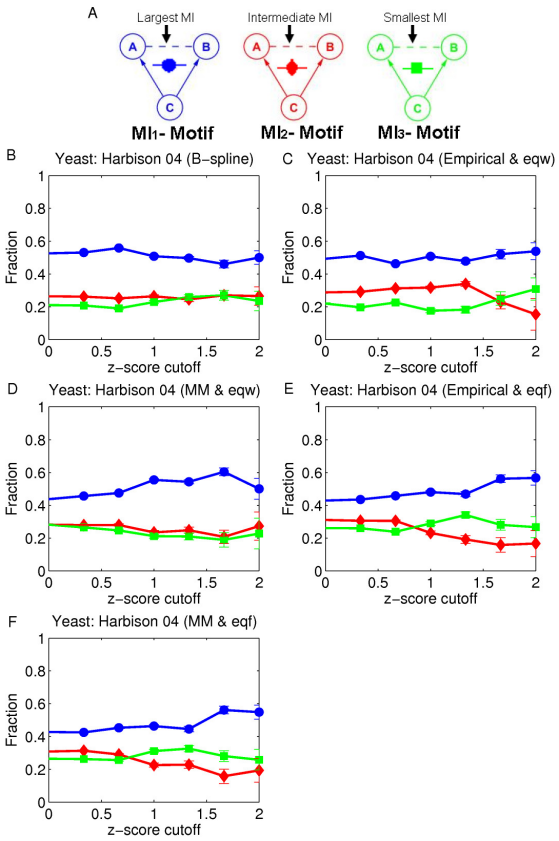

Figure 2.8: **Fractions of MI-motifs vs. the z-score cutoff of non-interacting gene pairs for Harbison 04 (Chip-chip/Sequence Motif) of yeast as in Fig. 2.4D.** A) MI-motifs in which the non-interacting gene pair has the largest, intermediate and smallest MI. Fractions of MI_1, MI_2 and MI_3 - motifs are plotted in blue, red and green respectively. The five subplots correspond to the use of different MI estimators and discretization method, B) B-spline (the same figure as in Fig. 2.4D), C) Empirical[9] and equal width (eqw), D) Miller- Madow (MM)[24] and equal width (eqw), E) Empirical and equal frequency (eqf) and F) Miller- Madow (MM) and equal frequency (eqf). These plots show that (in contrast to Fig. 2.6) the green plot is consistently below the blue plot independent of the MI estimator that is employed.

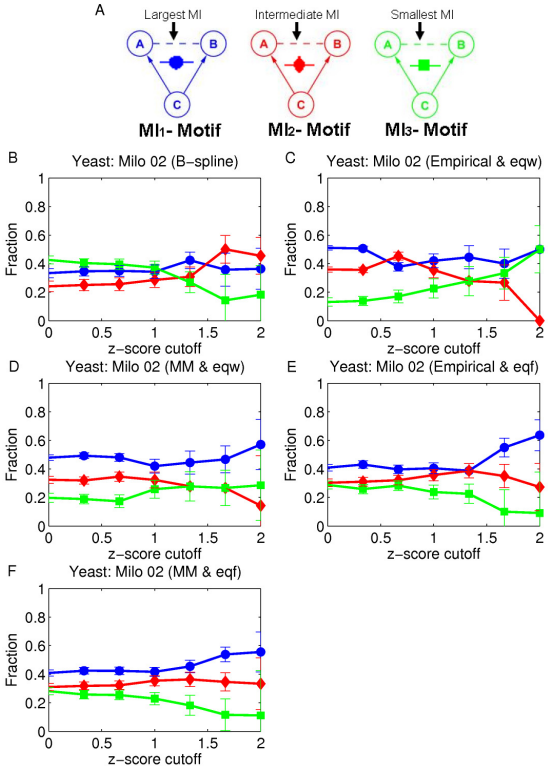

Figure 2.9: **Fractions of MI-motifs vs. the z-score cutoff of non-interacting gene pairs for Milo 02 (Compilation) of *yeast* as in Fig. 2.4E.** A) MI-motifs in which the non-interacting gene pair has the largest, intermediate and smallest MI. Fractions of MI_1, MI_2 and MI_3 - motifs are plotted in blue, red and green respectively. The five subplots correspond to the use of different MI estimators and discretization method, B) B-spline (the same figure as in Fig. 2.4E), C) Empirical[9] and equal width (eqw), D) Miller- Madow (MM)[24] and equal width (eqw), E) Empirical and equal frequency (eqf) and F) Miller- Madow (MM) and equal frequency (eqf). These plots show that (in contrast to Fig. 2.6) the green plot is consistently below the blue plot independent of the MI estimator that is employed.

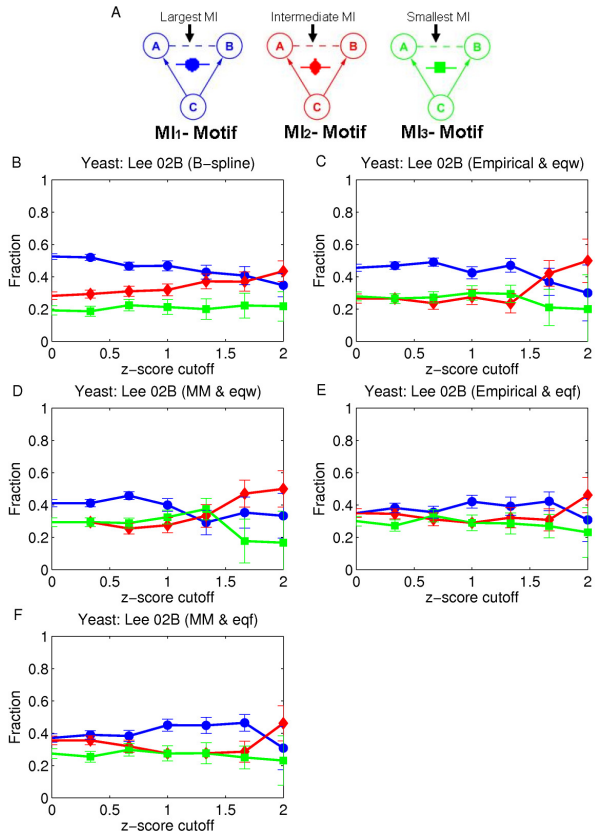

Figure 2.10: **Fractions of MI-motifs vs. the z-score cutoff of non-interacting gene pairs for Lee 02B (Compilation) of *yeast* as in Fig. 2.4F.** A) MI-motifs in which the non-interacting gene pair has the largest, intermediate and smallest MI. Fractions of MI_1, MI_2 and MI_3 - motifs are plotted in blue, red and green respectively. The five subplots correspond to the use of different MI estimators and discretization method, B) B-spline (the same figure as in Fig. 2.4F), C) Empirical[9] and equal width (eqw), D) Miller- Madow (MM)[24] and equal width (eqw), E) Empirical and equal frequency (eqf) and F) Miller- Madow (MM) and equal frequency (eqf). These plots show that (in contrast to Fig. 2.6) the green plot is consistently below the blue plot independent of the MI estimator that is employed.

use the z-score metric to measure the degree of coexpression. An initial question is whether the previously found pattern in expression correlation of coregulated gene pairs [10, 26] also appears when the z-score metric is used to quantify coexpression. Figure 2.11 shows a plot of Pearson correlation versus z-score for *E.coli*. In this figure, gene pairs that are coregulated and not coregulated according to RegulonDB compilation are plotted as blue and red dots respectively (plots for yeast turn out to show similar features to the plot for *E.coli* and are not shown here). To meaningfully represent relative densities of coregulated (blue) and not coregulated (red) pairs in the presence of overlapping of the printed points, we plot points one by one, alternating between blue and red and selecting the gene pairs in the chosen group (blue and red) randomly. This plot shows that a high z-score (z-score >6) is associated with positive correlation and that high z-score gene pairs are likely to be coregulated [the density of blue dots (coregulated gene pairs) is higher than that of red dots (gene pairs that are not coregulated) when the z-score is high]. Motivated by this finding, we consider the situation when a TF gene regulates two other genes, and we ask whether other patterns exist in expression correlation between the TF gene and each of the coregulated genes when coregulated gene pairs have a high degree of coexpression.

We refer to the TF gene and the two genes that it regulates as a coregulation subgraph and we identify these subgraphs from the established TRI databases. However, in contrast to fan-out motifs (discussed in section 2.3.2), coregulated genes in these coregulation subgraphs may or may not interact directly. To further explore the correlation and coexpression among genes in coregulation subgraphs, we

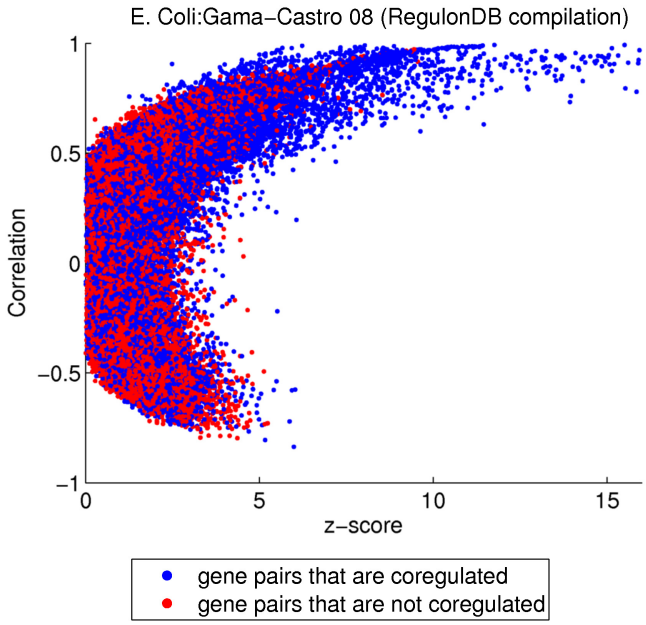

Figure 2.11: **Pearson correlation vs. z-score.** Gene pairs that are coregulated are represented by blue dots and those that are not coregulated are represented by red dots for *E.coli*.

define six correlation-motifs (C-motifs) by classifying the coregulation subgraphs into different types according to the combinations of the signs of Pearson correlation between the expression of coregulation subgraph genes. There are six such types as shown in Figs. 2.12A and 2.12G, where C denotes the TF gene and the other two genes are denoted A and B. The + and − signs on the links denote postive and negative Pearson correlation. We apply Fisher's **z**-transformation to the coefficients of Pearson correlation and obtain the 95% confidence intervals for all coefficients [27]. Among all coregulation subgraphs, we only consider cases where all Pearson correlation coefficients have confidence intervals indicating they have less than a 5% probability to be of the opposite sign.

Next we investigate how the relative abundances of the six C-motifs depends on the z-score between the A and B genes. We first generate different groups of coregulation subgraphs using different z-score cutoffs on the coregulated gene pairs, and for each group, we calculate the relative abundances of the six C-motifs amongst all coregulation subgraphs. Figures 2.12B-F show plots of the fractions of different C-motifs as a function of the z-score cutoff on coregulated gene pairs for both *E.coli* and yeast. Only the fractions of C_1, C_2 and C_3-motifs are shown (respectively plotted in red, blue and green) as those of the other C-motifs (Fig. 2.12G) are very small at all z-score cutoffs. For *E.coli* (Fig. 2.12B), when the z-score cutoff is above 2, the fractions of C_1 and C_2-motifs are always about 75% and 18% respectively, and the fraction of C_3-motifs are always lower than those of C_1 and C_2-motifs and decreases to near zero around a z-score cutoff of 5. For yeast (Figs. 2.12C-F), the C_1 and C_2-motifs are again the most abundant, while C_3-motifs are the least

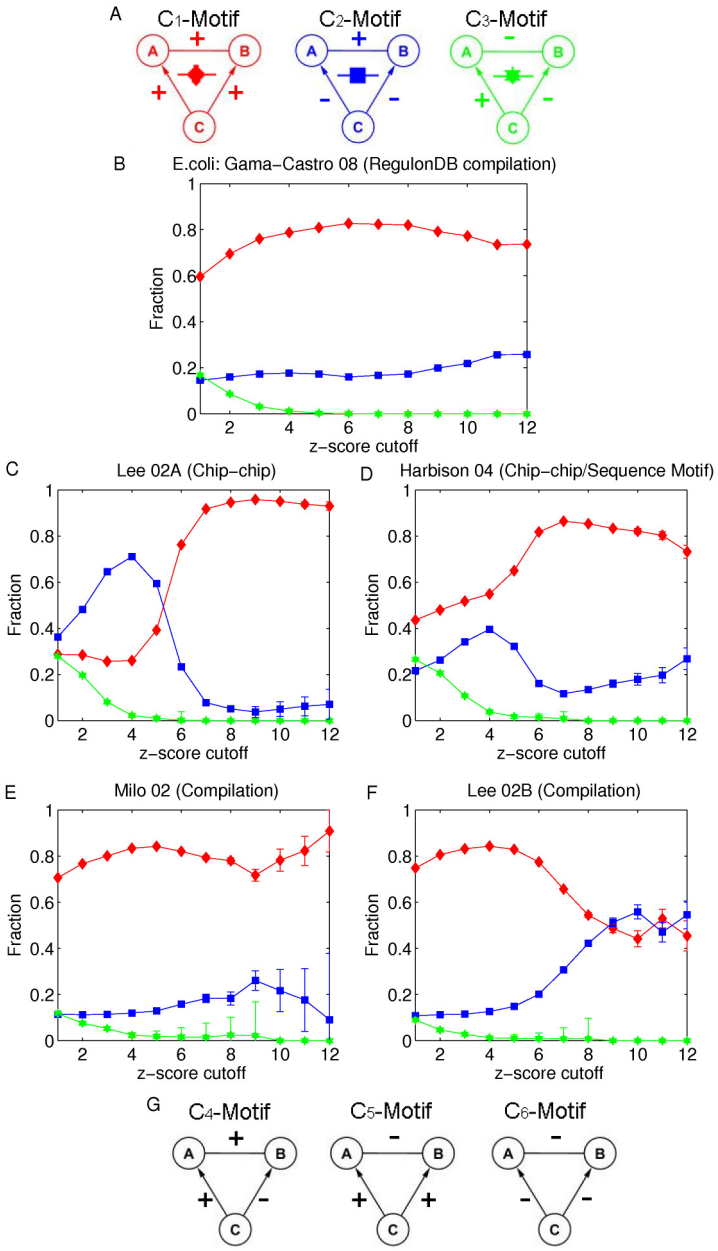

Figure 2.12: **Fractions of C-motifs in a group of subgraphs of coregulation vs. z-score cutoff on coregulated gene pairs in the group.** A) C_1, C_2 and C_3-motifs. B-F) The fractions of C_1, C_2 and C_3-motifs are plotted in red, blue and green (solid lines) respectively. The five subplots correspond to the five established TRI data sets for *E.coli* and *yeast*(Table.2.1), B) RegulonDB, C) Lee et al. 2002(Chip-chip), D) Harbison et al. 2004 (Chip-chip/sequence motif), E) Milo et al. 2002 (Compilation) and F) Lee et al. 2002 (Compilation). G) C_4, C_5 and C_6-motifs.

abundant and their fractions decrease to near zero when the z-score cutoffs are high enough (around 6). In particular, for the analysis using the Lee02A TRI data set (Fig. 2.12C), C_1-motifs are more abundant than C_2-motifs when the z-score cutoff is higher than about 5.5, but they are less abundant than C_2-motifs when the z-score cutoff is lower than 5.5. For the analyses using the other three TRI yeast data sets (Figs. 2.12D, 2.12E and 2.12F), C_1-motifs are generally more abundant than C_2-motifs (except for Fig. 2.12F for the cutoffs greater than 8, where they are approximately equal). The observed differences between the analyses of the four different yeast TRI data sets indicates that there may be significant differences in coregulated genes in different data sets. Overall, results from both *E.coli* and yeast are consistent with our Fig. 2.11 in that coregulated gene pairs with a high degree of coexpression are more likely to be positively correlated. In addition, these results also imply that when coregulated gene pairs have a large enough z-score, the correlations between the TF gene and the two other genes in the coregulation subgraphs both have the same correlation sign (i.e., they are C_1 or C_2 motifs).

We now further characterize the difference between the coregulated gene pairs in C_1 and C_2-motifs used in the plots of Figs. 2.12B-F. For each coregulated gene pair, we find their respective mutual information and z-score. Then we construct scatter plots of mutual information versus z-score for all these coregulated gene pairs for both *E.coli* and yeast (Fig. 2.13) where points corresponding to C_1-motifs are plotted in red and those corresponding to C_2 motifs are plotted in blue. There are more C_2-motifs (blue) than C_1-motifs (red). Since overlapping is present, the order in which we plot the points is significant (as for our previous figure, Fig. 2.11).

In the present case we proceed as follows. We first plot randomly selected blue (C_2-motifs) points until the number of remaining unplotted C_2-motifs is equal to the number of the C_1-motifs. After that, points are plotted one by one, alternating between randomly selected C_1-motifs and randomly selected C_2-motifs. For *E.coli*, data points for coregulated gene pairs in C_1-motif are well mixed with those for coregulated gene pairs in C_2-motif in Fig. 2.13B. Thus there is no apparent distinction observed between coregulated gene pairs in C_1 and C_2-motifs for *E.coli*. Our analyses of the Lee02A and Harbison04 yeast data sets (Figs. 2.13C and 2.13D) show that mutual information is approximately linearly related to z-score for both groups of coregulated gene pairs (corresponding to blue and red), and that, the slope of the linear relationship for C_2-motifs (blue)is larger than that for C_1-motifs (red). However, distinct slopes are not observed in the analyses of the other two yeast established TRI data sets (Figs. 2.13E and 2.13F). We do not presently have a good idea as to a mechanism leading to the observed distinctive C_1 and C_2 patterns seen in Figs. 2.13C and 2.13D.

Regarding a possible reason for the presence of the splitting observed in Figs. 2.13C and 2.13D versus the lack of such a splitting in Figs. 2.13E and 2.13F, we note that the links in the Milo 02/ Lee 02B databases (used for Figs. 2.13E and 2.13F) are very different from those in the Lee02A/ Harbison 04 databases (used for Figs. 2.13C and 2.5D). In particular, the Lee02A and Harbison 04 TRI databases are based on Chip-chip experiments, while links in Milo02 and Lee02B are inferred by several different methods such as changing in the expression of the target gene owing to the deletion (or mutation) of the TF gene. It has been shown

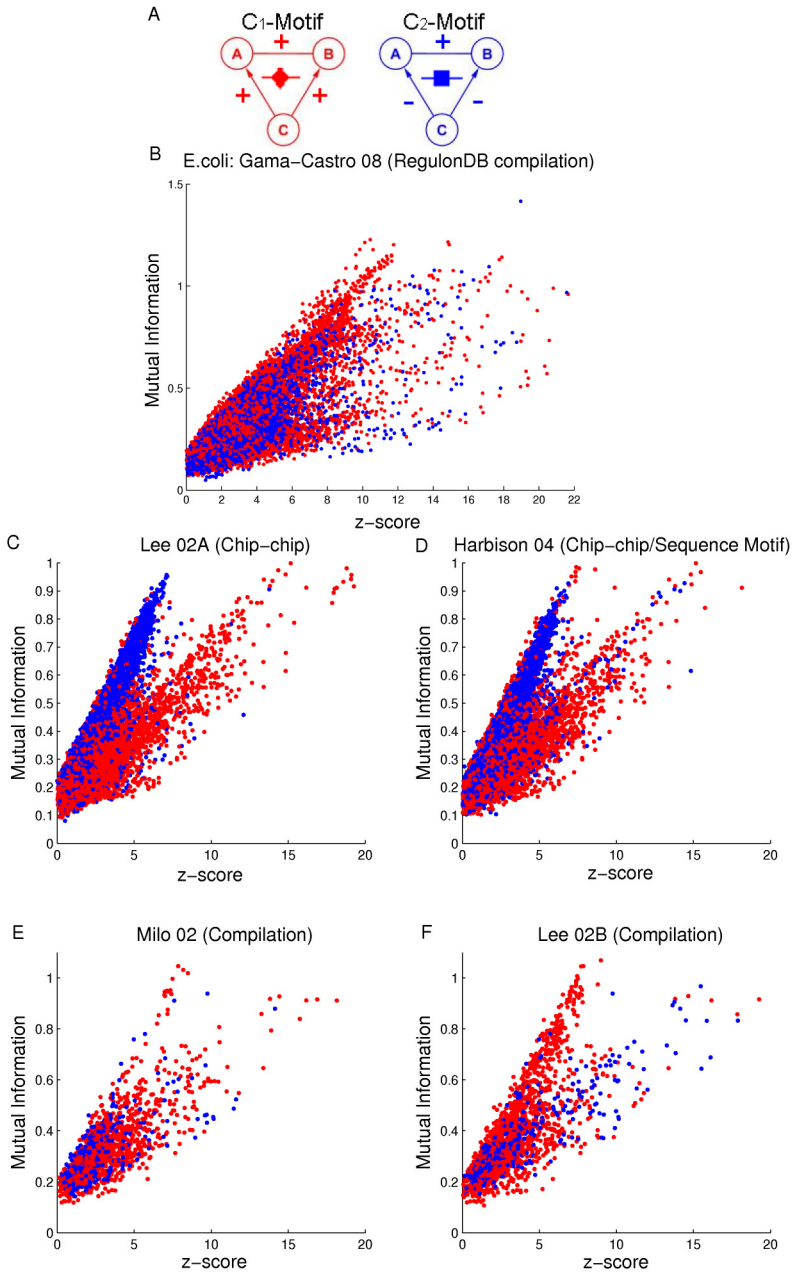

Figure 2.13: **Mutual information vs. z-score for coregulated gene pairs in C_1 and C_2-motifs.** A) C_1 and C_2-motifs. B-F) Data points for coregulated gene pairs in C_1 and C_2-motifs are plotted in red and blue respectively. The five subplots correspond to the five established TRI data sets for *E.coli* and *yeast*(Table.1), B) RegulonDB, C) Lee et al. 2002(Chip-chip), D) Harbison et al. 2004 (Chip-chip/sequence motif), E) Milo et al. 2002 (Compilation) and F) Lee et al. 2002 (Compilation).

that different TRI inference methods, such as Chip-chip, targeted gene disruption, and overexpression of TFs, capture distinct facets of the transcriptional regulatory program, and uncover disparate biological phenomena [28]. The fact that a splitting feature is observed in Figs. 2.13C and 2.13D but not in Figs. 2.13E and 2.13F may be because different biological processes are reflected in their database constructions.

2.4 Discussion

Our study demonstrates that the performances of predictions of coregulated gene pairs and transcriptional regulatory interactions determined by coexpression levels are organism dependent. For *Escherichia coli*, the prediction of transcriptional regulatory interactions outperforms prediction of coregulated gene pairs when the predictions are determined by coexpression with z-score greater than 3. However, the situation is very different for *Saccharomyces cerevisiae*, with the prediction of coregulated gene pairs outperforming the prediction of TRIs for all z-score cutoffs. Many methods of inferring transcriptional regulatory interactions or coregulated gene pairs have been developed and shown to give excellent performance in specific organisms. However, based on our study, applications of these method to other organisms should be conductd with caution as their predicting powers may depend on the organism studied.

The Data processing inequality(DPI) has been applied to the prediction of transcriptional regulatory interactions after excluding highly coexpressed gene pairs that do not interact directly. The results show that the application of the DPI to

Escherichia coli data works better than random prediction of gene pairs. However, the performance of the application of DPI in *Saccharomyces cerevisiae* is worse than that of random prediction. The strong failure of applying DPI to yeast data suggests that factors/mechanisms exist in yeast that lead to an imperfect correlation between the protein and mRNA levels of TFs.

In our study investigating patterns of expression correlation among genes in coregulation subgraphs, we find two distinct types of coregulated gene pairs: one in which the correlation between the expression of the TF gene and both its two target correlated genes are positive and another in which they are both negative. In particular, we present scatter plots of mutual information versus z-score for these two types of gene pairs. The plots for yeast reveal that the two types of coregulated gene pairs split into two parts, thus characterising the differences between these two types of gene pairs. Further studies are needed to explain the mechanism leading to this behavior.

