RICE UNIVERSITY

Spontaneous Emergence of Hierarchy in Biological Systems

by

Jiankui He

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE:

Michael W. Deem

Michael W. Deem, Chair
John W. Cox Professor, Department of
Bioengineering and Department of Physics
and Astronomy

Oleg Igoshin
Assistant Professor, Department of
Bioengineering

Jason H. Hafner
Associate Professor, Department of Physics
and Astronomy and Department of
Chemistry

Houston, Texas
November, 2010

ABSTRACT

Spontaneous Emergence of Hierarchy in Biological Systems

by

Jiankui He

Hierarchy is widely observed in biological systems. In this thesis, evidence from nature is presented to show that protein interactions have became increasingly modular as evolution has proceeded over the last four billion years. The evolution of animal body plan development is considered. Results show the genes that determine the phylum and superphylum characters evolve slowly, while those genes that determine classes, families, and speciation evolve more rapidly. This result furnishes support to the hypothesis that the hierarchical structure of developmental regulatory networks provides an organizing structure that guides the evolution of aspects of the body plan. Next, the world trade

network is treated as an evolving system. The theory of modularity predicts that the trade network is more sensitive to recessionary shocks and recovers more slowly from them now than it did 40 years ago, due to structural changes in the world trade network induced by globalization. Economic data show that recession-induced change to the world trade network leads to an increased hierarchical structure of the global trade network for a few years after the recession. In the study of influenza virus evolution, an approach for early detection of new dominant strains is presented. This method is shown to be able to identify a cluster around an incipient dominant strain before it becomes dominant. Recently, CRISPR has been suggested to provide adaptive immune response to bacteria. A population dynamics model is proposed that explains the biological observation that the leader-proximal end of CRISPR is more diversified and the leader-distal end of CRISPR is less diversifed. Finally, the creation of diversity of antibody repertoire is investigated. It is commonly believed that a heavy chain is generated by randomly combining V, D and J gene segments. However, using high throughput sequence data in this study, the naive

VDJ repertoire is shown to be strongly correlated between individuals, which suggest VDJ recombination involves regulated mechanisms.

Acknowledgments

First and foremost, I would like to thank my adviser Dr. Michael W. Deem for his guidance and support through my years at Rice. I am really fortune to have him as my advisor. I would also like to acknowledge the advice and support of Dr. Jun Sun and Dr. Ramdas Pophale. I am grateful to Keyao Pan and Dirk Lorenz for many discussions I had.

Finally, I would like to dedicate this thesis to my fiancée Yan Zeng.