Motivated by the increasing availability high-throughput gene expression data, a variety of approaches have been developed to infer TRIs or gene coregulation. Our studies in this chapter reveal that some approaches which apparently lead to useful prediction in some model organisms may fail in other organisms.

Chapter 3: Modeling the Dynamics of Bivalent Histone Modifications

This work in this chapter was published in PLoSOne in 2013 [2].

3.1 Introduction

Histones can undergo various types of covalent modifications, such as methylation and acetylation, which serve as an additional layer of transcriptional control by mediating the chromatin accessibility and by recruiting regulatory proteins [29, 30]. Experimental studies using chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) have suggested that different cell types can be characterized by different histone modification patterns [31].

The molecular mechanisms underlying chromatin state establishment, maintenance, and heritability remain incompletely understood. A number of mechanisms are implicated [32], including (1) sequence-specific recruitment through interactions between chromatin regulators and DNA binding factors; (2) recruitment of chromatin regulators to existing histone marks; (3) histone marks deposited by transcriptional machineries; (4) RNA mediated recruitment; and (5) stochasticity associated with DNA replication. However, any single mechanism alone is insufficient

for chromatin state establishment [32, 33].

One of the best characterized chromatin states is a bivalent domain, a segment of the nucleosome array, in which H3K4me3 (an active mark) and H3K27me3 (a repressive mark) coexist on most individual nucleosomes within the domain [34]. Bivalent domains are thought to be an important feature of stem cells. For example, bivalent domains have been discovered in the promoters of most lineage-control genes in embryonic stem cells, and most of these domains become monovalent upon cell differentiation [31, 34–37]. Also, a recent study observed that gene activation in the differentiation process occurs in conjunction with the decay of repressive marks in bivalent domains [38]. In particular, one prominent proposal [34] for the function of bivalent domains is that the H3K27me3 marks act to repress the lineage-control gene in stem cells, while the H3K4me3 marks poise these genes for activation upon cell differentiation. Thus this proposal suggests that activation of these genes in differentiated cells is determined by the existence of bivalent domains in stem cells. These findings indicate the importance of bivalent domains and motivate further study in order to illuminate the underlying principles and mechanisms involved in their formation and evolution.

It has been proposed that the formation of chromatin domains is consistent with a model that includes not only the chemical interactions between histone marks, but also nucleation sites where domains are more likely to form [39]. The dynamics of histone modifications have been studied both theoretically and experimentally for some time [39–43]. In general, histone methylation marks are catalyzed by a variety of methyltransferase enzymes which may act singly or cooperatively. For example,

H3K27me3 marks are catalyzed by Ezh2, a core member of the Polycomb group proteins. In addition to the normal stochastic conversion which would be expected from each of these individual enzymes, there is also a feedback process between the histone marks and the enzymes [44]. Existing H3K27me3 marks may attract Polycomb group complexes, which enhance nearby methylation [45, 46]. A similar recruitment mechanism has also been suggested for H3K4me3 via Trithorax protein complexes (TrxG) [47]. In addition, there exists experimental evidence supporting a negative feedback mechanism between H3K4me3 and H3K27me3 marks via the action of histone demethylases [48–52].

Certain specific DNA sequences may serve as the docking sites of modification enzymes and may therefore be associated with enhanced local attraction of histone marks [32, 53]. We refer to these as nucleation sites. For example, CpG islands are strongly enriched in bivalent domains in human and mouse embryonic stem cells [47], and appear to be required for Polycomb binding in certain cases [54].

Recently, *in silico* methods have provided important additional insights for chromatin state inheritance. Major contributions have been made by Dodd el al. [40] and Sedighi and Sengupta [55]. These paper considered 1-dimensional lattice models in which nucleosomes are allowed to have active or repressive modifications that evolve stochastically and by recruitment. They found that a bistable state with either mostly active nucleosomes or mostly repressive nucleosomes can appear and be heritable, consistent with experimental observations. Subsequently, Hodges and Crabtree [39] found that adding a nucleation site into a model of the above type produces a bounded chromatin domain. Also, in a more recent paper, Binder et

al. [56] proposed a model describing binding of catalytic enzymes to DNA and their interaction with histone marks with one aim being explaining length distributions of modified chromatin regions. These past studies are limited to a single type of histone mark on a nucleosome, whereas it is well-known that gene regulation is governed by combinatorial patterns of multiple histone marks [30, 57]. In this study, we extend previous studies by presenting an approach to model the dynamics of combinatorial chromatin states. This is achieved by allowing each individual nucleosome to carry both active and repressive marks simultaneously.

In the next section we describe our model. Then, in the Results section, we apply this model to investigate the dynamics of histone modification patterns with the focus on bivalent domains. Discussion and Conclusions are given at the end of the chapter.

3.2 Methods

General framework of our model. We consider a 1D lattice of N nucleosomes, where there is a nucleosome at each lattice site $i = 1, 2, ..., N$. An actual nucleosome consists of 8 histone protein moleclues, that can be regarded as two identical groups of four each. In what follows we only consider the state of one of these four histone group members, namely the H3 histone, which is specifically related to bivalency. Thus, in our model, we represent the state of a nucleosome as being determined by the states of its two H3 histone copies. There are two modification sites in each H3 histone, one which may have an active mark (such as H3K4me3) and the other

which may have a repressive mark (such as H3K27me3). Thus, there are 16 possible states of a nucleosome, and each of which is determined by 4 histone modification sites (see Figure 16 in Appendix A). As shown in the appendix A, this, together with the restriction obtained from experiment [58] that active and repressive marks do not occur simultaneously on the same H3 histone, leads to the six physically distinct nucleosome states depicted in Fig. 3.1. In Fig. 3.1 the circle represents a nucleosome and the vertical ellipses represent H3 histones. The lower case letters within each ellipse represent the states of the two modification sites of the H3 histone (u = unmodified, α = modified by an active mark, ρ = modified by a repressive mark). For convenience, we assign the symbols UU, AA, RR, AU, UR, and AR to the six possible states. From now on, when we say 'histone' it is to be understood that we mean an H3 histone. We note that the state AR will play a prominent role in subsequent considerations in Section 3.4, and we will call a nucleosome in this state a 'bivalent nucleosome'.

We then allow each nucleosome state to evolve according to a discrete time (t) model, in which from time t to time $t+1$, a nucleosome state changes from state σ to state $\sigma^{'}$ with probability $\pi_{\sigma\sigma'}$. Since the time step $t \to t+1$ is regarded as small, we assume that, at most, only one modification site may change on each time step. Thus, there are 12 possible transitions among the 6 distinct states (see Fig. 3.2 which shows the possible transitions).

Reduced model The general framework above can lead to a relatively complex class of models and has many parameters. Thus, for the simulations that we report in this study, we have adopted the somewhat modest goal of illustrating different types

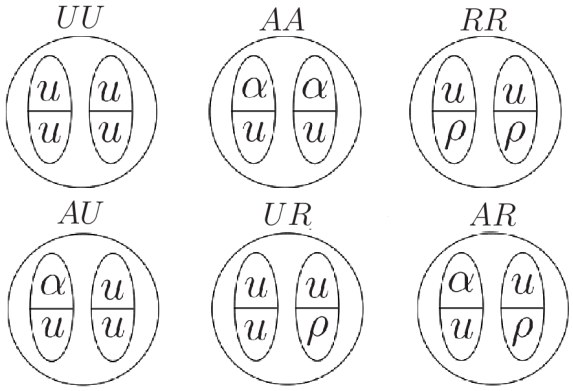

Figure 3.1: **6-state model.** Illustration of the states in the 6-state model. Circles represent nucleosomes. A nucleosome contains two histones copies represented by the vertically oriented ellipses. Each histone has two sites, one site (represented by the upper half of the ellipse) that can be either unmodified (symbolized by u) or have an active mark (symbolized by α), and another site (represented by the lower half of the ellipse) that can be either unmodified (symbolized by u) or have a repressive mark (symbolized by ρ). (Note that the physical nucleosome states labeled AU, UR and AR could be just as well depicted by interchanging the left and right ellipses within the respective circles.)

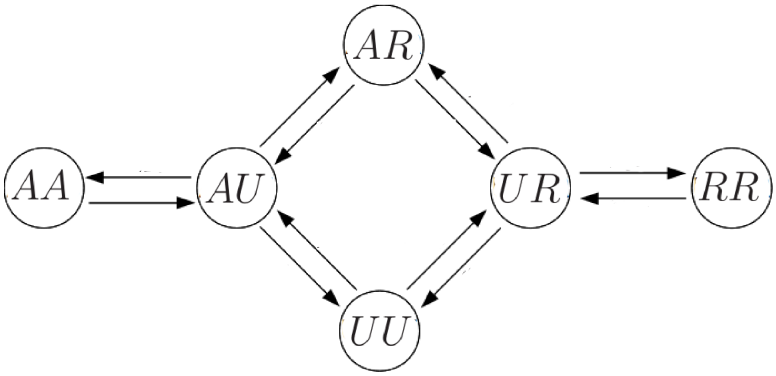

Figure 3.2: **Transitions for the 6-state model.** Transitions among the 6 distinct states in the 6-state model are indicated by arrows. The time step is supposed to be chosen small enough that only one site of the four nucleosome modification sites shown in Fig. 3.1 may change on each time step.

of dynamics that can arise when different nucleosome states interact and compete. With this goal in mind, we now seek an illustrative, but still somewhat plausible, reduction of our general 6-state model. Our reduction is based on the assumption, motivated in Appendix A, that the occurrence of nucleosome states having either active marks on both histones (AA in Fig. 3.1) or repressive marks on both histones (RR in Fig. 3.1) are unlikely. Thus we consider the idealized case where AA and RR states do not occur. Hence each nucleosome of the reduced model is in only one of 4 nucleosome states, namely AU, UR, AR and UU (see Fig. 3.3A). Referring to Fig. 3.1, we see that the four states have the following meanings.

AU: One histone has an active mark and the nucleosome's other three sites are unmodified.

UR: One histone has a repressive mark and the nucleosome's other three sites are unmodified.

AR: One histone has an active modification, while its other site is unmodified. The other histone has a repressive modification, while its other site is unmodified.

UU: All four sites of the nucleosome are unmodified.

Model Dynamics. During a cell cycle, we consider the time t states of modeled nucleosomes on our one dimensional lattice and update these states to new states at time $t+1$ through two probabilistic processes that we call "recruitment conversion" and "exchange conversion". At the conclusion of a cell cycle, "replication" occurs, following which a new cycle begins.

Table 3.1: Summary of parameters

Parameters	Physical description	Biological process simulated
r^i_{UR}, r^i_{UA}	Coefficient determining the probability of U converting to R/A via recruitment by the surrounding R/A marks	**Histone methylation spreading**: existing H3K27me3/H3K4me3 recruits methylase to methylate nearby nucleosomes.
r^i_{RU}, r^i_{AU}	Coefficient determining the probability of R/A converting to U via recruitment by the surrounding A/R marks	**Crosstalk between** A **and** R: existing H3K27me3/H3K4me3 recruits demethylase to demethylate nearby H3K4me3/H3K27me3.
p^i_{UR}, p^i_{UA}	Probability of U converting to R/A independent of the states of other nearby nucleosomes	**Nucleation :** continuous random histone marks placements at nucleosome site i
p^i_{RU}, p^i_{AU}	Probability of R/A converting to U independent of the states of other nearby nucleosomes	**Histone turnover rate** : histone marks can also be lost by random demethylation.
f^i_R, f^i_A	Fraction of R/A marks in nucleosomes within the recruitment range l of site i	We assume that the probability of recruitment (involved in the methylation spreading and crosstalk processes above) is proportional to the local density of the recruiting mark.
τ	The cell-cycle DNA replication period	**Cell cycle**
l	The nucleosome interaction distance	**Recruitment Range**

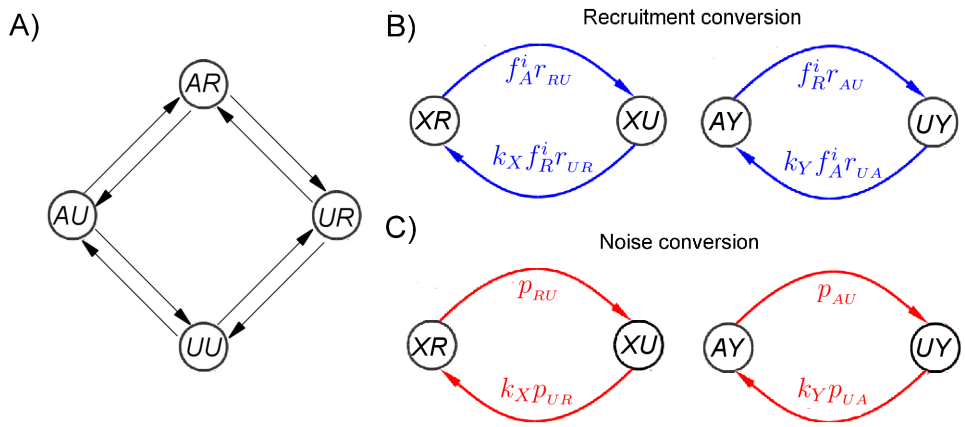

Figure 3.3: **Transitions for the reduced 4-state model.** (A) Transitions among the 4 distinct nucleosome states (i.e., AR, UR, AU, and UU) in the 4-state model. The time step is small enough that at most only one modification site of a nucleosome may change on each time step. (B) Transition probabilities between nucleosome states via recruitment conversions, where X can either be A or U while Y can either be R or U. k_X (k_Y)=2 if X (Y) is U, otherwise k_X (k_Y)= 1. Thus, as an example, the transition probability from the AR state to the AU state is the same as that from the UR state to the UU state. (C) Transition probabilities between nucleosome states via exchange conversions, where X can either be A or U while Y can either be R or U.

- *Recruitment.* This refers to the recruitment of histone marks to a nucleosome through interaction with neighboring nucleosomes. Recruitment at a site i depends on the states of the nucleosomes in an interval of length $2l$ centered at i, and we refer to l as the range of recruitment. We define f_X^i as the fraction of nucleosomes in this interval which carry a type-X histone mark, where the subscript f_x^i is $X = A, R$. If $l \leqslant i \leqslant N-l$, then the recruitment range will span $2l+1$ nucleosomes on our lattice. However, if i is too close to the beginning or the end of the lattice (i.e., $1 \leqslant i < l$ or $N - l < i \leqslant N$, respectively), then the recruitment range will include 'phantom' sites j not on the lattice ($j < 1$ and $j > N$, respectively), and for the purpose of determining f_X^i, we consider such phantom sites j to be in the UU state. The probability of recruitment conversion from U to X at site i is taken to be given by $f_X^i r_{UX}$, where r_{UX} is a constant describing the strength of the recruitment interaction. On the other hand, the probability of recruitment conversion from X to U (i.e., mark removal) depends on the concentration of histone marks which are opposite (rather than similar) to X (where we regard A and R as opposites). In this case, the conversion probability is taken to be given by $f_Y^i r_{XU}$, where Y is R if X is A and vice versa (see Fig. 3.3B). Note that in our model we allow r_{XU} to differ from r_{UX} because different enzymes are recruited for the addition and removal of histone marks.

- *Exchange.* Unlike the recruitment process, the exchange process refers to histone modifications which occur spontaneously, independent of the states of

nearby nucleosomes. The probabilities for exchange conversion are denoted by p_{UA}, p_{UR}, p_{AU}, and p_{RU} (see Fig. 3.2C). In particular, we think of p_{AU} and p_{RU} as corresponding to the histone turnover process, and p_{UA} and p_{UR} as corresponding to processes involving nucleation sites (See Table 3.1).

- *DNA replication.* When DNA replication occurs, we imagine that in the real situation the parental nucleosomes are randomly assigned to one of the two daughter strands at the same site as that which they occupied on the parental strand, while the corresponding site on the other strand is assigned an unmodified nucleosome (i.e., a nucleosome in the UU state). This scenario is supported by an experimental observation [59]. In our model, we do not follow both daughter strands. Rather we follow just one. Thus, with probability 1/2, our model replication process randomly replaces each nucleosome with an unmodified (UU) nucleosome. This model DNA replication occurs periodically with a period equal to the 'cell cycle time' τ. This is similar to how replication is modeled in [40].

In accord with the above recruitment and exchange processes, during a cell cycle, our model gives appropriate equations for the probabilities $P^i_{XY}(t+1)$ that nucleosome i is in state $XY = UU, AU, UR, AR$ at time $t+1$, given the state of the lattice at time t. After the probabilites $P^i_{XY}(t+1)$ are determined the state (UU, AU, UR or AR) of each nucleosome i is randomly chosen according to the probabilities $P^i_{XY}(t+1)$, thus determining the state at time $t+1$. Letting $\delta^i_{XY}(t) = 1$ if nucleosome i is in state XY, and $\delta^i_{XY}(t) = 0$ if nucleosome i is not in state XY, our model equations

for the probabilities are

$$
\begin{aligned}
P^i_{AU}(t+1) &= 2[f^i_A(t)r_{UA} + p^i_{UA}]\delta^i_{UU}(t) + [f^i_A(t)r_{RU} + p_{RU}]\delta^i_{AR}(t) \\
&+ \{1 - [f^i_R(t)(r_{AU} + r_{UR}) + p_{AU} + p^i_{UR}]\}\delta^i_{AU}(t), \\
P^i_{UR}(t+1) &= 2[f^i_R(t)r_{UR} + p^i_{UR}]\delta^i_{UU}(t) + [f^i_R(t)r_{AU} + p_{AU}]\delta^i_{AR}(t) \\
&+ \{1 - [f^i_A(t)(r_{RU} + r_{UA}) + p_{RU} + p^i_{UA}]\}\delta^i_{UR}(t), \\
P^i_{AR}(t+1) &= [f^i_R(t)r_{UR} + p^i_{UR}]\delta^i_{AU}(t) + [f^i_A(t)r_{UA} + p^i_{UA}]\delta^i_{UR}(t) \\
&+ \{1 - [f^i_A(t)r_{RU} + f^i_R(t)r_{AU} + p_{RU} + p_{AU}]\}\delta^i_{AR}(t), \\
P^i_{UU}(t+1) &= 1 - \{P^i_{AU}(t+1) + P^i_{UR}(t+1) + P^i_{AR}(t+1)\}.
\end{aligned}
$$

Consistent with our assumption that at most one site on a nucleosome can change state in one time step, our choice of parameters satisfies $r_{XY}, p_{XY} \ll 1$. Note that $f^i_A(t)$ and $f^i_R(t)$ depend on the lattice state in a neighborhood of site i within the range of recruitment specified in the second bullet above.

In section 3.3.3, where we treat localization of AR states, we allow the exchange transitions probabilities p^i_{XY} to vary from site to site, but everywhere else we consider p^i_{XY} to be the same at each site, $p^i_{XY} = p_{XY}$.

Simulation Parameters. To assign roughly reasonable values to the parameters r_{XY} and p_{XY}, we first consider that our model time step, $t \to t+1$, corresponds to a real time step $\Delta t = 2$ min. We have numerically verified that our simulation results are independent of our choice of Δt so long as Δt is sufficiently small. To estimate a rough range for the parameters r_{XY} and p_{XY}, we set p_{XY}, $r_{XY} \approx (\Delta t/T)$, where T is the characteristic time scale of the relevant process (see Table 3.1), and,

as required, the Δt that we have chosen is such that $\Delta t/T$ is small compared to one for all such processes. We fix as many parameters (Table 3.1) as possible using experimental information (see Table 3.2). Because the authors are not aware of any experimental measurements of the characteristic time for recruitment demethylation and methylation via exchange, we will consider these probabilities as free parameters in our numerical simulations below. Previous work [60] suggests that the loss of active marks is faster than the loss of repressive marks. In particular, it has been shown that nucleosome turnover is faster in regions bound by trithorax-group proteins. Therefore, we selected the model parameters so that all rates associated with active mark are faster than those associated with the repressive mark. Specifically, we assume that $r_{UR}/r_{UA} = p_{UR}/p_{UA} = r_{RU}/r_{AU} = p_{RU}/p_{AU} = 0.5$ in the simulation (when nonzero). Regarding the cell cycle, for embryonic stem cells the cell cycle length is about 12 hours, which, with our $\Delta t = 2$ min, corresponds to 360 time steps of our discrete time model per cell cycle. Finally, motivated by Ref. [59], we take $l = 2$, corresponding to a fairly short range of recruitment.

3.3 Results

We now illustrate the utility of our model by employing it to investigate dynamic changes of histone modification patterns. As described in the Introduction, both nucleation sites and recruitment of methylation may be involved in the establishment of bivalent domains. As noted above, we suggest that certain nucleosomes act as nucleation sites during the early stages of development. These nucleation

sites may be instrumental in the formation of bivalent domains. We incorporate nucleation sites into our model by assigning them a higher value of p_{UA} and p_{UR} than other sites, and we model the absence of nucleation sites by lowering its value of p_{UA} and p_{UR}.

In Sections 3.3.1 and 3.3.2, we discuss the formation and decay of AR states with different initial conditions in the absence of nucleation sites. In Section 3.3.3, we study the effect of nucleation sites on dynamics of the formation of AR states. Finally, in Section 3.3.4, we consider how varying the cell-cycle length affects AR states. Taken together, these analyses demonstrate the utility or our model for systematic investigation of the dynamic properties of bivalent domains.

3.3.1 Formation of AR States

The formation of bivalent domains has been experimentally observed in studies of the early stages of embryogenesis [64] and in studies of cell reprogramming [65]. In particular, studies of cell reprogramming observe this formation process to be gradual [66].

In this section we use our model to simulate the formation of regions that are dense with AR states, and we identify such regions with bivalent domains. In the simulations, we take $p_{UA} = p_{UR} = 0$ for all nucleosomes and fix r_{UA}= 0.046 (corresponding to an H3K4me3 methylation timescale of 30 mins) and $p_{AU} = 0.005$. Also, r_{AU} and r_{RU} are considered to be very small (for simplicity, we set r_{AU} = r_{RU} =0), so that the AR states can be established and persist for a long time.

For the initial state of the lattice in the simulations, we consider a situation where there are a relatively small number of nucleosomes in AR states, with all other nucleosomes initially in the UU state. In particular, we choose the initial number of AR nucleosomes to be five (out of the 80 nucleosomes on the lattice), and we study how AR states spread to other nucleosomes on the lattice. To investigate the effect of the initial spatial distribution of AR nucleosomes, we consider two extreme cases: *the localized case* in which all five initial AR state nucleosomes are located at five consecutive nucleosome sites in the center of the lattice, and *the delocalized case* in which the five initial AR state nucleosomes are located at equally spaced sites spanning the entire lattice (at sites 1, 20, 40, 60, 80).

Fig. 3.4 shows results for the space-time evolution of the distribution of nucleosomes for both *localized* (left column of figure panels) and *delocalized* (right column of figure panels) initial states. Fig. 3.4C shows space-time plots for the four types of nucleosomes in a typical single run, while Figs. 3.4A-B show average space-time plots of the level of AU and AR nucleosomes, that is, the fraction of runs for which the nucleosome is in the indicated state. The average level of UR nucleosomes (not plotted) is low everywhere all the time (dark blue, in terms of the color scale of Figs. 3.4A and B). Note that, in Figs. 3.4A-B, the regular drops of the levels of the indicated nucleosomes every 360 time steps (corresponding to the start of a new cell cycle) are due to the inserted of UU nucleosomes in the DNA replication process. In Fig. 3.4A, for the localized case (corresponding to the left panel figure), the AR nucleosomes spread over the lattice via a propagating front [55] manifested by the approximately straight lines of the color transition boundaries emanating from the

Table 3.2: Model parameters

Dynamical processes	Parameters	Characteristic time	References
Adding H3K4me3 marks via recruitment	r_{UA}	0.5 -6 hours	[61, 62]
Adding H3K27me3 marks via recruitment	r_{UR}	0.5- 6 hours	[61, 62]
Removing both H3K4me3 and H3K27me3 marks via exchange	p_{AU} and p_{RU}	1-24 hours	[61, 62]
Adding both H3K4me3 and H3K27me3 marks via exchange	p_{UA} and p_{UR}	not known	———
Removing both H3K4me3 and H3K27me3 marks via Recruitment	r_{AU} and r_{RU}	not known	———
Cell cycle length in human embryonic stem cells	τ	12 hours	[63]
Cell cycle length in human adult cells	τ	24 hours	[63]

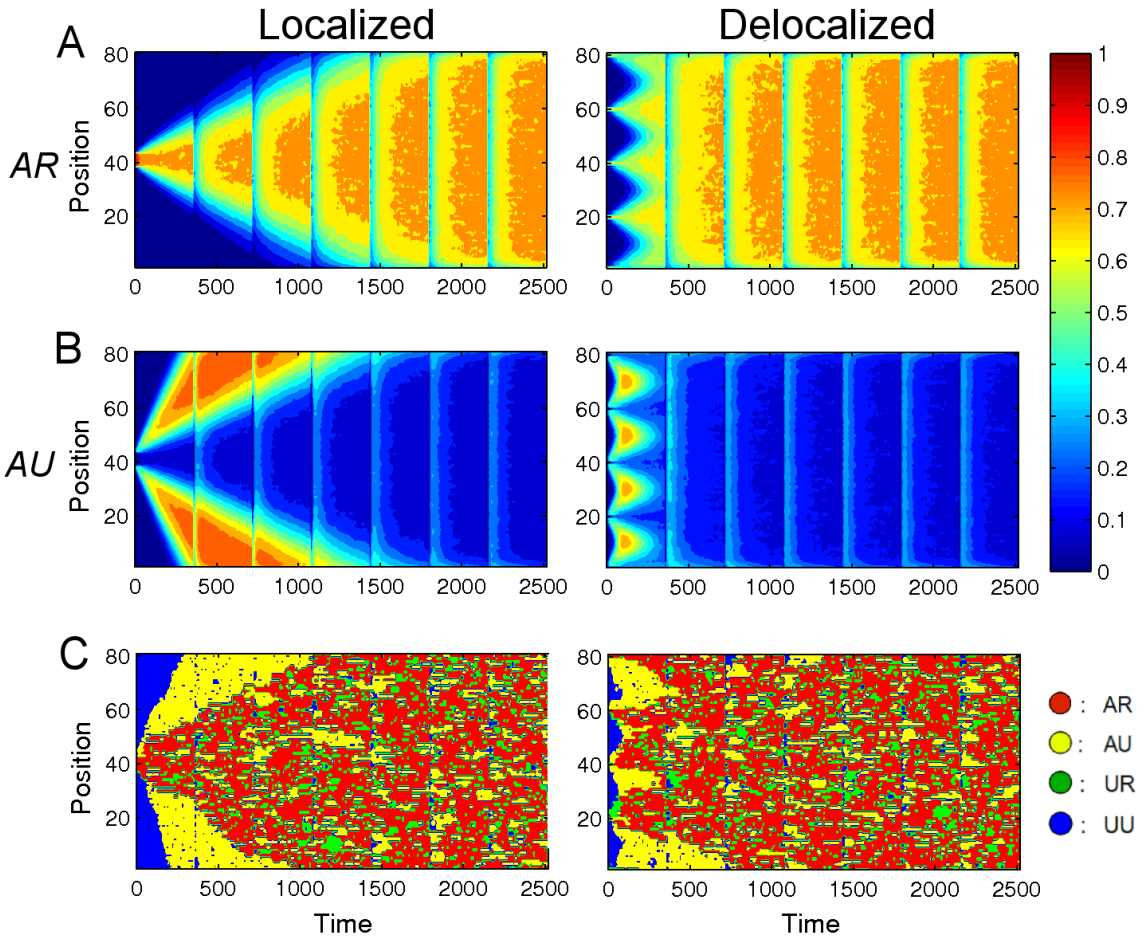

Figure 3.4: **Space-time plots for the formation of AR states.** Space-time plots of the average level of AR and AU nucleosomes for the *localized* and *delocalized* initial conditions are shown in (A) and (B). Here by 'level' we mean the fraction of runs for which the nucleosome is in the indicated state. These levels are computed by counting the indicated type of nucleosome in all runs at each position and time, and averaging over 2000 runs. The red color indicates a higher level of the indicated type of nucleosome while the blue color indicates a lower level of that type of nucleosome. (C) Space-time plots for a single run for the *localized* and *delocalized* initial conditions. AR, AU, UR, and UU nucleosomes are plotted in red, yellow, green, and blue, respectively.

space-time point at the center of the lattice at time $t = 0$. For the delocalized case (right panel of Fig. 3.4A), AR nucleosomes spread over the lattice via individual propagating fronts emanating from the five initial AR sites. These fronts merge near the end of the first cell-cycle (time ≈ 300), but the system takes longer time (time $\approx$ 1250) to reach a final equilibrium distribution. The model results show that, while the space time evolution of the distribution of AR nucleosomes is dependent upon the initial condition, the time it takes to establish a final equilibrium distribution is comparable and relatively long for both the localized and delocalized cases. This may have relevance to the experimental observation of Ref. [66] that the establishment of bivalent domains is gradual.

For the localized case, there appears to be two fronts, a fast $UU \rightarrow AU$ front (corresponding to the blue to yellow transition in the left panel of Fig. 3.4C), followed by a $AU \rightarrow AR$ front (yellow to red transition in the left panel of Fig. 3.4C) that propagates at a slower speed than the $UU \rightarrow AU$ front. The slow $AU \rightarrow AR$ front is clearly seen in the left panels of Figs. 3.4A and B, while the $UU \rightarrow AU$ front is evident in the left panel of Fig. 3.4B. These two fronts propagate symmetrically in space in the average space-time plots (Fig. 3.4A and 3.4B) but, due to fluctuations, more asymmetrically in space in the single run plot (see left panel of Fig. 3.4C). Examining a range of parameters, we find that the fastest front corresponds to either a $UU \rightarrow AU$ transition (as in Fig. 3.4B) or a $UU \rightarrow RU$ transition (not shown). For the delocalized case, we also observe that the spreading of the active marks is faster than that of the repressive marks. This can be easily seen from the typical single run plot in the right panel of Fig. 3.4C.

Finally, we also studied the effects of varying the number of AR nucleosomes in the initial condition on the above simulations. Using the same parameters values as above, we plot (Fig. 3.5) the final average fraction of AR nucleosome at the end of the final simulated cell cycle (10 cell cycles) as function of the initial number m of AR nucleosomes which are taken to occupy the m nucleosome sites in the center of the lattice. As shown in Fig. 3.5, the average fraction of final AR nucleosomes initially increases with increasing m. We observe that past $m \geq 4$ the value is essentially constant up to $m = 80$ with AR nucleosomes spanning the whole lattice. For a given m, each simulation can be categorized into two groups, (1) the final spatial average level of AR nucleosomes is approximately equal to the corresponding large m limiting value, or (2) all AR nucleosomes vanish. Thus at low m, the value plotted on the vertical axis of Fig. 3.5 can be thought of as the limiting larger-m value (basically the value at $m = 4$) multiplied by the fraction of runs in category (1). In the early stage of a simulation, the spreading of histone marks compete with the loss of histone marks via histone turnover. If either type of mark is lost totally, it cannot recover (i.e., the run is in category 2). On the other hand, we find that histone marks do not die out if there are enough of them on the lattice (the run is then in category 1). As a result, the average fraction of AR nucleosomes is larger with larger m, and with smaller p_{AU} and p_{RU} (compare the red and blue plots in Fig. 3.5). The above simulations suggest that in order for AR states to form when p_{UA} and p_{UR} are small, a sufficient number of initial AR nucleosomes is required. Taken together, these results have shown that the formation of bivalent domains undergoes two distinct phases: expansion and stabilization. In the expansion phase, the border

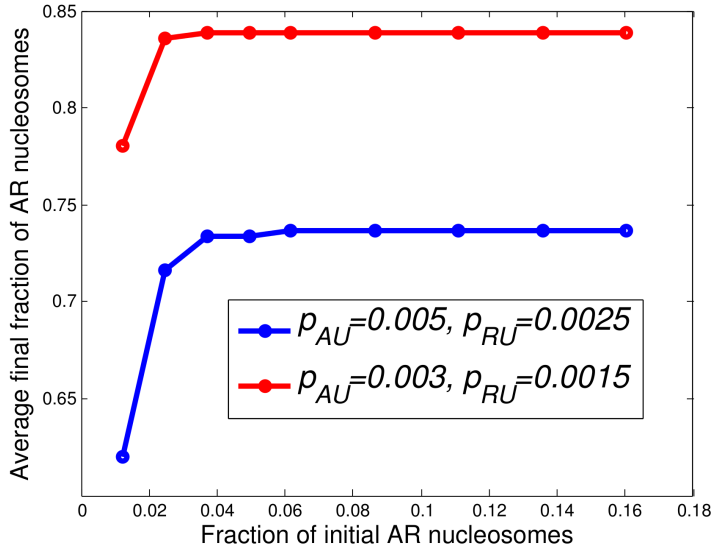

Figure 3.5: **Average final fraction of AR nucleosomes vs. the number of initial AR nucleosomes.** The average final fraction of AR nucleosomes is plotted as a function of m (the number of initial AR nucleosomes) for (p_{AU}, p_{RU}) being (0.003, 0.0015)(red) and (0.005, 0.0025)(blue). These levels of AR nucleosomes are computed by averaging the final number of AR nucleosomes in the simulations over 2000 runs.

of bivalent domains expands to neighboring nucleosomes. The expansion process is relatively fast (>10 nucleosomes per cell-cycle in our simulation) but quite noisy. As a result, only a sparse subset of nucleosomes are marked with the AR state. During the stabilization phase, the nucleosome state configuration is further refined and eventually reaches an equilibrium. Even then, the state of individual nucleosomes is still highly dynamic and equilibrium is only reached in the statistical sense.

3.3.2 Decay of AR States

In this section we use our model to simulate the decay of AR states. All parameters are the same as in section 3.3.1 except that r_{AU} and r_{RU} are taken to be non-zero. This is motivated by experimental findings that recruitment of demethylases is important for the decay of bivalent domains [50, 51], and occurs during cell differentiation. Also, we consider an initial condition in which all nucleosomes are in AR states. Results are shown in Figs. 3.6-3.8 for different values of r_{AU} and r_{RU} keeping their ratio fixed at $r_{AU}/r_{RU} = 2$.

Fig. 3.6 shows results for the space-time evolution of the distribution of all four nucleosome states AR, UR, AU, and UU for three values of $r_{AU} \equiv 2r_{RU}$. In Fig. 3.6A, for the case $r_{AU} = 0.004$, the initial level of AR nucleosomes rapidly (in about one cell-cycle) drops to a lower level of AR nucleosomes, but there still remains a substantial presence of AR nucleosomes which persists to the end of the run. In contrast, for both $r_{AU} = 0.016$ and $r_{AU} = 0.034$, where there is again similar very rapid decreases of the level of AR nucleosomes, now the final level is essentially

zero. In addition, it is seen that the level of AR nucleosomes takes longer to fully decay for $r_{AU} = 0.016$ than for $r_{AU} = 0.034$. The latter case is consistent with the experimental observations [38, 66] that an essentially complete loss of bivalent domain can occur very rapidly. To further explore how the decay of AR states depends on the recruitment demethylation rates, we plot the fraction of simulation runs that have at least one AR nucleosome on the lattice as a function of time in Fig. 3.7, and the final average fraction of AR nucleosomes (averaged over 1000 runs) as a function of r_{AU} in Fig. 3.8. Comparing Fig. 3.6A to Fig. 3.7, we observe that the fraction of runs with at least one AR nucleosome plotted in Fig. 3.7 shows a slower decay compared to the decay of AR levels in Fig. 3.6A. This suggests that lineage-control genes in bivalent domains may become active without the full destruction of repressive marks. In Fig. 3.8, as might be anticipated, we observe that, in general, smaller histone turnover (p_{AU}) and smaller recruitment demethylation rate give a higher final average fraction of AR nucleosomes. Also, the value of r_{AU} at which the average fraction of AR nucleosomes drops to zero is lower for larger p_{AU}. Our results suggest that a large recruitment demethylation rate in a cell is important for cell differentiation. This is consistent with experimental findings [50, 51].

In a real situation, a change from low to high values of the recruitment demethylation rates during cell differentiation will take place by processes not included in our model, and these processes may take some time. Thus our simulation use of constant non-zero initial r_{AU} and r_{RU} results in a determination of the characteristic decay time associated only with processes that are included in our model, and the true decay rate of AR state nucleosomes may be longer than this time due

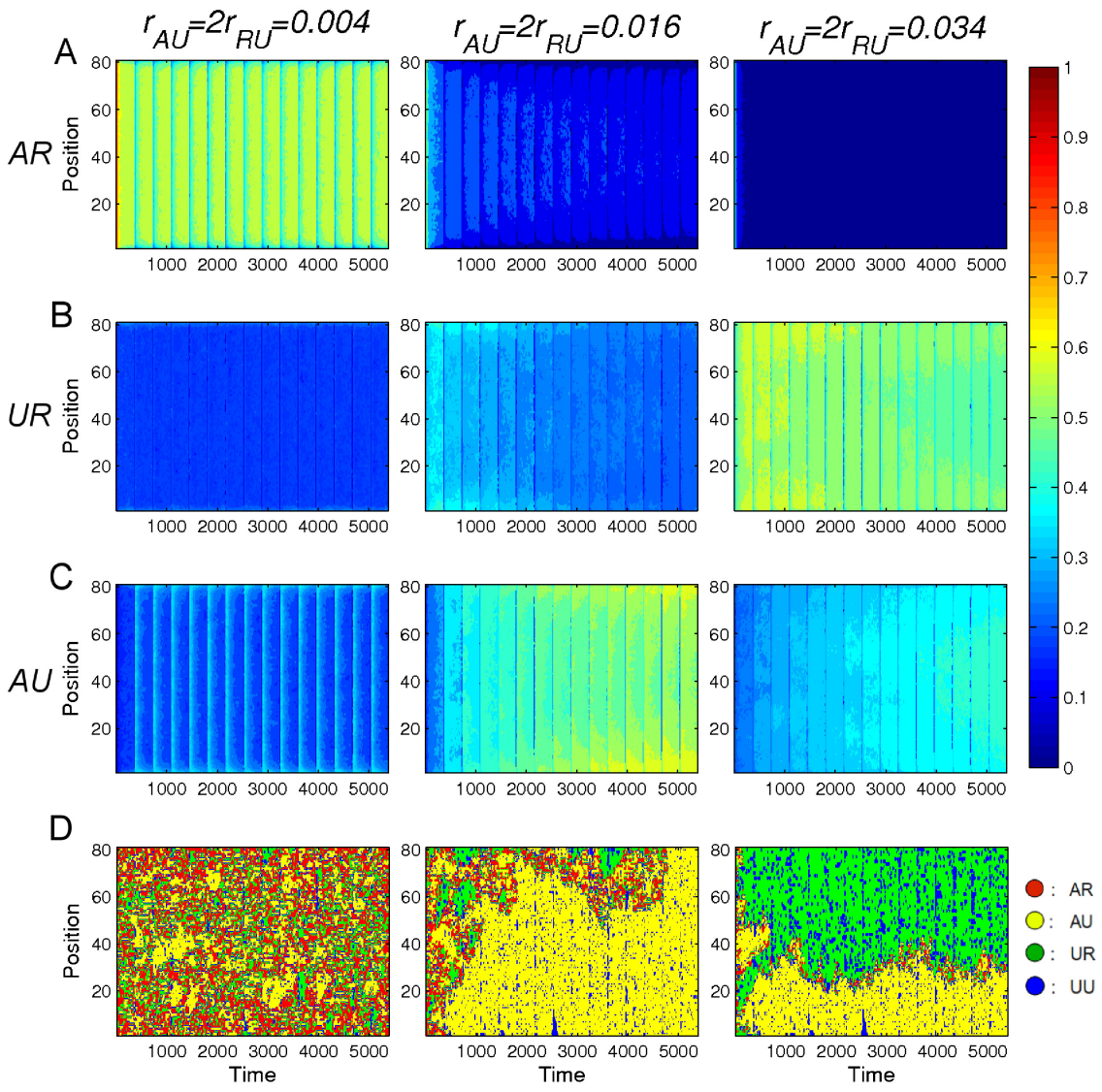

Figure 3.6: **Space-time plots for the decay of AR states.** In these plots, all nucleosomes are initially ($t = 0$) in the AR state. Space-time plots of the average level of AR, UR, and AU nucleosomes for $r_{AU} = 0.004, 0.016$, and 0.034 are shown in (A), (B), and (C), respectively. These plots are similar to Figure 3.4A and B. Here by level we mean the fraction of runs for which the nucleosomes is the indicated state. (D) Space-time plots for a single run with $r_{AU} = 0.004, 0.016, 0.034$. AR, AU, UR, and UU nucleosomes are plotted in red, yellow, green, and blue, respectively.

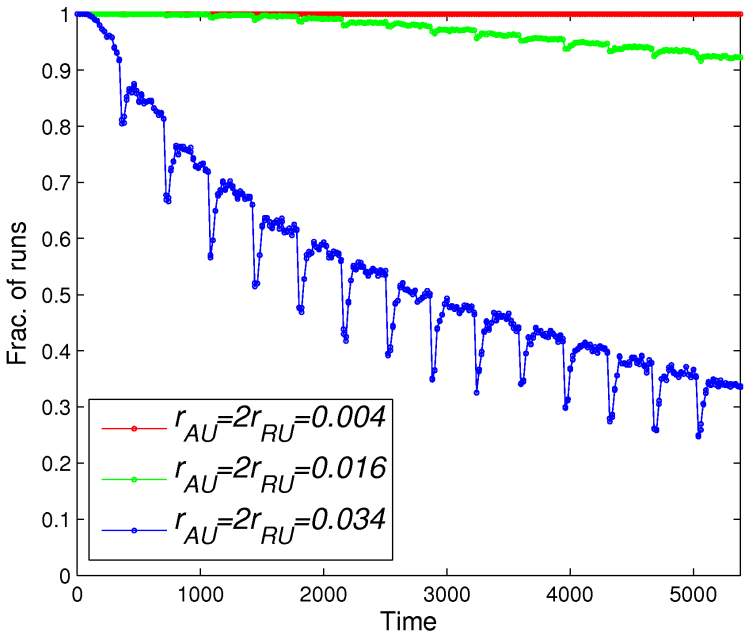

Figure 3.7: **Fraction of runs that have at least one AR nucleosome vs. time.** The fraction of runs that have at least one AR nucleosome on the lattice is plotted as a function of time for $r_{AU} = 0.004$, 0.016, and 0.034.

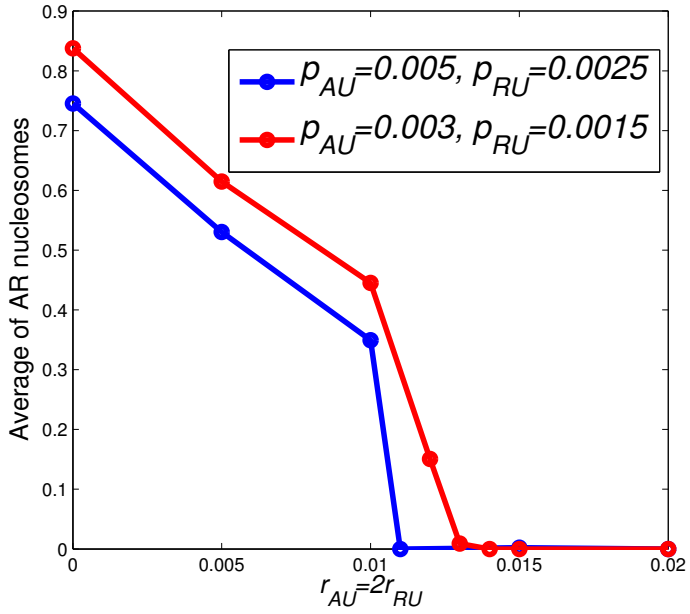

Figure 3.8: **Fraction of runs that have at least one AR nucleosome vs. time.** The average level of AR nucleosomes is plotted as a function of r_{AU} for both $p_{AU} = 0.003$ and 0.005. These levels of AR nucleosomes are computed by averaging the final number of AR nucleosomes in the simulations over all the runs and the whole lattice.

to the finite time for r_{AU} and r_{RU} to change. Overall, we observe that the decay determined from our model of AR state nucleosomes in response to high initial value of recruitment demethylation rate is relatively fast, as compared to the time that it takes to establish AR states spanning the lattice in Section 3.3.1. We conclude from this that processes included in our model do not prevent rapid decay of AR state nucleosomes, and that rapid decay, as seen in experiments [38], can occur in response to rapid increase of r_{AU} and r_{RU}.

In addition, it is interesting to emphasize the probabilistic nature of these results. For example, Fig. 3.6D shows results of typical single realizations. This figure also shows that the final state for $r_{AU} = 0.016$ is different from that for $r_{AU} = 0.034$. For the case $r_{AU} = 0.016$, we observe that AU nucleosomes are dominant in the lattice at the end of the simulation (see also the second panels of Figs. 3.6B and 3.6C). However, for the case of $r_{AU} = 0.034$ at long time, green regions of UR nucleosomes form at the upper edge (see third panel of Fig. 3.6D), while the AU nucleosomes are at the lower edges. This is because the AU and UR states can both be stable for this combination of parameters (see third panels of Figs. 3.6B and 3.6C). Our results suggest that the strength of the recruitment demethylation (i.e., the values of r_{AU} and r_{RU}) is not only important for the decay of bivalent domains, but also strongly influences the possible final state following decay.

3.3.3 The Localization of AR States

The next issue that we discuss is the effect of nucleation sites (i.e., in our model, $p_{UA} = p_{UR} > 0$ at these sites). The existence of such sites is suggested by the finding [32, 53] that DNA specific sequences can recruit protein binding factors like TF which in turn recruit histone marks to the DNA. In section 3.3.1, we took $p_{UA} = p_{UR} = 0$, and we found that AR nucleosomes either span the whole lattice or disappear. Although similar broad bivalent domains are observed, narrow bivalent domains are also detected in some experiments [31,52]. Although a recent model [39] has previously been used to simulate the dynamics of localized histone modification domains, that model allowed only a single type of histone modification, and therefore it cannot address the dynamics of bivalent domains. Using our model, we will be able to analyze interactions among the placements of active and repressive histone marks, histone turnover rate, and crosstalk between active and repressive histone marks. We consider p_{UA} and $p_{UR} > 0$ for the central nucleosome (corresponding to the case that the central nucleosome is a nucleation site). For the initial condition, we consider that there are five AR nucleosomes located at the five consecutive nucleosome sites in the center of the lattice, with all other nucleosomes initially in UU state. Using our previous parameter ratios (i.e., $r_{UR}/r_{UA} = p_{UR}/p_{UA} = r_{RU}/r_{AU} = p_{RU}/p_{AU} = 0.5$), we explore the parameter space regions for which our model reproduces narrow and broad distributions of AR nucleosomes.

We consider cases of both relatively small and relatively large recruitment demethylation rates (r_{AU} and r_{RU}). The former and latter choices are meant to

simulate cell environments far before, and during, cell differentiation, respectively. We run the simulations for four cell-cycles such that the averaged nucleosome state configuration reaches an equilibrium. In particular, a steady spatial distribution of AR nucleosomes seems to be reached within the first cell-cycle, and change very little thereafter. Therefore, the time for establishment of a highly localized AR distribution ($<$ 1 cell-cycle) is much shorter compared to that of establishing a very broad and uniform AR distribution (about 5 cell-cycles and see Figure 3.4A). For the case of small recruitment demethylation rate, Figs. 3.9A-B show plots of the fraction of AR nucleosomes averaged over 2000 simulations. Figs. 3.9A-B demonstrate narrow (left panels of Figs. 3.9A and 3.9B) and broad (right panels of Figs. 3.9A and 3.9B) distributions of AR nucleosomes. The widths of these bounded distributions reflect the balance between the continuous placement of histone marks on the nucleation site, the spreading of histone marks by the recruitment process, and the destruction of histone marks via exchange [39]. From the simulations, we find that the width of the distributions of AR nucleosomes depends more on $p_{_{AU}}$ and $p_{_{RU}}$, which they are inversely related to the width of the AR distribution. On the other hand, the amplitude of the distributions depends more on $p_{_{UA}}$ and $p_{_{UR}}$ (i.e., the continuous placements of histone marks on the center nucleosome) (Figs. 3.9A-B).

Next, we did simulations using the same parameters as in Fig. 3.9A but with larger recruitment demethylation rates ($r_{_{AU}}$ and $r_{_{RU}}$). The results are shown in Fig. 3.9C. Both of the corresponding distributions in Fig. 3.9A become narrower in Fig. 3.9C. In particular, the changes in the broad distribution (right panel) is particularly

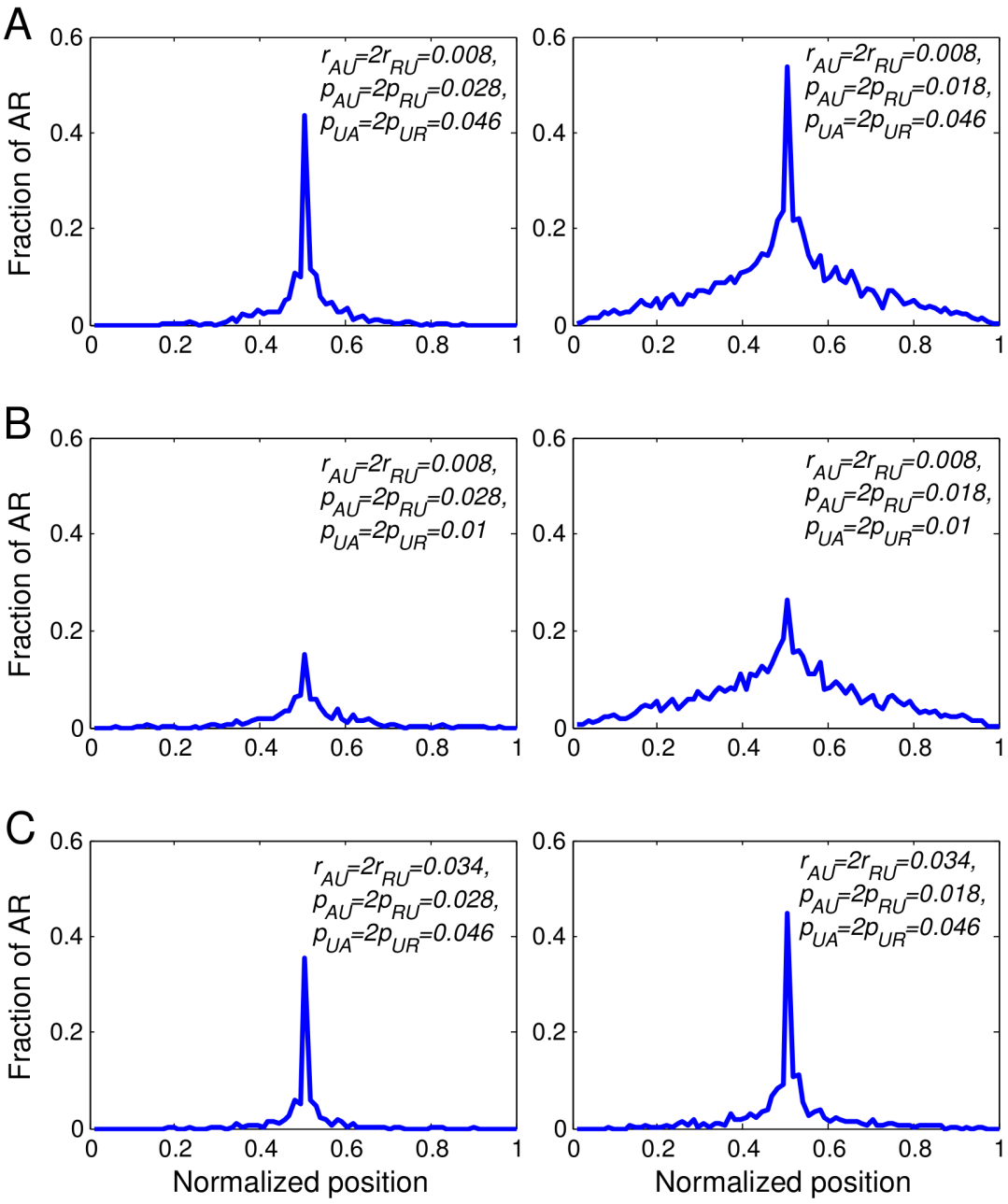

Figure 3.9: **Distributions of AR nucleosomes.** Distributions of AR nucleosomes are plotted at the final time of the simulations (time = 1800).

dramatic. This suggests that it may be easier to see changes in the broad bivalent domain than the narrow one during cell differentiation in experiments. Overall, our results demonstrate that nucleation sites can be responsible for the onset of bounded domains of AR nuclesomes. Also, narrow distributions can be obtained via either enhanced histone demethylation via exchange or via enhanced recruitment. We have also studied the distributions of AR nucleosomes, active marks, and repressive marks, using other reasonable parameter choices (see Figure 17 in Appendix A and the correponding texts). Taken together, these results suggest that highly localized bivalent domain patterns can be established surrounding nucleation sites, similar to the one-mark scenario described in previous study [39]. However, the local dynamics is more complex because multiple states are involved in the competition. The end configuration is an equilibrium resulting from the balance of multiple molecular forces.

3.3.4 The Effects of Cell-Cycle Length on the Stability of AR States

During DNA replication, the nucleosomes, along with their associated histone marks, must be dissociated from the mother strand. How these marks are reassembled to the newly synthesized strands remains poorly understood. Recent studies suggest that the nucleosome, along with their associated marks, are randomly distributed to daughter strands [59]. In this section, we use our model to study the impact of DNA replication on the level of AR nucleosomes.

We choose parameters which correspond to cell environments during the

formation of bivalent domains (see Fig. 3.10). Also, we assume that nucleation sites lose their properties at the very beginning of the simulations, so that there are no nucleation sites. We then vary the cell cycle lengths from 6 hours to 24 hours, which corresponds to varying the cell cycle length from that in stem cell to that in differentiated cells. We run the simulations for 10 cell cycles such that the average level of AR nucleosomes over a cell cycle reaches a stable value. Fig. 3.10 shows the average level of AR nuclesomes as a function of cell cycle length, where these levels are computed by averaging the number of AR nucleosomes at the end of the simulations over the lattice and over all simulation runs. Fig. 3.10 shows that the average level of AR nucleosomes is, in general, larger for longer cell-cycle. This result is expected, since there is more time for the lattice to recover from the loss of AR nucleosomes, caused by DNA replication, when the cell-cycle is longer. But the significance of cell-cycle length seems to be weaker for strong bivalent domains (blue curve in Fig. 3.10). This result is consistent with the experimental finding that higher levels of histone marking are observed when the length of the cell cycle increases [67].

3.4 Discussion

Development of computational models of bivalent domain dynamics can help to elucidate the mechanism of chromatin domain formation, and give insight for formulating and analyzing experimental studies. In this study we introduce a model that incorporates multiple histone marks on a nucleosome and the inter-

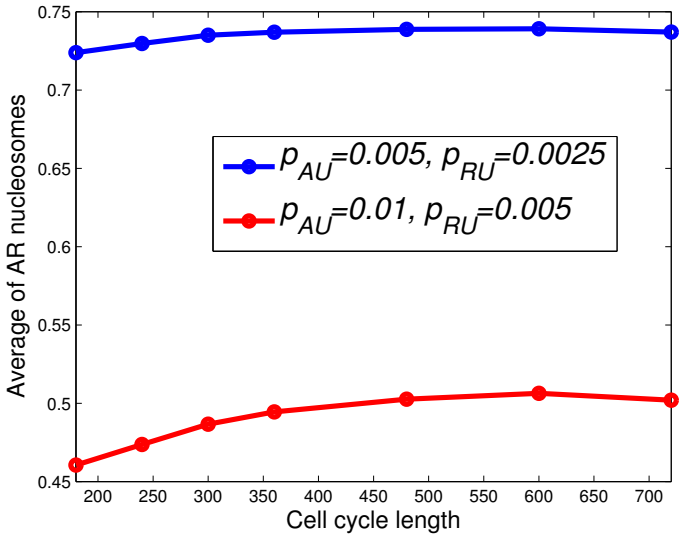

Figure 3.10: **Average AR nucleosome level vs. cell-cycle length.** The average AR nucleosome level is plotted as a function of cell-cycle length. These levels of AR nucleosomes are computed by averaging the final number of AR nucleosomes in the simulations over all the runs and the whole lattice. $r_{UA} = 2r_{UR} = 0.046$, $p_{UA} = p_{UR} = 0$, and $r_{AU} = r_{RU} = 0$.

actions among these marks. We have illustrated the potential use of our model by employing it to investigate the dynamics of bivalent domains, with the following results.

Our main conclusion is that the formation of bivalent domains are highly stochastic at individual nucleosomes, but reproducible patterns can been obtained by averaging a large number of simulations. Dynamic changes of these patterns are maintained by the subtle balance of the multiple factors including the exchange rate, recruitment, distribution of nucleation sites, and cell-cycle length, resulting high degree of plasticity which might be advantageous for facilitating smooth transitions between cell-states during development.

Our analysis suggests that the formation of bivalent domains is in general a slow, two-step process, which can be divided into an expansion and a stabilization phase. In contrast, the decay of bivalent domains, induced by demethylase activities, is much faster. This asymmetry between formation and decay dynamics may be an important feature for development control and perhaps needs to be taken into consideration into development epigenetic-based therapeutic approaches.

Specific epigenetic patterns can be established through targeted recruitment of chromatin regulators to specific genomic sequences. The effect of such nucleation sites on the establishment of highly localized epigenetic patterns has been studied via computational models in a number of previous studies [39, 68]. We have extended these investigations by considering multiple histone marks in our model. As expected, we found that the strength of a nucleation site plays an important role in maintenance of localized bivalent domains. In the absence of nucleation sites,

the bivalent domains either expands to the whole nucleosome array or disappears entirely. Our analysis is consistent with numerous experiment studies, which show that GC-rich DNA sequences are required for establishment of bivalent domains [54].

One limitation of our current model is that many kinetic parameters remain unknown, preventing us from making more quantitative predictions. Nevertheless, the major conclusions described above are robust with respect to parameter value changes therefore may reflect true biological principles. It will be interesting to test these principles by conducting quantitative experimental measurements.

Chapter 4: Dynamical Transitions in large Systems of Mean-Field-Coupled Landau-Stuart Oscillators: Extensive Chaos and Clumped States

4.1 Introduction

By a complex system we mean a system composed of a large number of interconnected dynamical units for which the overall macroscopic behavior is 'emergent' in that it is dependent crucially on interactions, and is not simply deducible from examination of the properties of the constituent uncoupled units. Understanding of the behavior of complex systems is a key issue in many fields, including physics, chemistry, neuroscience, social science, economics and biology. Thus there has been much activity in the quest for basic underlying phenomena, tools, and principles capable of advancing the study of such systems. One approach toward building up understanding is to investigate classes of systems that are particularly simple in some aspect. One of these classes is that of systems of N identical dynamical units ($N \gg 1$) that are coupled by a mean-field. For this class of systems, if each one of the coupled units has a real, time t, vector state denoted $\boldsymbol{x}_j(t)$ ($j = 1, 2, ..., N$), then the time evolution of $\boldsymbol{x}_j$ for $t \geq t_0$ depends on $\boldsymbol{m}(t)$ and $\boldsymbol{x}_j(t_0)$, where $\boldsymbol{m}(t)$ is

a mean field vector that is determined from some form of average of the $\boldsymbol{x}_j(t)$ over j. While the study of such systems can be viewed as a stepping-stone in the effort to understand complex systems with more complicated coupling, we also emphasized that mean-field-type coupling is a good approximation to many real situations (e.g., see [69–83]). In general, systems of identical mean-field coupled units can be represented as

$$\dot{\boldsymbol{x}}_j = \boldsymbol{F}(\boldsymbol{x}_j(t), \boldsymbol{m}(t), \boldsymbol{p}); j = 1, 2, ...N, \tag{4.1}$$

where $\boldsymbol{p}$ is a parameter vector.

Here we will study a particular instance of Eq.(4.1). However, we believe that the phenomena we find may be typical to many systems of the form (4.1). In particular, the system we study is that of mean-field-coupled Landau-Stuart oscillators [69], previously considered, e.g., in Refs. [84–92],

$$\dot{W}_j = W_j - (1 + iC_2)|W_j|^2 W_j + K(1 + iC_1)(\bar{W} - W_j), \tag{4.2}$$

where W_j is a complex number (corresponding to $\boldsymbol{x}_j$ in (4.1) being two dimensional), and the parameter vector corresponds to $\boldsymbol{p} = [C_1, C_2, K]^T$. $\bar{W}$ represents the mean field (analogous to $\boldsymbol{m}$ in (4.1)),

$$\bar{W} = N^{-1} \sum_j W_j. \tag{4.3}$$

A fundamental question that we address is that of whether the dynamics is intensive (also referred to as low dimensional) or extensive; i.e., whether the attractor dimension D remains limited by a constant bound as $N \to \infty$ (intensive), or whether, in contrast, D/N approaches a constant as $N \to \infty$ (extensive). For the

case of coupled Landau-Stuart oscillators, different dynamics which we claim can be viewed as including both intensive and extensive attractors, has been observed. More generally relevant to the dynamics of identical mean field coupled systems, Kaneko [93, 94], who considered large systems of identical coupled maps, found a intensive collective behavior called 'clustering', in which all the state components split into a small number of different clumps and in which the components in each clump behave identically. The dynamics in this clumped phase can be regarded as low dimensional (intensive) since the system state can be specified by giving the states of the small number of clumps. Also, for some parameter values, collective behavior can emerge in which each component behaves differently and in an irregular manner (e.g., Refs. [84–87] for Eqs. (4.2)), which we identify (Sec. 4.5) as corresponding to extensive chaos .

One key issue is the possible existence of dynamical phase transitions from an intensive phase (clustering) to an extensive chaotic phase. The possibility of this type of dynamical phase transition was originally pointed out in the early 1990's by Nakagawa and Kuramoto [84–86] in the particular context of coupled Landau-Stuart oscillators (see also [87]), but, to the best of our knowledge, it has not received further attention. In particular, in our paper, we will be interested in *following a specific identified attractor as a parameter is continuously varied* with the goal of seeing how this identified attractor evolves as the parameter varies. We emphasize that the question of how our identified attractor evolves with continuous parameter change cannot be fully addressed by the common procedure of investigating the attractor (or attractors) that result from some given initial condition (or set of

initial conditions) that remains fixed as many simulations are independently run from $t = 0$ with different parameter values.

Related to the above point, another fundamental question for such systems concerns the clumped dynamical phase. While clump dynamics is inherently low dimensional, for large N, even considering the number of clumps as fixed, there can be very many attractors corresponding to different population fractions of the N dynamical units in each clump. One might then ask whether there are circumstances that lead to selection of particular population distributions among clumps (i.e., selection of a particular attractor). Here we will show that, when there are two clumps and the system is subject to slow adiabatic parameter change, such population distribution selection can occur by a mechanism that we refer to as 'marginal stability'. Furthermore, we show that this mechanism is the key ingredient needed for understanding an explosive transition from low dimensional behavior to extensive chaos. Finally, we use an analogy to low-dimensional randomly forced systems [95, 96] to apply the Kaplan-Yorke dimension formula [97, 98] to a suitable reduced set of Lyapunov exponents, and we show that the resulting prediction of the extensive dimensionality (D/N for large N) is consistent with numerical computations of the information dimension of the attractor.

4.2 Background and Formulation

In this study, we consider mean-field coupling of a large number of identical Landau-Staurt oscillators, as described by Eqs. (4.2) and (4.3) with the oscillators

all identical (i.e., K, C_1 and C_2 are the same for all j). In our numerical experiments, we explore the types of attractors that occur and how the system behavior changes with change of a parameter. Specifically, we set $C_1 = -7.5$ and $C_2 = 9.0$, and vary K.

We now give a brief overview of our numerical experiments and main findings. Our numerical experiments reveal system behaviors similar to the previous studies [84–87]. Some representative results are given in Fig. 4.1 which shows the states of each of the $N = 3000$ oscillators in the complex plane for three different parameter values $K = 0.1$, $K = 0.74$ and $K = 0.95$, plotted at some fixed time (a 'snapshot'). For K smaller than about 0.4, the system is in an incoherent state (i.e., $\bar{W} \cong 0$) which is shown in Fig. 4.1(a). In the incoherent state, $|W_j| \cong (1 - K)$ for each oscillator $j = 1, 2, ..., N$, and $\sum W_j = \bar{W} \cong 0$, since the phases of the oscillators are apparently distributed randomly with uniform density in [0, 2π]. In contrast, for K very large, a single locked state exists where all oscillators have the same identical behavior (i.e., $W_j = e^{iC_2 t} = \bar{W}$ for all j). These incoherent states and locked states have been discussed in previous studies [86, 87]. In particular, the stability of these states can be calculated analytically. At $K = 0.74$, we observed the existence of what we call the extensively chaotic state in which all oscillators behave differently (Fig. 4.1(b)) and the macroscopic mean field $\bar{W}$ varies irregularly in time. As shown in Sec. 4.5, the extensively chaotic state is high-dimensional, and we can observe what appears to be a fractal distribution in the snapshot of the oscillator states (for a more detailed discussion see Sec. 4.5). (Previous work in Refs. [85–87] considered parameter values for which fractal structure was much less apparent and was not

explicitly noted.) At $K = 0.95$, we observed the existence of a clumped state (Fig. 4.1(c)). In the clumped state of Fig. 4.1(c), there are two clumps, where oscillators in the same clump all behave identically. We will discuss and analyze clumped states in the next section [97].

The dynamics of the attractors can be quantified by Lyapunov exponents. Consider a system that is governed by Eq. (4.2), and has a solution

$$W_j(t) = W_{j,0}(t). \tag{4.4}$$

To calculate its Lyapunov exponents, we initially perturb $W_{j,0}$ to $W_{j,0} + \delta W_j$. Considering δW_j to be infinitesimal, we obtain a set of perturbation equations for δW_j,

$$\begin{aligned} \delta\dot{W}_j = \quad & [1 - 2(1 + iC_2)|W_{j,0}|^2 - K(1 + iC_1)]\delta W_j \\ & -(1 + iC_2)W_{j,0}^2\delta W_j^* + K(1 + iC_1)\delta\bar{W}, \end{aligned} \tag{4.5}$$

where $j = 1, 2, ..., N$ and $\delta\bar{W} = N^{-1}\sum_j \delta W_j$. The Lyapunov exponents (λ) are given by

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \ln \frac{\delta(t)}{\delta(0)}, \tag{4.6}$$

where $\delta(t) = \sqrt{\sum_j |\delta W_j(t)|^2}$. Depending on the initial set of perturbations $\{\delta W_j(0)|j = 1, 2, ..., N\}$, the Lyapunov exponent in (4.6) can in principle take on $2N$ possible values. However, for a typical random choice of the initial condition $\delta W_j(0)$ ($j = 1, 2, ..., N$), Eq. (4.6) will give the largest Lyapunov exponent.

In the clumped states, we also divide the Lyapunov exponents into two types, one of which determines the internal stability of a clump, while the other determines the stability of the clump orbits. For the case of two clump states which we will

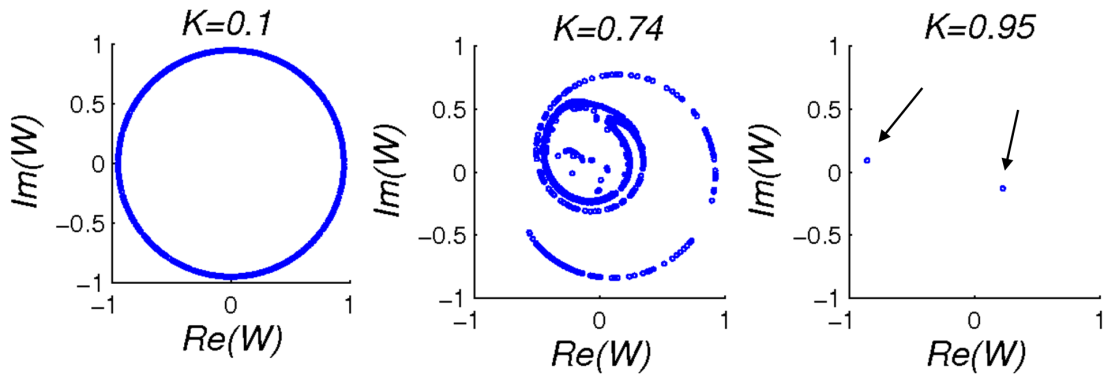

Figure 4.1: These figures show three snapshot attractors which are simulated by using different values of K; (a) correspond to an incoherent state at $K = 0.1$; (b) corresponds to an extensive chaotic state at $K = 0.74$; (c) corresponds to a two-clump states at $K = 0.95$.

henceforth focus on, we distinguish the two clumps by the labels a and b, where we take the larger clump (i.e., the clump with the most oscillators) to be clump a, while we take the smaller one to be clump b, and W_j is either equal to W_a or W_b for all j. As a result, Eqs. (4.2) and (4.3) show that the motions of these clumps are governed by a reduced set of two equations,

$$\begin{aligned} \dot{W}_a &= W_a - (1+iC_2)|W_a|^2 W_a + (1+iC_1)(\bar{W} - W_a), \\ \dot{W}_b &= W_b - (1+iC_2)|W_b|^2 W_b + (1+iC_1)(\bar{W} - W_b), \end{aligned} \tag{4.7}$$

where $\bar{W} = f_a W_a + f_b W_b$. Here f_a and f_b are the fractions of oscillators in clumps a and b, respectively (i.e., $f_{a,b} = N_{a,b}/N$ where N_a and N_b are the numbers of oscillators in clumps a and b, and $f_a > f_b$). In the following, we call the system described by Eq. (4.7) the 'two-clump system', while the system described by Eqs. (4.2) and (4.3) is called the 'full system'.

To determine the internal stability of a clump, say clump a, we perturb the states of each oscillator in clump a, $W_j(t) = W_a(t) + \delta W_j(t)$, and we choose the initial perturbations to oscillators in clump a to statisfy $\sum_j \delta W_j(0) = 0$ with $\delta W_j = 0$ for all oscillators in clump b. Inserting this into the full system Eq. (4.3) and linearizing with respect to δW_j, we find, by summing over j in clump a, that $\delta\bar{W} = N^{-1}\sum_j \delta W_j$ remains zero for all time, and that each of the δW_j satisfies the *same* equation. Using the notation $\delta\widetilde{W}_a$ to denote any one of these oscillator perturbations for clump a, the evolution of $\delta\widetilde{W}_a(t)$ is governed by the equation,

$$\dot{\delta\widetilde{W}}_a = \delta\widetilde{W}_a - K(1+iC_1)\delta\widetilde{W}_a - (1+iC_2)[2|W_a|^2\delta\widetilde{W}_a + W_a^2\delta\widetilde{W}_a^*], \tag{4.8}$$

where $\delta\widetilde{W}_a^*$ denotes the complex conjugate of $\delta\widetilde{W}_a$. It is convenient to regard $\delta\widetilde{W}_a$

and $\delta\widetilde{W}_a^*$ as if they were independent and to rewrite Eq. (4.8) in the form,

$$\begin{pmatrix} \dot{\delta\widetilde{W}_a} \\ \dot{\delta\widetilde{W}_a^*} \end{pmatrix} = M \begin{pmatrix} \delta\widetilde{W}_a \\ \delta\widetilde{W}_a^* \end{pmatrix}. \tag{4.9}$$

Similarly, we can derive the same perturbation equation for $\delta\widetilde{W}_b$ corresponding to pertubations of oscillators in clump b. We call the Lyapunov exponents derived from Eq. (4.9), the clump integrity exponents, λ_{CI}^{σ}, where $\sigma = a$ or b corresponding the exponents for clumps a or b, respectively. λ_{CI}^{a} and λ_{CI}^{b} each have two values for the two-clumped states (because $\delta\widetilde{W}_a$ and $\widetilde{W}_b$ are complex and hence two-dimensional). We find (see next section) that, for the two clump solutions that we investigate, there are two types of clumped states: (i) a state in which $W_a(t) = D_a \exp(i\Omega t)$, $W_b(t) = D_b \exp(i\Omega t)$ where Ω is a real constant and $D_{a,b}$ are complex constants; this case corresponds to a fixed point solution in the frame rotating with the frequency Ω; and (ii) a solution in which $|W_{a,b}(t)|$ varies periodically with time, and, again transforming to a suitable rotating frame at some frequency Ω, the transformed $W_a(t)$ and $W_b(t)$ are periodic. Assuming that this type of rotation transformation has been performed, M is constant (periodic) in time for case (i) (case (ii)). For the case where M in Eq. (4.9) is time-independent, the two λ_{CI} are equal to the magnitudes of the eigenvalues of M. For the case that M is time-dependent, the largest λ_{CI}^{σ} can be computed by Eq. (4.6) with $\delta(t) = |\delta W_\sigma|$. The sum of the larger and smaller λ_{CI}^{σ} is equal to the time average of the divergence of the 'flow' given by Eq.(9). This divergence is

$$\frac{\partial\dot{\delta\widetilde{W}_\sigma}}{\partial\delta\widetilde{W}_\sigma} + \frac{\partial\dot{\delta\widetilde{W}_\sigma^*}}{\partial\delta\widetilde{W}_\sigma^*} = 2(1-K) - 4|W_\sigma|^2. \tag{4.10}$$

Therefore, we can calculate the smaller λ_{CI}^{σ} by subtracting the larger λ_{CI}^{σ} from $2(1-K) - 4\langle|W_\sigma|^2\rangle$, where $\langle ...\rangle$ denotes the time average. Note that if the larger Lyapunov exponent for internal clump stability satisfies $\lambda_{CI}^{\sigma} > 0$, then clump σ tends to fly apart (loose its integrity). (Referring back to Eqs. (4.5) and (4.6) where we noted that there were $2N$ solutions for λ, and observing that $\sum_j \delta W_j = 0$ for j in clump a represents two real constraints on the $2N_a$ real variables $Re(\delta W_j)$ and $Im(\delta W_j)$, we conclude that λ_{CI}^{a} has multiplicity $(2N_a - 2)$, and similarly that λ_{CI}^{b} has multiplicity $(2N_b - 2)$, thus together accounting for $(2N-4)$ of the $2N$ possible Lyapunov exponents.)

For the other type of Lyapunov exponent, we derive the perturbation equation similar to the derivation of Eq. (4.8), but now setting all the δW_j in a clump to be equal, $\delta W_j = \delta W_a$ for all oscillators j in clump a, and $\delta W_j = \delta W_b$ for all oscillators in clump b. In this case, $\delta \bar{W} = f_a \delta W_a + f_b \delta W_b \neq 0$, and we can interpret δW_a and δW_b as displacement perturbations of the *whole* clump a and of the *whole* clump b, respectively. We call these Lyapunov exponents the clump system orbit stability exponents (λ_{SO}). There are four possible values of λ_{SO}, corresponding to the four real perturbation variables $Re(\delta W_{a,b})$ and $Im(\delta W_{a,b})$.

Rather than working directly with Eq. (4.7), to calculate all the λ_{SO} for the two clump states, we first reduce the number of real equations from four to three. We let $W_a = \rho_a e^{i\theta_a}$ and $W_b = \rho_b e^{i\theta_b}$, where ρ_a, ρ_b, θ_a, and θ_b are all real. Also, we define the relative phase difference $\phi = \theta_a - \theta_b$. As a result, Eq. (4.7) yields three coupled equations (as opposed to the four coupled equations that would result from

taking the real and imaginary parts of Eq. (4.7)),

$$\begin{aligned}
\dot{\rho_a} &= [Kf_a - K + 1]\rho_a - \rho_a^3 + Kf_b\rho_b(\cos\phi + C_1\sin\phi), \qquad (4.11)\\
\dot{\rho_b} &= [Kf_b - K + 1]\rho_b - \rho_b^3 + Kf_a\rho_a(\cos\phi - C_1\sin\phi),\\
\dot{\phi} &= KC_1(f_a - f_b) - C_2(\rho_a^2 - \rho_b^2) + KC_1\cos\phi\left(\frac{f_b\rho_b}{\rho_a} - \frac{f_a\rho_a}{\rho_b}\right) - K\sin\phi\left(\frac{f_b\rho_b}{\rho_a} + \frac{f_a\rho_a}{\rho_b}\right).
\end{aligned}$$

Similar to Eq. (4.8), we can derive the perturbation equations for $\delta\dot{\rho_a}$, $\delta\dot{\rho_b}$, and $\delta\dot{\phi}$. There are three λ_{SO} that result, corresponding to the three equations in (4.11). (There is also a forth Lyapunov exponent of zero for the original four dimensional system (4.7) that corresponds to an infinitesimal rigid phase rotation of the system $(\delta\theta_a,\ \delta\theta_b)\rightarrow(\delta\theta_a + \delta\eta,\delta\theta_b + \delta\eta)$ which we note, does not change the value of $\delta\phi$. This extra exponent does not affect our discussion and will henceforth be ignored. Correspondingly, we also note that, by use of the variable $\phi = \theta_a - \theta_b$, any constant rotation of W_a and W_b in the complex plane (i.e., a common factor of $e^{i\Omega t}$) is removed.) The largest λ_{SO} is computed by Eq. (4.6), with $\delta(t) = \sqrt{\delta\rho_a^2 + \delta\rho_b^2 + \delta\phi^2}$. To calculate the negative of the smallest λ_{SO}, we integrate the perturbation equation derived from (4.11) with a typical initial perturbation following a saved forward unperturbed orbit on the attractor backwards in time. Similar to the calculation of λ_{CI}^{σ}, we can compute the divergence for the perturbation equations derived from (4.11), and the middle λ_{SO} can then be obtained by subtracting the sum of the largest and smallest λ_{SO} from the time average of the divergence.

4.3 Two-Clump State Attractors

In this section, we focus on the 'two-clump system' described by Eqs. (4.11). In particular, we study the possible two clump attractors in the (f_a, K) parameter space. To do this, we solve the two-clump system in Eq. (4.11) numerically and compute λ_{SO} for these solutions. We observed both fixed-point solutions and periodic-orbit solutions, but no chaotic solutions. We emphasize that such solutions of the two clump system may be unphysical, since the individual clumps may or may not be internally stable; i.e., it may be the case that one of the λ_{CI}^{a} or λ_{CI}^{b} is positive. In this section we do not consider λ_{CI}^{σ}. Thus, when we refer to stability in this section, we are refering to stability as determined by λ_{SO} (clump internal stability, as determined by λ_{CI}, is considered in Sec. 4.4).

To find the fixed point solutions of Eq. (4.11), we set $\dot{\phi} = \dot{\rho_a} = \dot{\rho_b} = 0$, for which Eqs. (4.11) become

$$
\begin{aligned}
0 &= [Kf_a - K + 1]\rho_a - \rho_a^3 + Kf_b\rho_b(\cos\phi + C_1\sin\phi), \qquad (4.12)\\
0 &= [Kf_b - K + 1]\rho_b - \rho_b^3 + Kf_a\rho_a(\cos\phi - C_1\sin\phi),\\
0 &= KC_1(f_a - f_b) - C_2(\rho_a^2 - \rho_b^2) + KC_1\cos\phi\left(\frac{f_b\rho_b}{\rho_a} - \frac{f_a\rho_a}{\rho_b}\right) - K\sin\phi\left(\frac{f_b\rho_b}{\rho_a} + \frac{f_a\rho_a}{\rho_b}\right).
\end{aligned}
$$

Note that a possible solution to Eqs. (4.12) occurs for $\rho_a = \rho_b = 1$, $\phi = 0$, which corresponds to a single clump fixed point solution. However, we are interested in solutions of (4.12) representing two clump states. We reduce the number of equations in Eq. (4.12) by eliminating the variable ϕ. To do this, we first solve for $\cos\phi$ and $\sin\phi$ from the first two equations in (4.12). We then substitute these solutions into

the relation, $\cos^2\phi + \sin^2\phi = 1$, to obtain

$$\begin{aligned} 4C_1^2K^2f_a^2f_b^2xy &= C_1^2[f_ax^2 + f_by^2 - f_ax - f_by + Kf_af_b(x+y)]^2 \\ &+ [f_ax^2 - f_by^2 - f_ax + f_by + Kf_af_b(x-y)]^2, \end{aligned} \tag{4.13}$$

where we have introduced $x = \rho_a^2$ and $y = \rho_b^2$. Also, we substitute the solutions of $\cos\phi$ and $\sin\phi$ into the third equation in Eq. (4.12), to obtain

$$\begin{aligned} & 2C_1f_af_bxy[C_2(x-y) - KC_1(f_a - f_b)] \\ = \; & C_1^2(f_by - f_ax)[f_ax^2 + f_by^2 - f_ax - f_by + Kf_af_b(x+y)] \\ - \; & (f_by + f_ax)[f_ax^2 - f_by^2 - f_ax + f_by + Kf_af_b(x-y)]. \end{aligned} \tag{4.14}$$

Two clump fixed point solutions can occur at the intersection of y versus x plots of Eqs. (4.13) and (4.14). An example with $f_a = 0.82$, $K = 0.78$ is shown in Fig. 4.2 in which Eqs. (4.13) and (4.14) are plotted in red and blue, respectively. There is an intersection point at $x = 1$ and $y = 1$ corresponding to a single clump state. There are three other intersection points (shown in the figure as black dots) that are also consistent with Eqs. (4.12) and that thus correspond to two-clump state solutions. Stability analysis reveals that only the two intersection points labeled A and C are stable solutions of the two clump system (4.11), i.e., all λ_{SO} for these solutions are negative.

We determine fixed point solutions (e.g., as done in Fig. 4.2) and their stability (i.e., by calculating λ_{SO}) for different K and f_a. Results are shown in Fig. 4.3, where we denote the fixed point solutions A or C by fp_A or fp_C, respectively. Referring to Fig. 4.3(a), fp_A is stable in the region above the solid and dashed blue

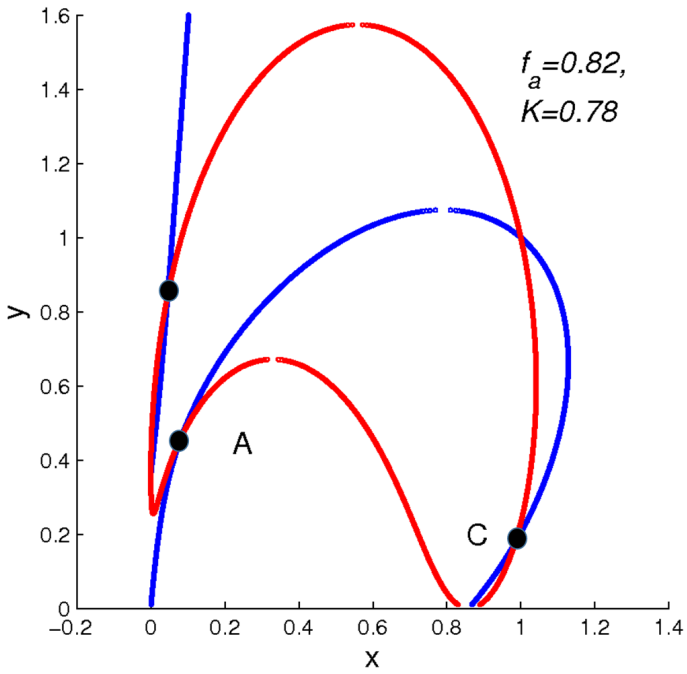

Figure 4.2: Plots of Eq. (4.13) (red) and (4.14) (blue) for $f_a = 0.82$ and $K = 0.79$.

lines. Below these blue solid and dashed lines, a stable solution for fp_A does not exist. In particular, the solid blue line corresponds to a saddle-node bifurcation of fp_A, while the dashed line corresponds to a Hopf bifurcation of fp_A. [A co-dimension two bifurcation occurs at the point where the saddle-node bifurcation coincides with the Hopf-bifurcation. At this point, one λ_{SO} is zero while the real parts of the other two λ_{SO} are zero.] Similarly, as shown in Fig. 4.3(b), fp_C is stable (unstable) in the region above (below) the solid green line, at which a Hopf-bifurcation occurs.

Our computational procedure for investigating periodic orbit attractors is as follows. We first obtain numerical solutions of Eqs. (4.11) using many different initial conditions for every selected pair of f_a and K. Next, we numerically track our discovered periodic orbit attractors with the system undergoing 'slow adiabatic parameter change'. In our implementation of what we call slow adiabatic parameter change, after we run the numerical code solving Eqs. (4.11) for a long enough time that the orbit has settled onto a periodic orbit attractor, we then change the parameters by a small amount, $K \to K + \delta K$, $f_a \to f_a + \delta f_a$, and we perform a new simulation with these shifted parameters, using for the initial condition the system state (ρ_a, ρ_b, ϕ) at the end of the previous run. By repeating this procedure through many parameter shifts, we continuously track an identified attractor through a path in the (K, f_a) parameter space. By doing this and computing the λ_{SO} of the solutions, we have explored the stability boundaries of our periodic orbit attractors. The results are shown in Fig. 4.3. In particular, we find that all the periodic attractors observed in our simulations can be thought of as originating from bifurcations of fixed points solutions. In Fig. 4.3(a), a periodic orbit attractor denoted po_A is

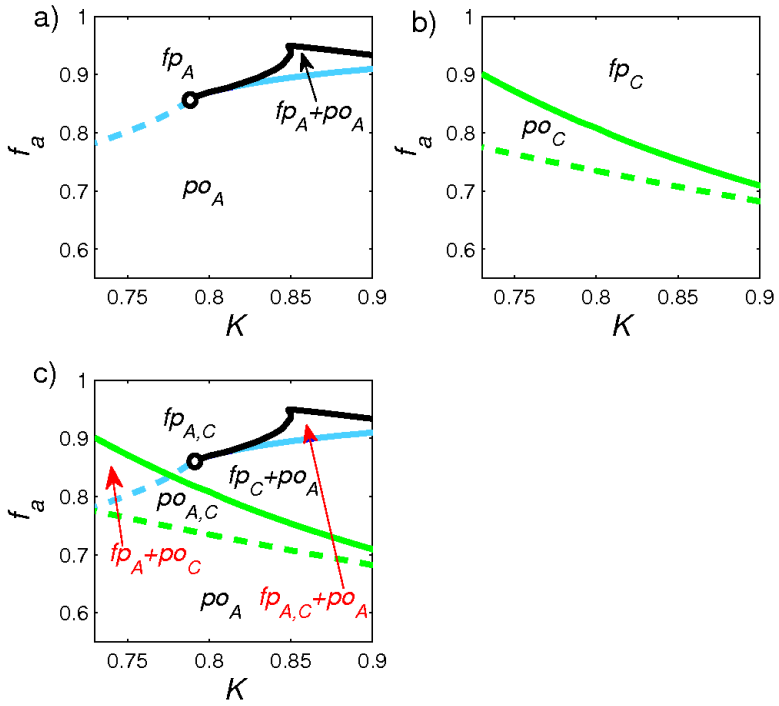

Figure 4.3: These figures show the solutions of the two-clump system described by Eq. (4.11) in the phase space of f_a versus K.

produced via a Hopf-bifurcation of fp_A, occurring as the dashed blue line is crossed from above. We found that po_A is stable in a region below the dashed blue curve and the black curve. In Fig. 4.3(b), there is another periodic attractor denoted po_C, which is produced by a Hopf-bifurcation of fp_C as the solid green curve is crossed from above. The orbit po_C is stable in the region between the solid green and dashed green curves. The regions of these stable solutions are plotted in Fig. 4.3(c) which essentially overlays Figs. 4.3(a) and 4.3(b) (the region where both fp_A and fp_C are stable is labeled $fp_{A,C}$.)

4.4 Transitions Between the Extensively Chaotic States and the Clumped States

In this section, we will consider the internal stability of the two clump state, and we will discuss the transitions between the clumped state and extensive chaos. In particular, we focus on transitions between the two-clump state and the extensively chaotic state with slow adiabatic change in K. To investigate this, we numerically solve the full system described by Eqs. (4.2) and (4.3), and also compute the internal stability exponents λ_{CI}^{σ} of the two clump states of Eqs. (4.7) and (4.11). We find that, if the full system is initially in a two-clump state and we decrease the coupling strength slowly, the population of clumps changes in such a way as to keep the clump state marginally stable with respect to the internal stability of the clumps. Eventually, with further decrease of K, the two clump state reaches a critical coupling strength at which adaptation to a marginally stable state is not

possible and the clumps explode, leading to a state of extensive chaos. In what follows, we will first discuss the internal stability analysis followed by the results of the numerical experiments of the full system. In addition, starting at a lower K value, in an extensively chaotic state, we will investigate how the state evolves with adiabatic *increase* of K and transitions to a two-clump state. We find that this transition is discontinuous and hysteretic, and that following this transition there is also a type of clump population readjustment occurring for increasing K, which is different from the marginal-internal-clump-stability-readjustment process for decreasing K.

4.4.1 Internal Clump Stability

The internal stability of a clump is determined by λ_{CI}^{σ} which is obtained from Eqs. (4.7)-(4.10). Clumps a and b are both stable if all λ_{CI}^{σ} are negative for $\sigma = a$ and b. We computed these exponents for all the two clump states in Fig. 4.3(c). The results are displayed in Fig. 4.4, which shows regions where the two clump state system solutions are stable ($\lambda_{SO} \leq 0$), and both clumps are internally stable (λ_{CI}^{a}, $\lambda_{CI}^{b} < 0$). Note that the blue curves in Fig. 4.4 represents the boundary above which there exists a stable fixed point solution fp_A of Eqs. (4.11) (same as Fig. 4.3(a)). Above the blue curves, fp_A orbit is stable according to Eqs. (4.11) (i.e., the values of λ_{SO} are negative). On the other hand, both clumps are internally stable only in the grey region bounded by the red solid and the blue curve. For the periodic orbit po_A, the clumps are not internally stable anywhere above the blue

curve, and po_A has internally stable clumps only in the green region below the blue curve, bounded by the red solid and dashed curves. In this figure, the red solid, red dashed, and the blue curves represent different ways that the clumps become internally unstable. The solid red curve corresponds to the boundary where one of the λ_{CI}^a is zero, which is where the larger clump a becomes unstable. The dashed red curve corresponds to the boundary where one of the λ_{CI}^b is zero, which is where the smaller clump b becomes unstable. Different from the red solid and dashed curves, all λ_{CI} for fp_A on the dashed blue curves are negative. As the orbit fp_A becomes unstable (i.e., one of the λ_{SO} becomes positive), the two-clump system Eqs. (4.11) goes to another attractor (either po_A or fp_C), for which, however, one of the clumps is internally unstable.

4.4.2 Marginal Stability and the Explosive Transition from the Clumped State to Extensive Chaos

Below we discuss the results of numerical experiments for the full system with $N = 1000$ oscillators. We first set $K = 0.9$ and run our numerical code long enough that the full system (Eqs. (4.2)) settled on a two-clump state (fp_A, see Fig. 4.4). We then track how this state varies as we decrease K adiabatically. The tracking method is similar to that described in Sec. 3 [where we searched for the stability boundary of the periodic orbit solutions in the two-clump system (Eqs. (4.11))]. In partitular, the initial condition of each successive simulations is the final oscillator states of the previous simulation together with small random noise ($\approx 10^{-7}$) added

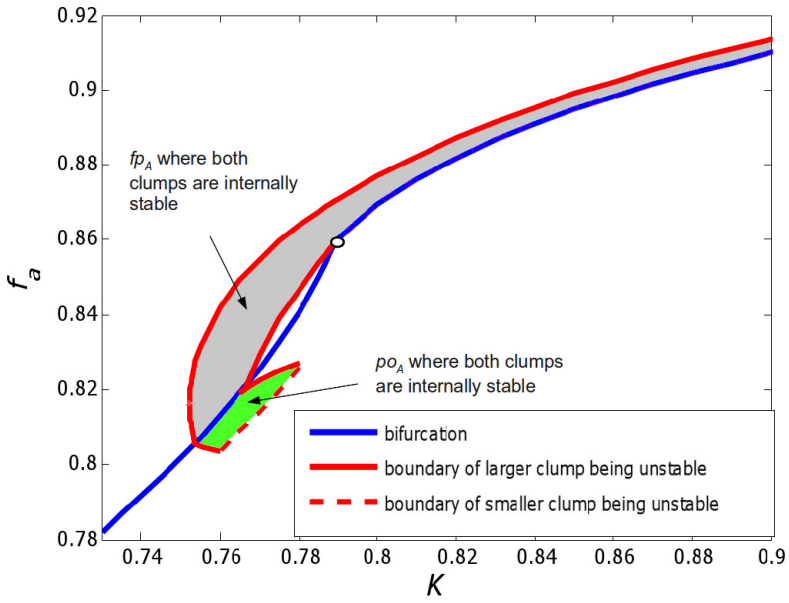

Figure 4.4: The blue curves represents the boundary where fp_A becomes unstable (same as Fig. 4.3(a)). fp_A is stable with internally stable clumps in the grey region. po_A is stable with internally stable clumps in the green region. The dashed red, and solid red are boundaries where the large clump a and the smaller clump b become unstable, respectively.

to each oscillator. Note, however, that in Eqs. (4.11) the clump populations f_a and $f_b = 1 - f_a$ are fixed, while, in contrast, in the full system Eqs. (4.2) we follow all the individual states W_j (for $j = 1, 2, ..., N$). Thus in Eqs. (4.2) the clump populations (f_a, f_b) may change dynamically upon change of the coupling $K \to K + \delta K$.

Results are shown in Fig. 4.5 in which the state of the full system is plotted in green in the $f_a - K$ space. Figure 4.5 also re-plots the results of Fig. 4.4, which shows the boundary where the orbit fp_A becomes unstable (the blue lines) and the boundary where the clumps become internally unstable (the dashed red lines). In the range $K = 0.9$ to $K \approx 0.75$, the full system is in the two-clump fixed point state fp_A. As we decrease K from 0.9, the population of clumps is redistributed in a way described by the green curve in Fig. 4.5a, which nearly matches the upper section of the red dashed curve. We refer to the mechanism of population redistribution between the clumps as 'marginal stability'. To understand this in more detail, imagine that the full system is in a two-clump state with $K = K^{'}$ and $f_a = f_a^{'}$ such that, in the $f_a - K$ space, it is located at a point within the grey internally stable region in Fig. 4.4. When $K^{'} \to K^{'} + \delta K$, $\delta K < 0$ ($\delta K \cong -10^{-5}$ in Fig. 4.5), the population of clumps of the full system remains unchanged if $K^{'} + \delta K$ and $f_a^{'}$ is still inside the grey region. This corresponds to the horizontal green lines shown Figs. 4.5b and 4.5c. On the other hand, if ($K^{'} + \delta K$, $f_a^{'}$) crosses the upper boundary of the grey region (i.e., the upper red dashed curve), clump a becomes internally unstable. We have made a movie of the time evolution of all the oscillators plotted in the complex W-plane when the green line crosses the red dashed curve. In the movie, we observe that the oscillators in clump a spread apart and interact with

clump b which in turn leads to oscillators in clump b spreading apart. The oscillators move in a complex manner until the system reassembles onto a new two-clumped state. Although the process of the redistribution of oscillators appears to involve complex chaotic dynamics, we observe that the net effect is that only one oscillator is transferred from clump a to clump b. In order to see expulsion of oscillators, we cannot allow all the oscillators in a clump to have the exact same states to machine round-off of our numerical computations. Thus, based on physical considerations and to prevent this from occurring, we have added the previously mentioned very tiny amount of random noise ($\approx 10^{-7}$) to the state of each oscillator. After the transfer of an oscillator from clump a to clump b, the new location in (K, f_a) space becomes $K = K^{'} + \delta K$ and $f_a = f_a^{'} - \frac{1}{N}$ which is now in the grey region. This $1/N$ decrease in f_a corresponds to the regular drop steps of the green line in the blow-ups shown in Figs. 4.5(b) and 4.5(a). Thus for $N \to \infty$ and $\delta K \to 0$, we expect that the drop steps of the green line tend to zero and that the green path followed by the system will converge to the dashed red curve. This process of redistribution of the clumps is repeated until $K \approx 0.75$, at the 'nose' of the red dashed line (the point at which df_a/dK becomes infinite). For K less than this critical value, clump a cannot restore its stability by the transfer of an oscillator to clump b. Consequently, we find that the two clumps solution explodes as K is reduced past the critical value $K_c \approx 0.75$, and an extensive chaotic attractor emerges. We defer discussion of the structure and properties of the extensively chaotic attractor to Sec. 4.5. Note that, in order to see the $(1/N)$ drop steps of the green lines, $|\delta K|$ should be small enough. A smaller $|\delta K|$ and longer numerical integration times between increments of K are

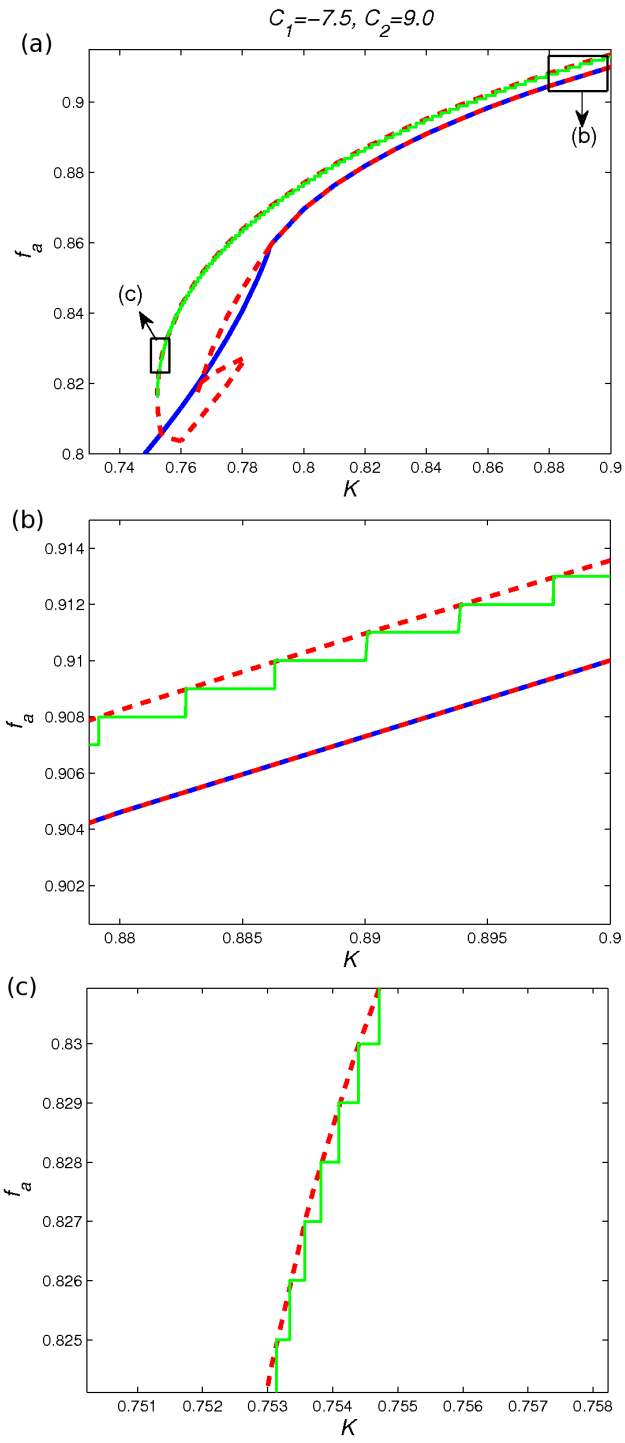

Figure 4.5: The full system is initially in a two-clump state at $K = 0.9$, and the system undergoes slow adiabatic change of K until $K \approx 0.7$. The trajectory of how the full system state changes is plotted in green. Also, the results in Fig. 4.4 is replotted similarly but replacing the the dashed red, solid red, and dashed magneta curves (Fig. 4.4) by the dashed red curves only in Fig. 4.5.

used in the 'nose' region near $K = K_c$ (Fig. 4.5(c)) due to the increase of the slope of the dashed red line as K decreases toward K_c.

4.4.3 The Discontinuous Transition from Extensive Chaos to Clumps with Increasing K

We now consider the evolution of the extensively chaotic attractor with slowly increasing K. We first choose $K = 0.7$ and run the numerical code long enough such that the system settles on an extensively chaotic attractor. We then track the attractor similarly as we did for the case of decreasing K. We typically find that the extensively chaotic attractor is destroyed at a coupling value K well in excess of the value K_c found for decreasing K (previous subsection). Furthermore, the K value where this occurs varies somewhat randomly when we repeat the computations under conditions that are very slightly different, and, on average, tends to be smaller for slower sweeping. Following destruction of the extensively chaotic state, a two-clump attractor emerges. Thus the situation is hysteretic since the transitions between the two-clump state and extensive chaos occurs at higher (lower) K when K is slowly increased (decreased).

In order to more clearly understand the nature of the transition from the extensively chaotic state to the two clump state with increasing K, we investigate the evolution of the system from random initial conditions where the $W_j(0)$ as uniformly sprinkled in the disc $|W| < 1$. For a specific value of $K > K_c$ in an appropriate range (e.g., $K = 0.86$), the system rapidly comes to a state where it

behaves chaotically, as in the infinite lifetime extensively chaotic state that exists, e.g., in $K < K_c$. However, after a (possibly quite long) finite time τ, the motion rather suddenly settles onto a two-clump state. Further, upon many repeats of this simulation procedure for the same K value, but with many different random initial conditions, we find that the time τ at which the system settles onto a two-clump state is different for different trials. Figure 4.7 shows semilog plots of histogram estimates of the probability distribution $P(\tau)$ of the lifetimes τ of the transient extensive chaos for several different K values in the range $0.845 \leq K \leq 0.870$. We observe that τ is approximately exponentially distribued for large τ; i.e., the semilog plots can be approximated fitted by a straight line at large τ. Performing such fits to the data in Fig. 4.7, we compute for each K a characteristic time $\langle\tau\rangle$ taken to be the inverse of the slope of the fitted lines. Figure 4.8 shows $\langle\tau\rangle$ versus K for $0.845 \leq K \leq 0.87$. We see that these characteristic settling times can be extremely long and increase monotonically with decreasing K. We do not currently have any principled basis for independently deducing the functional form of the dependence of $\langle\tau\rangle$ on K. However, we note that crises in the low dimensional chaotic systems [99, 100] can lead to chaotic transients with exponentially distributed lifetimes, roughly analogous to what we see in our system. Motivated by this observation, we try fitting our data for the dependence of $\langle\tau\rangle$ on K to a functional form that has been found to apply to typical crises in low dimensional systems [99,100], $\langle\tau\rangle \cong (\text{const.})(K - K^*)^{-\gamma}$, which we rewrite as

$$(1/\langle\tau\rangle)^{\gamma} \cong B(K - K^*), \tag{4.15}$$

where B is a constant, and K^* is a parameter value at which the lifetime of the chaotic transient diverges to infinity as $K \to K^*$ (from above), with the extensive chaos assumed to become perpetual (infinite τ) for $K < K^*$. In the low dimensional context γ is called the critical exponent of the crisis and has been theoretically analyzed in Ref [101]. Figure 4.9 shows $(1/\tau)$ versus K obtained from our data. This data seem to roughly conform to an approximately linear dependence (dashed line in Fig. 4.9) consistent with Eq. (4.15) and $\gamma \cong 1$. The dashed line intercept corresponds to a K^* value slightly less than 0.85 and substantially larger than $K_c \approx 0.74$.

4.4.4 Clump Population Redistribution with Increasing K

As discussed above, as K increases, there is a crisis-like transition of extensive chaos to a two-clump attractor. We now study the post-crisis evolution of this two clumps attractor with increasing K. Using approximately the same step size $|\delta K|$ as that in the case of decreasing K, for K between 0.84 and 0.9 (c.f., Fig. 4.5), we examine the evolution of the two clump attractor with $\delta K > 0$. To explain what we observe for increasing K, assume that the system is initially in a stable two-clump state at $K = K^{'}$ and $f_a = f_a^{'}$ and K is shifted to $K = K^{'} + \delta K$, where $\delta K > 0$. If the point, $K = K^{'} + \delta K$ and $f_a = f_a^{'}$, is to the right of the bottom boundary of the grey stability region shown in Fig. 4.4 (dashed red curve in Fig. 4.5 and Fig. 4.6), the *full* system state becomes unstable (i.e., a positive value of λ_{SO} emerges). From the analysis of Fig. 4.4, in the range of K above the transition out of the

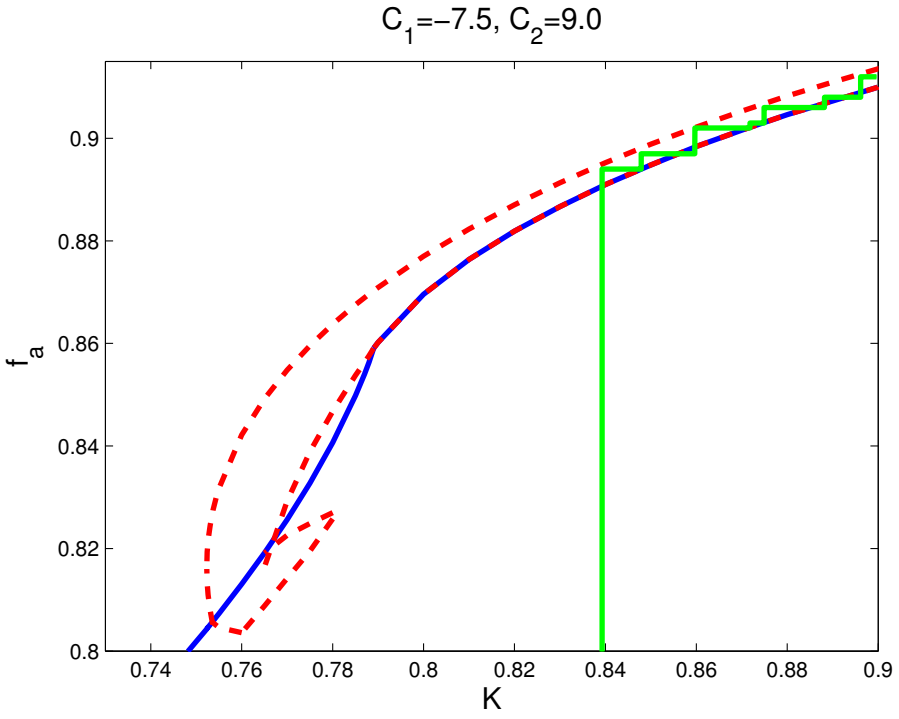

Figure 4.6: The full system is initially in the exntesive chaotic state at $K = 0.7$, and the system undergoes slow adiavatic change of K until $K \approx 0.9$. The trajectory of how the full system state changes is plotted in green. Also, the results in Fig. 4.4 is replotted similarly but replacing the the dashed red, solid red, and dashed magneta curves (Fig. 4.4) by the dashed red curves only in Fig. 4.6.

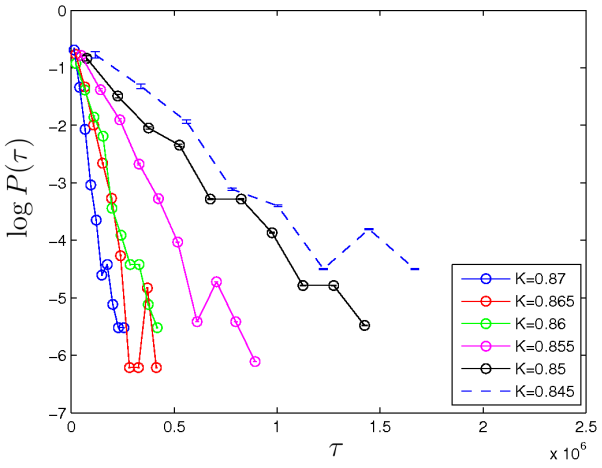

Figure 4.7: The natural logarithm of the probability distribution $P(\tau)$ versus the life time τ for different K, where fraction is the fraction of trials that have the corresponding life time.

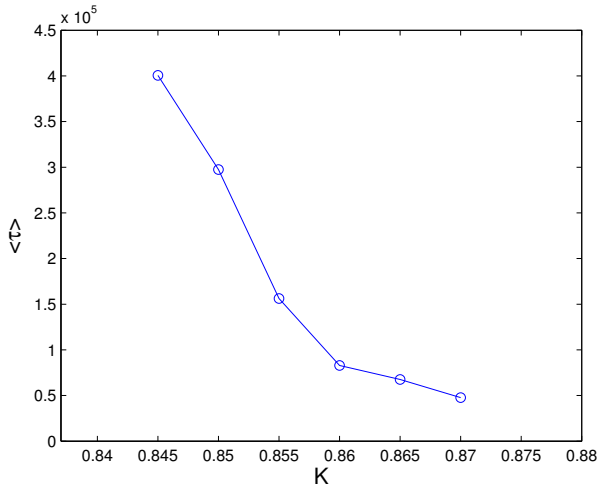

Figure 4.8: $\langle \tau \rangle$ is plotted versus K.

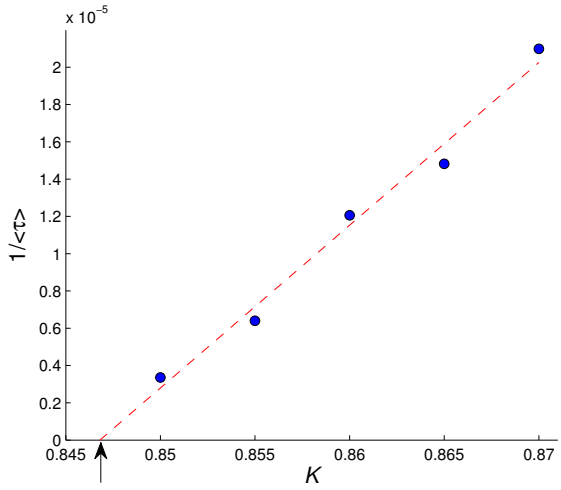

Figure 4.9: $1/\langle\tau\rangle$ versus K.

extensively chaotic state ($K > 0.84$), the only stable attractor from $K = 0.84$ to 0.9 is a two-clump fixed point state. As a result, we find that, as K is increased, the full system settles on one of these attractors, by making a number of successive irregular rises of f_a (the green curve in Fig. 4.6). We observe, however, that, unlike the evolution for decreasing K (Fig. 4.5), these rises are typically greater than $1/N$ and that their magnitudes are somewhat random.

4.5 Structure and Fractal Dimension of the Extensive Chaotic Attractors.

4.5.1 Snapshot Attractors

An attractor of a dynamical system with a fractal structure in its state space is called a *strange attractor*. In our case, the state space of our dynamical system of N oscillators is $2N$-dimensional, corresponding to specification at each time t of $Re(W_j)$ and $Im(W_j)$ for $j = 1, 2, ..., N$. Projecting this $2N$-dimensional state onto the two-dimensional complex W-plane by plotting the points $W = W_j(t)$ for $j = 1, 2, ..., N$, at a specific time t, we obtain a 'snapshot' of this projection. Our numerical experiments for extensively chaotic cases with large N show that the points in these snapshot projections appear to form a fractal distribution.

Let $\hat{D}$ denote the fractal dimension of such a time t projected pattern for an orbit that is *on the attractor* (here we will use the well-known information dimension as our definition of dimension [102]). The fractal attractor projection at any subsequent time $t + T$ is related to the time t attractor projection by a *smooth* mapping

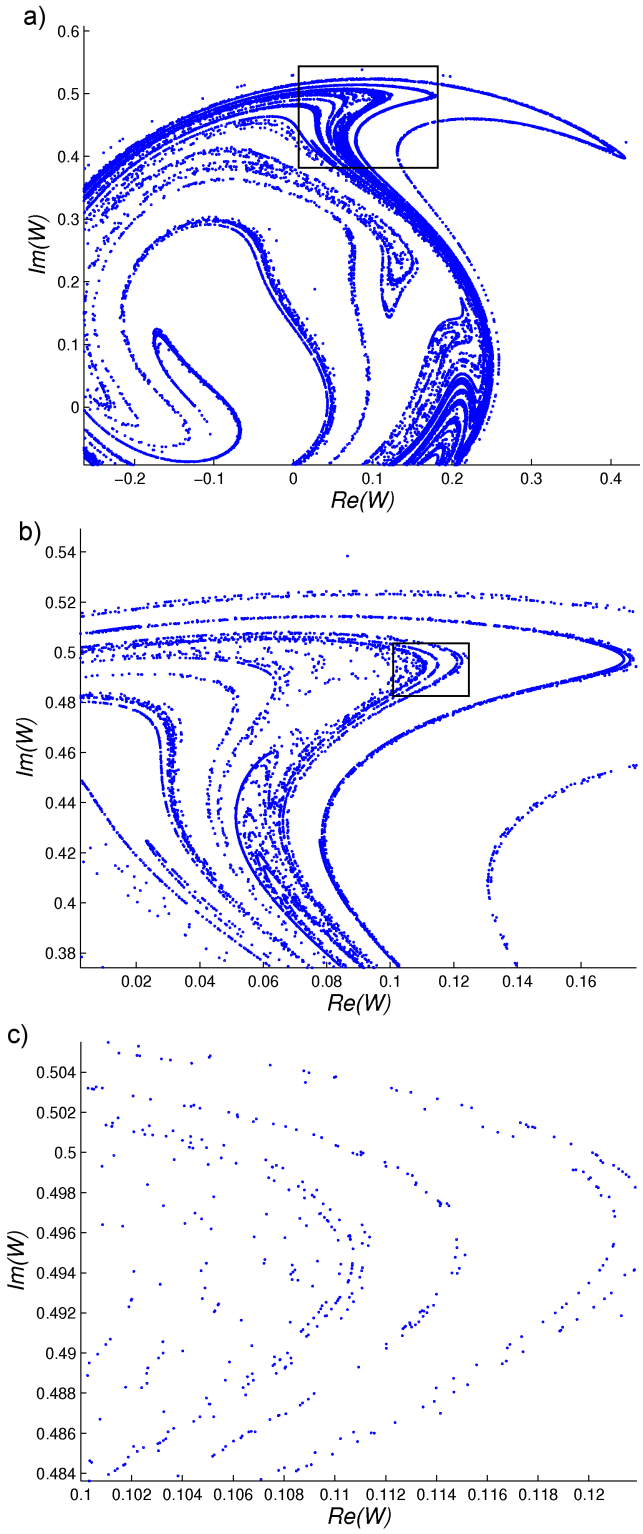

Figure 4.10: (a) A snapshot of the attractor is plotted with $K = 0.8$ and $N = 50000$. (b) The blow-up of the rectangles in (a). (c) The blow-up of the rectangle in (b).

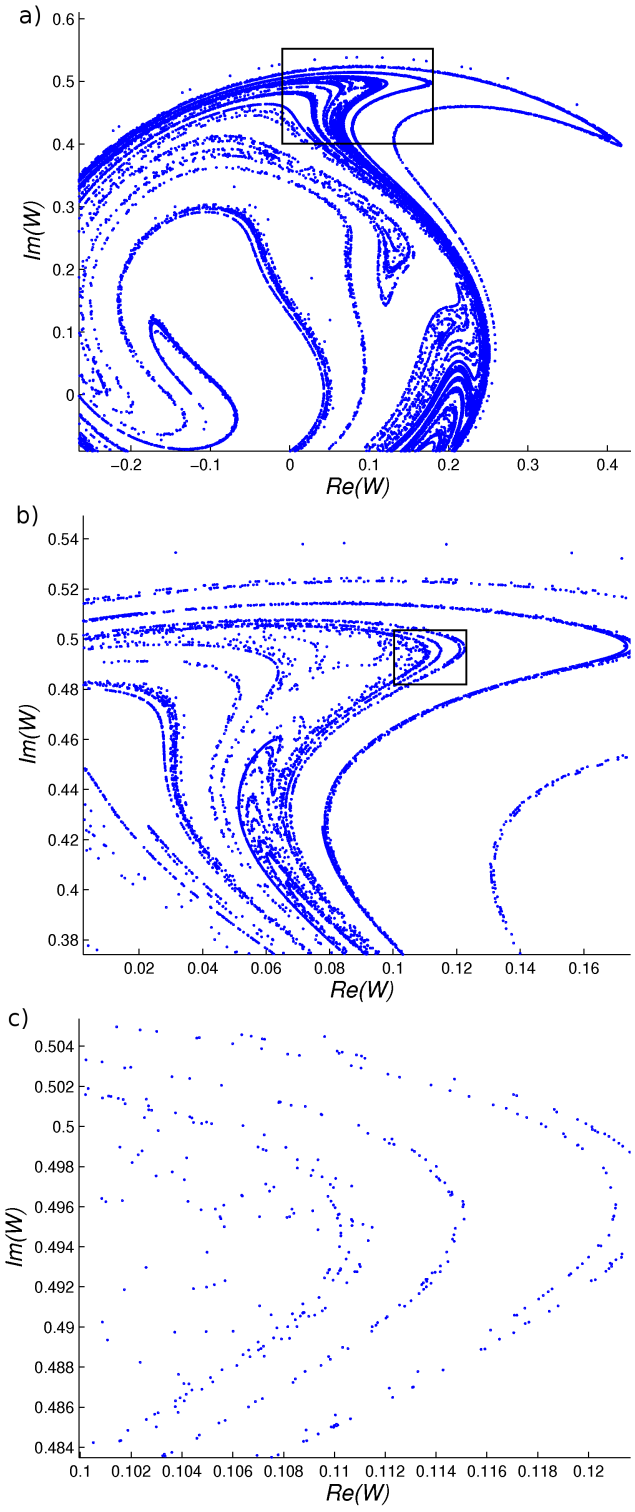

Figure 4.11: Snapshot attractor with externally imposed $\bar{W}(t)$. (a) A snapshot of the attractor is plotted similar to Fig. 4.10(a). (b) The blow-up of the rectangles in (a). (c) The blow-up of the rectangle in (b).

of the $W-$plane that follows from the $2N$-dimensional flow specified by Eqs. (4.2) and (4.3). Thus the dimension of the fractal attractor's projection must be the same at time t and $t+T$. That is, $\hat{D}$ is constant with time. Furthermore, we will argue later in this section that the attractor dimension D_A in the full $2N-$dimensional state space satisfies

$$D_A = N\hat{D} \tag{4.16}$$

as $N \to \infty$. Thus we confirm that such an attractor is indeed extensive.

An example of the observation of the fractal structure of the extensively chaotic attractors is given in Fig. 4.10 which shows the state of each of the $N = 50000$ oscillators in the complex $W-$plane for $K = 0.8$. Figure 4.10(a) shows a part of the snapshot attractor. Figure 4.10(b) displays a blow-up of the rectangle in Fig. 4.10(a), which, like Fig. 4.10(a), reveals that there exists fine-scale structure appearing as a number of curved "lines". Figure 4.10(c) displays a blow-up of the rectangle in Fig. 4.10(b), which (to within the resolution due to finite N) shows structure qualitatively similar to that in Figs. 4.10(a) and 4.10(b). We believe that, for $N \to \infty$, continuation of this blow-up procedure would show that the snapshot attractor has similar structure on arbitrarily small scale. Also, as previously mentioned, we have made a movie of the time evolving fractal-like pattern formed by the $N = 1000$ oscillators as they move in the complex W-plane. This movie shows continual stretching and folding dynamics, thus illustrating the mechanism by which the chaotic dynamics is produced.

As we will soon show, a useful way of thinking about snapshots like that

in Fig. 4.10 is to regard the time dependence of $\bar{W}(t)$ as being like an externally imposed random forcing in the equation for each oscillator j (Eq. (4.2)). That is, we ignore the self-consistent nature of $\bar{W}(t)$ which in reality is the average over all the $W_j(t)$. This view can be motivated as follows. Consider a specific oscillator $j = l$, and delete this one oscillator from the mean field to form

$$\bar{W}' = \frac{1}{N}\sum_{j\neq l} W_j = \bar{W} - (W_l/N). \tag{4.17}$$

Now consider the dynamics of the original system (4.2), but with $\bar{W}$ replaced by $\bar{W}'$. With this replacement the dynamics of the $(N-1)$ oscillators $j \neq l$ is uncoupled from the dynamics of oscillator l, and $\bar{W}'$ appearing in the equation for oscillator $j = l$ in Eq. (4.2) is thus effectively an imposed external forcing. Furthermore, for large N, the difference $\bar{W} - \bar{W}' = W_l/N$ is small and approaches zero as $N \to \infty$. Thus we expect that, for appropriate considerations of the oscillator dynamics in the case $N \gg 1$, the behavior of an individual oscillator of the system can be regarded as being like that of an isolated oscillator driven by an external $\bar{W}(t)$. In order to validate the view of $\bar{W}(t)$ as acting like an externally imposed forcing, we save in computer memory the time series of $\bar{W}(t)$ that resulted in the snapshot of Fig. 4.10. Next we choose $N = 50000$ random initial conditions $W_j(0)$ that are different from the 50000 random initial conditions used in generating Fig. 4.10. We then use Eqs. (4.2) to evolve these new initial conditions, but, in doing this, we replace the self-consistent $\bar{W}(t)$ by the previously computed and saved $\bar{W}(t)$. Thus, in this new computation, $\bar{W}(t)$ really is externally imposed. Figure 4.11 shows the resulting snapshot for this case determined at the same time t as in Fig. 4.10. We observe

that the macroscopic fractal-like patterns in Figs. 4.10 and 4.11 are the same. The only difference is that the exact placements of individual points are not the same. Our interpretation is that, associated with the saved time-dependent $\bar{W}$, there is an underlying time-dependent multifractal measure and that Figs. 4.10 and 4.11 represent two independent random $N = 50000$ samplings from this measure.

Next we consider the $N-$dependence of the dynamics. We have seen that the snapshot can be regarded as resulting from an external forcing $\bar{W}(t)$ and that this determines the overall snapshot pattern. We, therefore, examine the statistical properties of $\bar{W}(t)$. Figures 4.12(a) and 4.12(b) show $|\bar{W}(t)|$ for extensively chaotic dynamics ($K = 0.8$) for $N = 10000$ and $N = 50000$. These look qualitatively similar, but, to make the comparison qualitative, we show in Fig. 4.13 the correlation function,

$$C(\tau) = \langle [|\bar{W}(t)| - \langle |\bar{W}(t)| \rangle][|\bar{W}(t+\tau)| - \langle |\bar{W}(t)| \rangle] \rangle, \tag{4.18}$$

where the angle brackets denote a time average. In Fig. 4.13 the results for $N = 10000$ and $N = 50000$ are plotted as solid dots and crosses, respectively. The good agreement between these two results indicates that, at these large N values, the statistical properties of $\bar{W}(t)$ have essentially attained their $N \to \infty$ limiting form. As a consequence, we also conclude that the measure corresponding to the distributions in Figs. 4.10 and 4.11 has also essentially attained its $N \to \infty$ form.

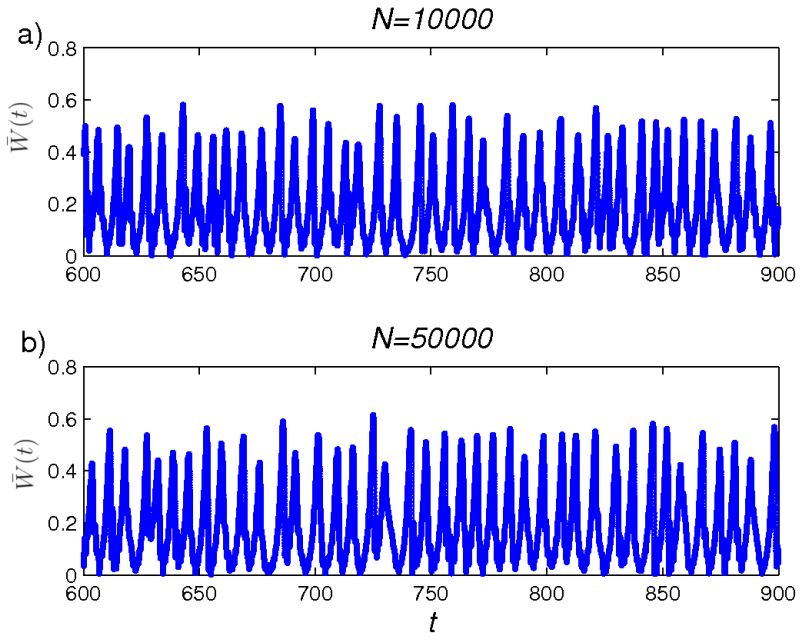

Figure 4.12: $\bar{W}(t)$ versus t for (a) $N = 10000$, and (b) $N = 50000$.

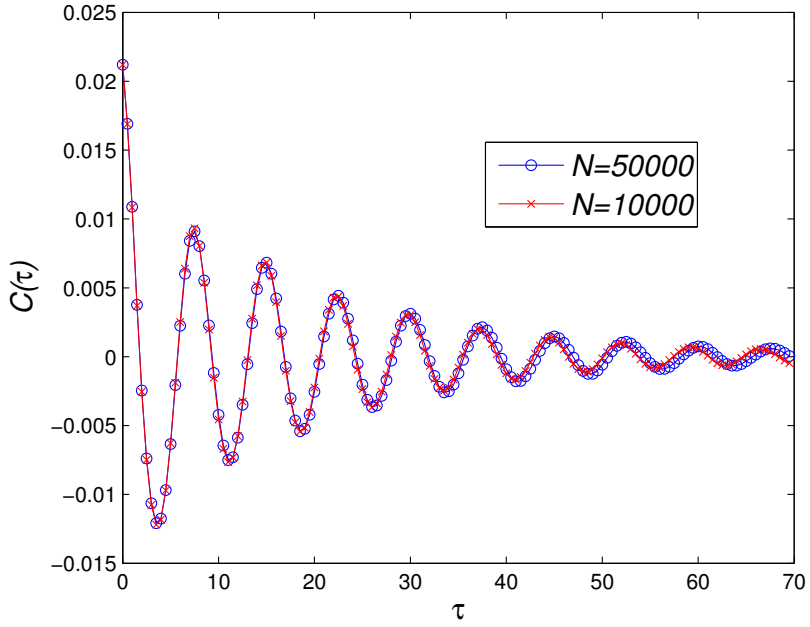

Figure 4.13: Correlation function $C(\tau)$ (Eq. 4.18) versus τ for N=10000 (red crosses) and N=50000 (blue circles).

4.5.2 Fractal Dimension

The usual definition of the information dimension D_I of an attractor in an M-dimensional space is [102]

$$D_I = \lim_{\epsilon \to 0} \frac{\sum_{i=1}^{\widetilde{N}(\epsilon)} \mu_i \ln \mu_i}{\ln \epsilon}, \tag{4.19}$$

where it is supposed that the space has been divided by a rectanglular grid into equal size M-dimensional cubes of edge length ϵ, and μ_i is the frequency with which a typical orbit on the attractor visits the ith cube. The information dimension may be thought of as quantifying how the average information content $I(\epsilon) = \sum \mu_i \ln \mu_i^{-1}$, of a measurement of the system state scales with the resolution, ϵ, of the measurement, $I(\epsilon) \sim \epsilon^{-D_1}$.

We now wish to numerically estimate the information dimension $\hat{D}$ for the measure corresponding to the distributions in our snapshots. In order to accomplish this, rather than numerically implementing a procedure based directly on division of the W-plane into an ϵ grid, as in the definition of Eq. (4.19), we find it convenient to use an alternate procedure that does not require formation of an ϵ-grid. Our procedure is a variant of the idea of Grassberger and Procaccia [98] for computing the correlation dimension, but adapted to yield the information dimension. We proceed as follows. We consider a snapshot attractor plot of N points in the complex W-plane at time t. We denote by $B_{\epsilon,j}^t$ the disc of radius ϵ centered at the point W_j, and by $\mu(B_{\epsilon,j}^t)$ the fraction of oscillator state points in the snapshot that fall within

the disc $B^t_{\epsilon,j}$. We let

$$Z_\epsilon = \langle \frac{1}{N} \sum_{j=1}^{N} \ln \mu(B^t_{\epsilon,j}) \rangle, \tag{4.20}$$

where $\langle ... \rangle$ again represents an average over time t (i.e., over snapshots).

Next we plot Z_ϵ versus $\ln \epsilon$. Assuming the existence of a reasonably large linear scaling range dependence for ϵ small compared to the diameter of the snapshot pattern, yet large compared to the average minimum distance between points, we estimate the snapshot pattern's information dimension ($\hat{D}$) as the slope of a straight line fit to this dependence in the appropriate range. We have numerically computed Z_ϵ versus $\ln \epsilon$ for the extensively chaotic attractors at $K = 0.7$ and $K = 0.8$ (Fig. 4.14), using 300 snapshots (corresponding to 300 times) each. The results are shown as the blue curves in Fig. 4.15. We see that there is indeed a reasonable scaling range of linear dependence. The red straight lines are obtained from a theory that we discuss in section 4.5.3. The red line plots have slopes corresponding to values of the information dimension of $\hat{D} = 1.26$ (Fig. 4.15(a) for $K = 0.7$) and $\hat{D} = 1.30$ (Fig. 4.15(b) for $K = 0.8$). As is evident from Fig. 4.15 these theoretical slope values are consistent with the blue curve plots in the scaling range.

We have also repeated this calculation for patterns obtained as in Fig. 4.11 (i.e., with $\bar{W}(t)$ replaced by an externally imposed time series obtained and saved from a previous self-consistent computation with different random initial condition). The results (not shown) are virtually identical to those obtained for our self-consistent calculations shown in Fig. 4.15, thus providing quantitative support for the view that $\bar{W}$ acts as an externally imposed driver for large N.

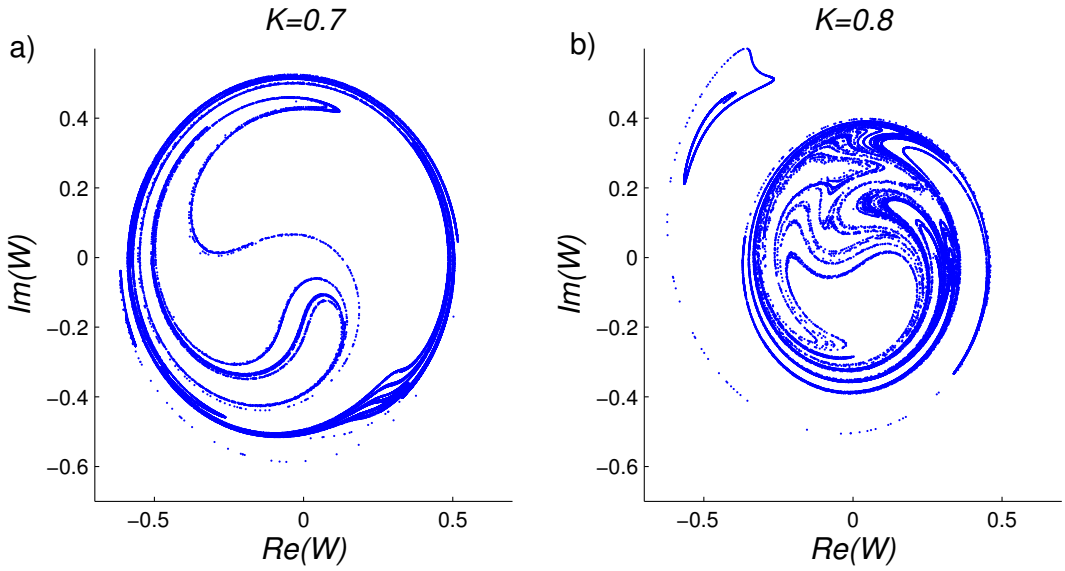

Figure 4.14: Snapshot attractors for N= 50000. (a) K=0.7 and (b) $K = 0.8$.

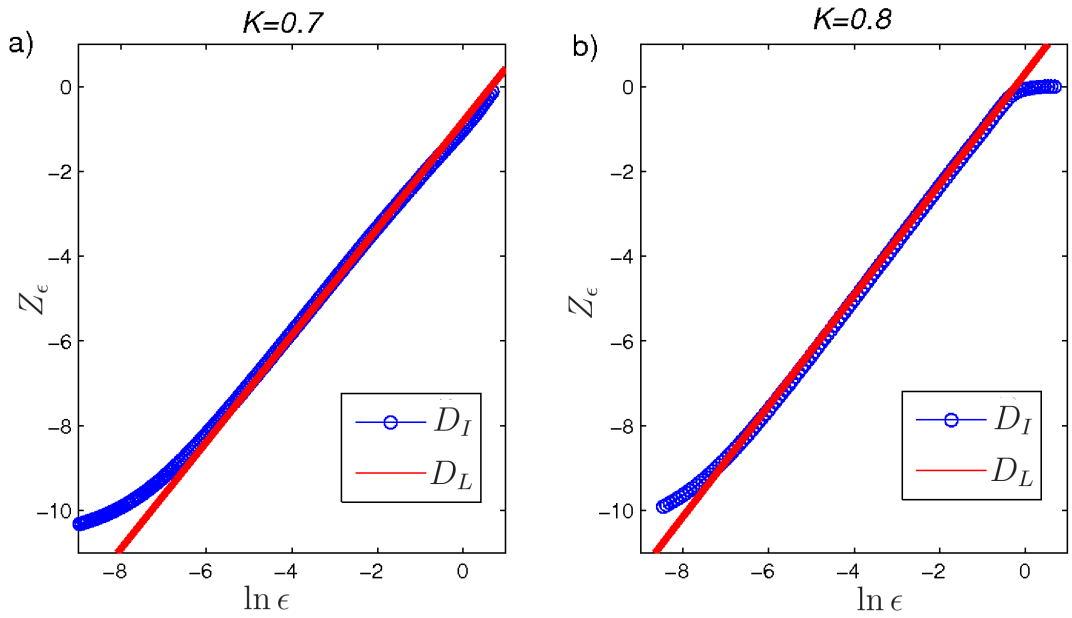

Figure 4.15: $\ln Z_\epsilon$ versus ϵ for (a) $K = 0.7$ and (b) $K = 0.8$. The information dimension D_I can be estimated by the slopes of fitted straight lines for the linear region of blue curves. The Lyapunov dimension is shown by the slop of the red curves.

4.5.3 Lyapunov Dimension

Since $\bar{W}(t)$ can be regarded as an externally imposed forcing and varies chaotically, we can view each of the N equations (Eqs. (4.2)) as being identical random dynamical systems of the general type considered by Yu *et al.* [95, 96]. For such systems Yu *et al.* show that the Kaplan-Yorke conjecture applies to snapshots (see also Young and Ledrappier [103]).

The Kaplan-Yorke conjecture [104–106] relates the information dimension of an attractor to the Lyapunov exponents. Consider an M dimensional system with Lyapunov exponents $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_M$. Let Q be the largest integer such that

$$\sum_{q=1}^{Q} \lambda_q \geq 0. \tag{4.21}$$

The Lyapunov dimension is defined by

$$D_L = Q + \frac{1}{|\lambda_{Q+1}|} \sum_{q=1}^{Q} \lambda_q. \tag{4.22}$$

(Note that the second term in (4.21) is between 0 and 1.) The Kaplan-Yorke conjecture is that $D_I = D_L$ for 'typical attractors'.

To apply the Kaplan-Yorke conjecture to a randomly forced system like (2) with $\bar{W}$ regarded as externally imposed, we calculate the two Lyapunov exponents for the variations δW_j with $\delta\bar{W} \equiv 0$ (because, for large N, $\bar{W}$(t) is regarded as externally imposed). Two Lyapunov exponents $\lambda_1 > 0 > \lambda_2$ are obtained, where for the cases we consider $\lambda_1 + \lambda_2 < 0$. Thus from (21)

$$D_L = 1 + \lambda_1/|\lambda_2|. \tag{4.23}$$

Note that for sufficiently long calculation times essentially the same numerical values of the exponents are found for all $j = 1, 2, ..., N$. The red lines on Fig. 4.15 have the slope given by (4.22).

4.5.4 Extensivity

We now return to the issue of how the information dimension of the snapshot attractors $\hat{D}$ is related to the attractor dimension D_A of the full system in its $2N$ dimensional state space. In particular, we give an argument supporting Eq. (4.16).

Going back to the definition of the information dimension in terms of the ϵ scaling of the information associated with an ϵ-accuracy state measurement, we note that a state measurement of all the W_j at any time t would determine the points W_j in a snapshot. Also most of the volume of a $2N$-dimensional ϵ-edge cube in the full $2N$-dimensional state space projects to an area of the W-plane with a diameter $\sim \epsilon$. Thus ϵ-accuracy measurements of the positions $\{W_j\}$ in the W-plane are approximately equivalent to an ϵ-accuracy measurement of the a state in the full state space. There are N positions of the W_j in the complex W-plane that must be measured to determine the full state. Further, if $\bar{W}(t)$ is regarded as imposed for large N, these W_j can be regarded as uncoupled (c.f. Eq. (4.2)) when considering the snapshot pattern and its dimension. Thus for large N the information associated with an ϵ-accuracy measurement of the full state is N times the information of an ϵ-accuracy measurement of one of the W_j. Hence, Eq. (4.16) follows, and we conclude that, for our system Eqs. (4.2) and (4.3), or indeed for any system of the mean-field

type Eq. (4.1) with $N \gg 1$, observation a fractal pattern in a snapshot corresponds to extensive chaos.

Another way of understanding Eq. (4.16) is as follows. Our system Eqs. (4.2) and (4.3) has $2N$ Lyapunov exponents. For very large N each oscillator equation can be approximately regarded as driven by an externally imposed $\bar{W}(t)$, and has Lyapunov exponents $\lambda_+ > 0 > \lambda_-$. Thus, at finite large N, we expect that a histogram of the $2N$ Lyapunov exponent values will be sharply peaked at $\lambda = \lambda_+$ and $\lambda = \lambda_-$, approaching delta functions at these two values in the limit as $N \to \infty$. This expectation is consistent with our extensivity result $D_A \cong N\hat{D}$, for $N \gg 1$. This follows from the Kaplan-Yorke formula for the Lyapunov dimension. In particular, considering that N of the exponents are approximately λ_+ and N are approximately λ_-, we have that $|\lambda_-| \geq N\lambda_+ - (Q-N)|\lambda_-| \geq 0$. Thus, $D_A \cong Q \cong N(1+\lambda_+/|\lambda_-|) = N\hat{D}$, consistent with out previous argument.

4.6 Conclusions

In this paper we have considered the dynamics of large systems of many identical Landau-Stuart oscillators coupled by their mean field (Eqs. (4.2) and (4.3)). We have obtained results that we believe should also apply to other types of mean-field coupled systems of many identical dynamical units (e.g., Eq. (4.1)). Our results were of two types. One type of result concerned dynamical transitions: the question of how an identified attractor evolves and bifurcates with slow adiabatic change of a system parameter (Secs. 4.3 and 4.4). The other type of result concerned

the structure of what we have called extensive chaos in these types of system (Sec. 4.5). We now summarize our main conclusions in these two areas.

Our conclusions with regard to dynamical transitions are the following.

1. Adiabatic variation of a parameter in a clumped state regime can lead to redistribution of oscillators between the clumps, and two mechanisms inducing such redistribution are marginal stability of clump integrity (as for the case of decreasing K in the range $K \geq 0.75$ in Fig. 4.5) and the crossing of the existence boundary for stable solutions of the clump motion equations, Eqs. (4.11) (as for the case of increasing K in the range $K \geq 0.85$ in Fig. 4.6).

2. An apparently typical explosive type of dynamical transition from a clumped state to an extensively chaotic state has been found to occur at a critical coupling value past which maintainence of clump internal stability becomes impossible ($K \leq 0.75$ in Fig. 4.5).

3. A transition by which an extensively chaotic attractor can be destroyed has been identified as bearing similarity to the crisis transition mechanism [99,100] of chaotic attractors of low dimensional systems; specifically, with variation of a system parameter, it appears that the extensively chaotic motion can assume a transient character whereby the extensive chaos exists only for a finite time, before, rather abruptly, moving to another type of motion (Fig. 4.8.)

Our conclusions with regard to the structure of extensive chaos in mean-field coupled systems of many identical dynamical units (large N) are the following.

1. These systems behave essentially like a collection of uncoupled components with a common random-like external drive.

2. Snap-shots of the component states of the system projected onto the state space of an individual unit (e.g., for Eqs. (2) and (3), the complex W-plane) can display fractal structure (e.g., Fig. 4.10) whose information dimension can be predicted by use of the Kaplan-Yorke formula.

3. The attractor dimension in the full phase space is N times the fractal dimension of a snapshot projection.

A.1 Dynamical Model of Bivalent Histone Modification

6-state model. Refer to Fig. 16. Circles in the figure represent nucleosomes. A nucleosome contains two histone copies represented by the vertically oriented ellipses. Each histone has a site (represented by the upper half of the ellipse) that can be either unmodified (symbolized by u) or have an active mark (symbolized by α) and another site (represented by the lower half of the ellipse) that can be either unmodified (symbolized by u) or have a repressive mark (symbolized by ρ). Each of the four modification sites in a nucleosome can be in one of two states (modified or unmodified), yielding the $2^4 = 16$ possibilities that are shown in the figure panels, (a)-(p). Panels grouped together by the curly brackets in the figure represent the same physical nucleosome state, e.g., panels (e) and (f) are considered to represent the same physical nucleosome state since (f) results from (e) by interchange of the left and right histone ellipses. There are six such pairs. Thus there are 10 physically distinct nucleosome states. In addition, experiments indicate that active and repressive marks do not occur simultaneously on the same histone [58] (i.e., α and ρ do not occur in the same ellipse). This eliminates the possibilities depicted in panels (d) and (k)-(p). Thus we arrive at 6 possible states. In these 6 possible states, each histone has three distinct configurations, and we assign symbols A, U, R to them. They have the following meanings.

A: The histone has an active mark and the other site is unmodified.

U: All two sites are unmodified.

R: The histone has a repressive mark and the other site is unmodified.

As a result, the six possible states can be depicted by 2 letters instead of 4 letters, which we label UU, AA, RR, AU, UR, and AR as shown in Fig. 16.

Reduced model. We now introduce a reduction of the above 6-state model to a more simple model. Our reduction is motivated by a limited number of simulations of the 6-state model in which we found that the experimentally observed bivalent nucleosome state (AR) tended to be absent unless the AA and/or RR states were suppressed (i.e., $\pi_{\sigma\sigma'}$ is low for the transition $\sigma = AU \rightarrow \sigma^{'} = AA$ and the transition $\sigma = UR \rightarrow \sigma^{'} = RR$). This is consistent with a recent experimentally motivated hypothesis that the existence of the asymmetrically modified nucleosome states, AU and UR, are important for the formation of bivalent domains [58].

One way of understanding this is to note from Fig. 3.2 that the AA state competes with the AR state for conversion from the AU state, and the RR state similarly competes with the AR state for conversion from the UR state. This suggests that if we want to allow for the occurrence of the experimentally observed AR state, we could chose parameters in our six state model such that the transition rate from AU to AA is sufficiently smaller than the transition rate to AR. Similarly we would want the transition rate from UR to RR to be sufficiently smaller than the transition rate to AR. Thus, to make the model more tractable, we employ a further simplification and consider the idealized case in which AA and RR are completely

suppressed. That is, in terms of our 6-state model, we set $\pi_{\sigma\sigma'} = 0$ for the transition $\sigma = AU \rightarrow \sigma' = AA$ and the transition $\sigma = UR \rightarrow \sigma' = RR$. In this formulation, AA and RR states do not occur, and the 6-state model reduces to a 4-state model.

Another example of localization of AR states related to our results in Section 4.4.3

We note that our result in Fig. 3.9 is not consistent with experiment in that in Fig. 3.9 the active marks are more extensive than the repressive marks, while Ref. [34] shows that the opposite situation holds in experiment. We note, however, that, as shown in Fig. 17, for other reasonable parameter choices, we can also obtain states for which the repressive marks are more extensive than the active marks (consistent with [34]).

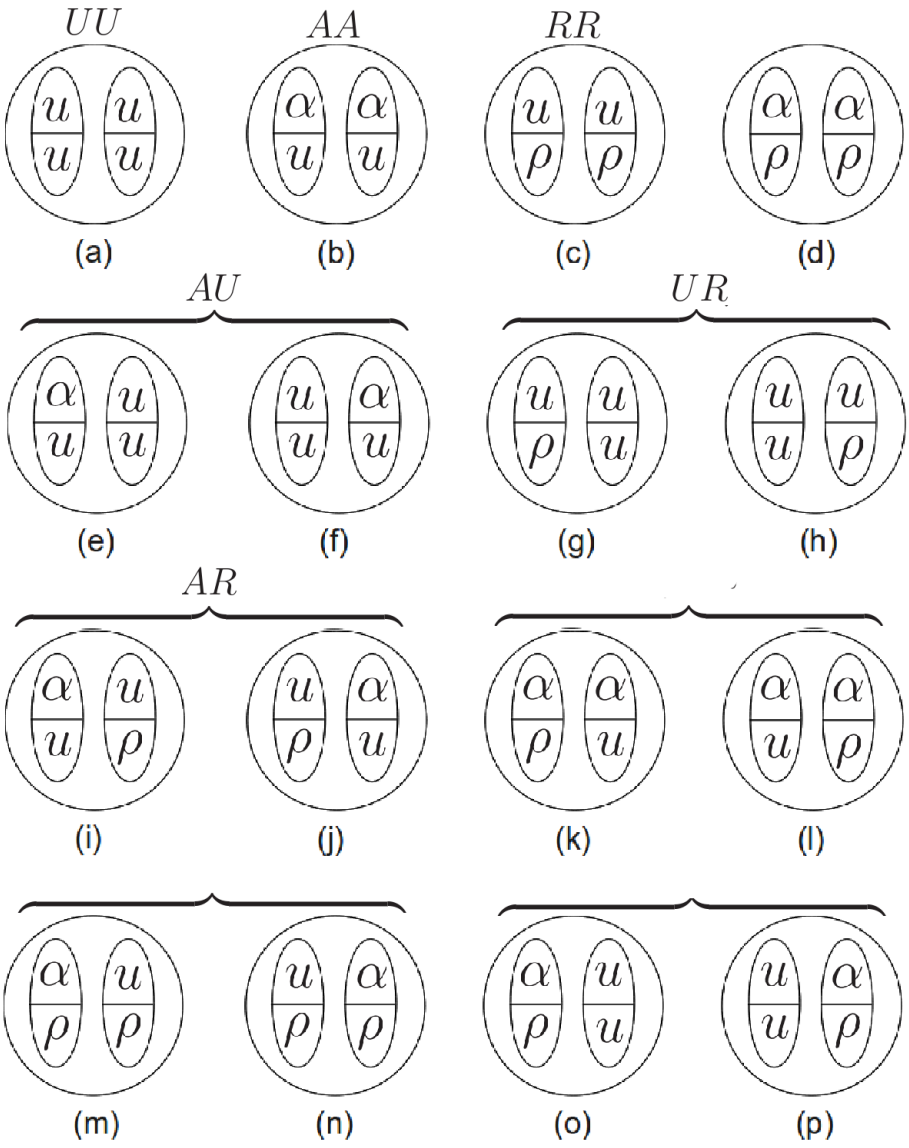

Figure 16: **Illustration for the explanation of the states of the 6-state model.**

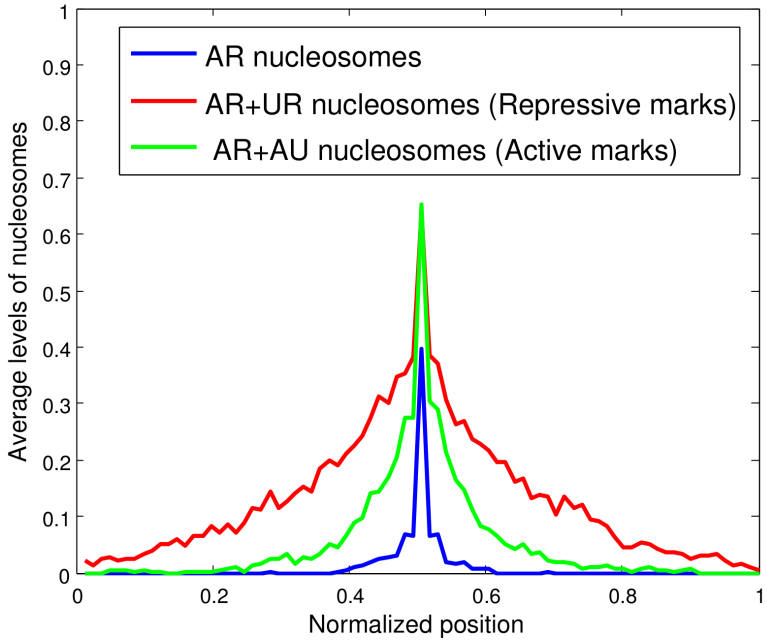

Figure 17: **Figure S2. An example of the distribution of AR nucleosomes, active, and repressive marks.** This plot illustrates that the 4-state model described in the main text can simulate bivalent domains (blue) in which the active mark (green) is less extensive than the repressive mark (red) (i.e., the bivalent domains (blue) are buried in the repressive domains (red)). The details of simulation can be referred to Section 3.3.3. Here, distributions of AR nucleosomes (blue), $AR + UR$ nucleosomes (red), and $AR + AU$ nucleosomes (green) are plotted at the end of the simulation runs (time =1800). The average levels of nucleosomes are averaged over 1000 simulation runs. In the simulation, $p_{UA}^{i=40} = 0.03$ and $p_{UR}^{i=40} = 0.015$. The other parameters are $r_{UA} = 0.029$, $r_{UR} = 0.021$, $p_{AU} = 0.025$, $p_{RU} = 0.015$, $r_{AU} = 0.004$ and $r_{RU} = 0.002$.

Bibliography

[1] W. L. Ku, G. Duggal, Y. Li, M. Girvan, and E. Ott. Interpreting patterns of gene expression: Signatures of coregulation, the data processing inequality, and triplet motifs. *PLoS ONE*, 7(2):e31969, 02 2012.

[2] W. L. Ku, M Girvan, G.-C. Yuan, F. Sorrentino, and E. Ott. Modeling the dynamics of bivalent histone modifications. *PLoS ONE*, 8(11):e77944, 11 2013.

[3] U. Alon. *An introduction to systems biology: design principles of biological circuits*. CRC Press, 2007.

[4] J. J. *et al.* Faith. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, January 2007.

[5] A. *et al.* Margolin. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1:S7, January 2006.

[6] F. Mordelet and J. Vert. SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76–i82, 2008.

[7] D. F. T. Veiga, F. F. R. Vicente, M. Grivet, A. de la Fuente, and A. T. R. Vasconcelos. Genome-wide partial correlation analysis of Escherichia coli microarray data. *Genetics and molecular research : GMR*, 6(4):730–42, January 2007.

[8] P. Zoppoli, S. Morganella, and M. Ceccarelli. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC bioinformatics*, 11:154, January 2010.

[9] O. Elemento, N. Slonim, and S. Tavazoie. A universal framework for regulatory element discovery across all genomes and data types. *Molecular Cell*, 28(2):337 – 350, 2007.

[10] D.J. Allocco, I.S. Kohane, and A.J. Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, 5(1):18, 2004.

[11] P . E. Meyer, F. Lafitte, and G. Bontempi. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, 9:461, 2008.

[12] K. C. Liang and X. D. Wang. Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008.

[13] D. C. Kim, X. Y. Wang, C. R. Yang, and J. Gao. Learning biological network using mutual information and conditional independence. *BMC Bioinformatics*, 11(Suppl 3):S9, 2010.

[14] J. Watkinson, K.-C. Liang, X. Wang, T. Zheng, and D. Anastassiou. Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann. N. Y. Acad. Sci.*, 1158:302–313, March 2009.

[15] M. S. *et al.* Carro. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, 01 2010.

[16] A. Margolin and A Califano. Theory and limitations of genetic network inference from microarray data. *Ann. N. Y. Acad. Sci.*, 1115:51–72, 2007.

[17] C. Olsen, P. E. Meyer, and G. Bontempi. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009, 2009.

[18] J. J. *et al.* Faith. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucl. Acids Res.*, 36(Database issue):D866–70, January 2008.

[19] S. *et al.* Gama-Castro. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucl. Acids Res.;*, 2010.

[20] T. I. *et al.* Lee. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, October 2002.

[21] C.T. *et al.* Harbison. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[22] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–7, October 2002.

[23] M. C. *et al.* Costanzo. YPDTM, PombePDTM and WormPDTM: model organism volumes of the BioKnowledgeTM Library, an integrated resource for protein information. *Nucl. Acids Res.*, 29(1):75–79, 2001.

[24] C. Daub, Ralf Steuer, J. Selbig, and S. Kloska. Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5(1):118, 2004.

[25] L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15:1191–1253, June 2003.

[26] N. Slonim, G. S. Atwal, G. Tkacik, and W. Bialek. Estimating mutual information and multi-information in large networks. *arXiv*, page cs/0502017, February 2005.

[27] L. D. Fisher and G. V. Belle. *Biostatistics*. New York: John Wiley & Sons, 1993.

[28] S. Balaji, L. M. Iyer, M. M. Babu, and L. Aravind. Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal? *Trends in Genetics*, 24(7):319 – 323, 2008.

[29] Tony K. Chromatin modifications and their function. *Cell*, 128(4):693 – 705, 2007.

[30] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, 2001.

[31] T. S. *et al.* Mikkelsen. *Nature*, 448(7153):553–560, 2007.

[32] D. Moazed. Mechanisms for the inheritance of chromatin states. *Cell*, 146(4):510–518, 2011.

[33] T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693 – 705, 2007.

[34] Bradley E. B. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125:315, 2006.

[35] X. D. *et al.* Zhao. Whole-genome mapping of histone H3 lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell*, 1(3):286 – 298, 2007.

[36] G. J. *et al.* Pan. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell*, 1(3):299 – 312, 2007.

[37] N. L. Vastenhouw and A. F. Schier. Bivalent histone modifications in early embryogenesis. *Current Opinion in Cell Biology*, 24(3):374 – 386, 2012.

[38] H. Chakravarthy, B. D. Ormsbee, S. K. Mallanna, and A. Rizzino. Rapid activation of the bivalent gene sox21 requires displacement of multiple layers of gene-silencing machinery. *The FASEB Journal*, 25(1):206–218, 2011.

[39] C. Hodges and G. R. Crabtree. Dynamics of inherently bounded histone modification domains. *Proc. Nat. Acad. of Sci.*, 109(33):13296–13301, 2012.

[40] I. B. Dodd, M. A. Micheelsen, K. Sneppen, and G. Thon. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell*, 129(4):813 – 822, 2007.

[41] D. David-Rus, S. Mukhopadhyay, J. L. Lebowitz, and A. M. Sengupta. Inheritance of epigenetic chromatin silencing. *Journal of Theoretical Biology*, 258(1):112 – 120, 2009.

[42] M. Sedighi and A. M Sengupta. Epigenetic chromatin silencing: bistability and front propagation. *Physical Biology*, 4(4):246, 2007.

[43] L. Ringrose and R. Paro. Polycomb/trithorax response elements and epigenetic memory of cell identity. *Development*, 134(2):223–232, 2007.

[44] M. C. *et al.* Ku. Genomewide analysis of prc1 and prc2 occupancy identifies two classes of bivalent domains. *PLoS Genet*, 4(10):e1000242, 10 2008.

[45] R. *et al.* Margueron. Role of the polycomb protein eed in the propagation of repressive histone marks. *Nature*, 461(7265):762–767, 2009. cited By (since 1996)196.

[46] K.H. *et al.* Hansen. A model for transmission of the H3K27ME3 epigenetic mark. *Nature Cell Biology*, 10(11):1291–1300, 2008. cited By (since 1996)201.

[47] D. A. Orlando, Guentherm M. G., G. M. Frampton, and R. A. Young. Cpg island structure and trithorax/polycomb chromatin domains in human cells. *Genomics*, 100(5):320 – 326, 2012.

[48] G. G. *et al.* Welstead. X-linked H3K27ME3 demethylase utx is required for embryonic development in a sex-specific manner. *Proc. Nat. Acad. of Sci.*, 109(32):13004–13009, 2012.

[49] S.M. Kooistra and K. Helin. Post-translational modifications: Molecular mechanisms and potential functions of histone demethylases. *Nature Reviews Molecular Cell Biology*, 13(5):297–311, 2012. cited By (since 1996) 18.

[50] D. *et al.* Pasini. Coordinated regulation of transcriptional repression by the rbp2 H3K4 demethylase and polycomb-repressive complex 2. *Genes Development*, 22(10):1345–1355, 2008.

[51] P. A.C. Cloos, J. Christensen, K. Agger, and K. Helin. Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. *Genes Development*, 22(9):1115–1140, 2008.

[52] S. W. Kim, S. M. Yoon, E Chuong, C Oyolu, A. E. Wills, R. Gupta, and J. Baker. Chromatin and transcriptional signatures for nodal signaling during endoderm formation in hESCs. *Developmental Biology*, 357(2):492 – 504, 2011.

[53] G.-C. Yuan. Linking genome to epigenome. *WIREs Syst Biol Med*, 4:297 – 309, 2012.

[54] E. M. *et al.* Mendenhall. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet*, 6(12):e1001244, 12 2010.

[55] M. Sedighi and A. M. Sengupta. Epigenetic chromatin silencing: bistability and front propagation. *Physical Biology*, 4(4):246, 2007.

[56] H. Binder, L Steiner, J Przybilla, T Rohlf, S Prohaska, and J Galle. Transcriptional regulation by histone modifications: towards a theory of chromatin reorganization during stem cell differentiation. *Physical Biology*, 10(2):026006, 2013.

[57] K. Sneppen and I. B. Dodd. A simple histone code opens many paths to epigenetics. *PLoS Comput Biol*, 8(8):e1002643, 08 2012.

[58] Philipp Voigt, Gary LeRoy, WilliamJ. Drury III, BarryM. Zee, Jinsook Son, DavidB. Beck, NicolasL. Young, BenjaminA. Garcia, and Danny Reinberg. Asymmetrically modified nucleosomes. *Cell*, 151(1):181 – 193, 2012.

[59] M. *et al.* Radman-Livaja. Patterns and mechanisms of ancestral histone protein inheritance in budding yeast. *PLoS Biol*, 9(6):e1001075, 06 2011.

[60] R. B. Deal, J. G. Henikoff, and S. Henikoff. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science*, 328(5982):1161–1164, 2010.

[61] B. Zee, R. Levin, P. DiMaggio, and B. Garcia. Global turnover of histone post-translational modifications and variants in human cells. *Epigenetics and Chromatin*, 3(1):22, 2010.

[62] B. M. *et al.* Zee. In vivo residue-specific histone methylation dynamics. *Journal of Biological Chemistry*, 285(5):3341–3350, 2010.

[63] K.W. Orford and D.T. Scadden. Deconstructing stem cell self-renewal: Genetic insights into cell-cycle regulation. *Nature Reviews Genetics*, 9(2):115–128, 2008. cited By (since 1996)221.

[64] N.L. Vastenhouw, Y. Zhang, I.G. Woods, F. Imam, A. Regev, X.S. Liu, J. Rinn, and A.F. Schier. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature*, 464(7290):922–926, 2010. cited By (since 1996)77.

[65] M. *et al.* Wernig. In vitro reprogramming of fibroblasts into a pluripotent es-cell-like state. *Nature*, 448(7151):318–324, 2007.

[66] Jose M. P *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell*, 151(7):1617 – 1632, 2012.

[67] A. *et al.* Calder. Lengthened g1 phase indicates differentiation status in human embryonic stem cells. *Stem Cells and Development*, 22(2):279 – 295, 2013.

[68] N. A. Hathaway, O. Bell, C. Hodges, E. L. Miller, D. S. Neel, and G. R. Crabtree. Dynamics and memory of heterochromatin in living cells. *Cell*, 149(7):1447–1460, 2012.

[69] A. Pikovsky, M. Rosenblum, and J. Kurths. Synchronizaion: A Universal Concept in Non-linear Sciences, chapter 7. Cambridge University Press, 2004.

[70] A. F. Taylor, M. R. Tinsley, F. Wang, Z. Huang, and K. Showalter. Dynamical quorum sensing and synchronization in large populations of chemical oscillators. *Science*, 323(5914):614–617, 2009.

[71] S. H. Strogatz. Sync: The Emerging Science of Spontaneuos Order. Hyperion, 2003.

[72] G. Kozyreff, A. G. Vladimirov, and P. Mandel. Global coupling with time delay in an array of semiconductor lasers. *Phys. Rev. Lett.*, 85:3809–3812, 2000.

[73] K. Wiesenfeld, C. Bracikowski, G. James, and R. Roy. Observation of antiphase states in a multimode laser. *Phys. Rev. Lett.*, 65:1749–1752, 1990.

[74] J. Zamora-Munt, C. Masoller, J. Garcia-Ojalvo, and R. Roy. Crowd synchrony and quorum sensing in delay-coupled lasers. *Phys. Rev. Lett.*, 105:264101, 2010.

[75] S. A. Marvel and S. H. Strogatz. Invariant submanifold for series arrays of Josephson junctions. *Chaos*, 19(1), 2009.

[76] S. Nichols and K. Wiesenfeld. Ubiquitous neutral stability of splay-phase states. *Phys. Rev. A*, 45:8430–8435, 1992.

[77] S. Dano, F. Hynne, S. De Monte, F. d'Ovidio, P. G. Sorensen, and H. Westerhoff. Synchronization of glycolytic oscillations in a yeast cell population. *Faraday Discuss.*, 120:261–275, 2002.

[78] S. De Monte, F. d'Ovidio, S. Dan, and P. G. Srensen. Dynamical quorum sensing: Population density encoded in cellular dynamics. *Proc. Nat. Acad. Sci.*, 104(47):18377–18381, 2007.

[79] D. C. Michaels, E. P. Matyas, and J. Jalife. Mechanisms of sinoatrial pacemaker synchronization: a new hypothesis. *Circulation Research*, 61(5):704–14, 1987.

[80] S. H. Strogatz, D. M. Abrams, A. McRobie, B. Eckhardt, and E. Ott. Crowd synchrony on the millennium bridge. *Nature*, 438(7064):43–44, 2005.

[81] B. Eckhardt, E. Ott, S. H. Strogatz, D. M. Abrams, and A. McRobie. Modeling walker synchronization on the millennium bridge. *Phys. Rev. E*, 75:021110, 2007.

[82] M. M. Abdulrehem and E. Ott. Low dimensional description of pedestrian-induced oscillation of the millennium bridge. *Chaos*, 19(1), 2009.

[83] I. Z. Kiss, Y. Zhai, and J. L. Hudson. Emerging coherence in a population of chemical oscillators. *Science*, 296(5573):1676–1678, 2002.

[84] N. Nakagawa and Y. Kuramoto. Collective chaos in a population of globally coupled oscillators. *Prog. Theor. Phy.*, 89(2):313–323, 1993.

[85] N. Nakagawa and Y. Kuramoto. From collective oscillations to collective chaos in a globally coupled oscillator system. *Physica D*, 75:74 – 80, 1994.

[86] N. Nakagawa and Y. Kuramoto. Anomalous Lyapunov spectrum in globally coupled oscillators. *Physica D*, 80(3):307 – 316, 1995.

[87] V. Hakim and W. J. Rappel. Dynamics of the globally coupled complex Ginzburg-Landau equation. *Phys. Rev. A*, 46:R7347–R7350, 1992.

[88] M. Shiino and M. Frankowicz. Synchronization of infinitely many coupled limit-cycle type oscillators. *Phys. Lett. A*, 136(3):103 – 108, 1989.

[89] P. C. Matthews and S. H. Strogatz. Phase diagram for the collective behavior of limit-cycle oscillators. *Phys. Rev. Lett.*, 65:1701–1704, 1990.

[90] P. C. Matthews, R. E. Mirollo, and S. H. Strogatz. Dynamics of a large system of coupled nonlinear oscillators. *Physica D*, 52(2-3):293 – 331, 1991.

[91] H. Daido and K. Nakanishi. Aging transition and universal scaling in oscillator networks. *Phys. Rev. Lett.*, 93:104101, 2004.

[92] H. Daido and K. Nakanishi. Diffusion-induced inhomogeneity in globally coupled oscillators: Swing-by mechanism. *Phys. Rev. Lett.*, 96:054101, 2006.

[93] K. Kaneko. Globally coupled circle maps. *Physica D*, 54(1-2):5 – 19, 1991.

[94] K. Kaneko. Clustering, coding, switching, hierarchical ordering, and control in a network of chaotic elements. *Physica D*, 41(2):137 – 172, 1990.

[95] L. Yu, E. Ott, and Q. Chen. Fractal distribution of floaters on a fluid surface and the transition to chaos for random maps. *Physica D*, 53(1):102 – 124, 1991.

[96] L. Yu, E. Ott, and Q. Chen. Transition to chaos for random dynamical systems. *Phys. Rev. Lett.*, 65:2935–2938, 1990.

[97] In addition to these attractors, it is interesting to note that we also observed the exitence of 'amplitude chimera states' which were discovered recently (G. C. Sethia and A. Sen, "Chimera States: The Existence Criteria Revisited," Phys. Rev. Lett **112**, 144101 (2013)). In this state, we observe the coexistence of a clump and a group of oscillators that behave differently and incoherently. However, in our paper we will not discuss this state further. Rather we will focus on the clump and extensive chaotic states (Figs. 1(b, c)), and their evolution with changing K.

[98] P. Grassberger and I. Procaccia. Characterization of strange attractors. *Phys. Rev. Lett.*, 50:346–349, 1983.

[99] C. Grebogi, E. Ott, and J. A. Yorke. Crises, sudden changes in chaotic attractors, and transient chaos. *Physica D*, 7(13):181 – 200, 1983.

[100] C. Grebogi, E. Ott, and J. A. Yorke. Chaotic attractors in crisis. *Phys. Rev. Lett.*, 48:1507–1510, 1982.

[101] C. Grebogi, E. Ott, F. Romieras, and J. A. Yorke. Critical exponents forcrisis-induced intermittency. *Phys. Rev. A*, 36(0):5365–5380, 1987.

[102] E. Ott. Chaos in Dynamical systems*, 2nd ed.*, chapter 3. Cambridge University Press, New York, 2002.

[103] F. Ledrappier and L.-S. Young. Dimension formula for random transformations. *Comm. Math. Phys.*, 117(4):529–548, 1988.

[104] J. L. Kaplan and J. A. Yorke. C*haotic Behavior of Multidimensional Difference Equations, in* Functional Differential Equations and Approximation of Fixed Points , volume 730, page 204. Springer, Berlin, 1979.

[105] P. Frederickson, J. L. Kaplan, E. D. Yorke, and J. A. Yorke. The Liapunov dimension of strange attractors. *J. Diff. Eq.*, 49(2):185 – 207, 1983.

[106] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Physica D*, 7(1-3):153 – 180, 1983.