Contents

Abstract ii

Acknowledgments v

List of Figures xii

List of Tables xxi

1 Introduction 1

1.1 The big pictures 1

1.2 Organization of the thesis 6

I Hierarchy in evolving systems 12

2 Spontaneous emergence of modularity 13

2.1 Introduction 13

2.2 A definition of compositional age 15

2.3 Compositional age and evolutionary rate 19

2.4 Growth of modularity in the protein-protein interaction network 19

2.5 Growth of modularity in the domain-domain interaction network 27

2.6 Conclusion 31

3 Hierarchical evolution of animal body plans 33

3.1 Introduction 33

3.2 Materials and Methods 38

3.2.1 Sea urchin gene regulatory network 38

3.2.2 Ratio of nonsynonymous substitution to synonymous substitution of genes 39

3.3 Results 42

3.3.1 Evolutionary rate in hierarchy 42

3.3.2 Generative entrenchment 45

3.3.3 Time of appearance of regulatory genes 48

3.4 Discussion 50

4 Structure and Response in the World Trade Network **54**

4.1 Introduction 54

4.2 Results 57

4.2.1 Measure of hierarchy by CCC 57

4.2.2 Evolution of structure in world trade 60

4.2.3 Response of world trade to recessions 66

4.3 Conclusion 71

II Influenza virus evolution **73**

5 Prediction of incipient dominant influenza strain by clustering **74**

5.1 Introduction 74

5.2 Results 78

5.2.1 Evolutionary path of 2009 A(H1N1) influenza 78

5.2.2 Low-dimensional clustering 81

5.2.3 H3N2 virus evolution for 38 years 87

5.2.4 Influenza vaccine strain selection 89

5.2.5 Demonstration of low-dimensional sequence clustering method. . 95

5.2.6 Prediction for H3N2 influenza in 2009–2010. 98

5.2.7 Comparison with previous results. 101

5.2.8 Detecting A/Wellington/1/2004 in the 2004 flu season in the Southern hemisphere 106

5.2.9 Detecting A/California/4/2004 as a future dominant strain 108

5.3 Discussion 112

5.4 Materials and Methods 114

5.4.1 Data sources 114

5.4.2 Geographical spread pattern of 2009 A(H1N1) 115

5.4.3 Multidimensional scaling 116

5.4.4 Biases in the data 121

III Bacterial and animal immune systems 122

6 Heterogeneous Diversity of Spacers within CRISPR 123

6.1 Introduction 123

6.2 Results 125

6.2.1 Differential equation model 126

6.2.2 Stochastic simulation model 133

6.2.3 Experimental results 137

6.3 Conclusion 140

7 Regulated mechanism in antibody VDJ recombination 142

7.1 Introduction 142

7.2 Results 144

7.2.1 Correlation in the naive VDJ repertoire 144

7.2.2 Regulated model of VDJ recombination 152

7.3 Conclusion and discussion 159

8 Conclusion **161**

References **163**

List of Figures

2.1 Distribution of conserved sequences with compositional age to find (a) age of LUCA, and (b) divergence time of fungi. 16

2.2 The dN/dS and compositional age of proteins in *S. cerevisiae*. As measured by the average dN/dS, newer genes are evolving more rapidly than older genes. The correlation coefficient is $R^2 = 0.82$ 20

2.3 The degree distribution of the *S. cerevisiae* domain-domain interaction network. 21

2.4 The reordered topological overlap matrix of the *E. coli* protein interaction network constructed from proteins whose compositional age are larger than 12.8 (a), 12.6 (b), and 12.2 (c). (d) The linear relationship between compositional age and real age. (e) and (f), The banded modularity evolution of *E. coli* and *S. cerevisiae*, respectively. The lines of different color in (e) and (f) correspond to different band sizes (W). 23

2.5 Evolution of module modularity of protein interaction network in *E. coli* (a) and *S. cerevisiae* (b). 27

2.6 (a) Average number of proteins in a module at different compositional ages, (b) Size of network in different compositional age network. 28

2.7 Evolution of banded modularity of domain-domain interaction network in *E. coli* (a) and *S. cerevisiae* (b). 28

2.8 Domain interaction network modularity evolution in *E. coli* (a) and *S. cerevisiae* (b). Score is the inverse of modularity. 31

3.1 The gene regulatory network of sea urchin endomesoderm specification up to 30 hours. The top five genes form the kernel. 40

3.2 The hierarchy of the gene regulatory network and functions at different levels of development of the body plan. 41

3.3 The ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions for different components of the gene regulatory networks that control the development of animal embryos. 44

3.4 The earliest appearance time of regulatory genes for kernels, plug-ins and I/Os as experimental data available(Davidson et al., 2002). 47

3.5 The distribution of dN/dS for each hierarchical level. P(dN/dS) is the probability of a gene with the dN/dS in specific hierarchical level. 48

4.1 Dendrogram representation of trade networks for selected countries at 1969 (top figure) and 2007 (bottom figure). 58

4.2 The CCC from 1969 to 2007. The upper right insert is the ratio of total world trade to world GDP. The lower left insert is the total world trade in units of US dollar. 61

4.3 The trade share matrix $S_{ij} = M_{ij} / \left(\sum_{m=1}^{N} M_{im} + \sum_{n=1}^{N} M_{jn} \right)$ after hierarchical clustering between countries in 2007. 64

4.4 (a) The ratio of the total world excluding the USA GDP change (percentage) to the change of the USA GDP (percentage) in 5 recessions, F in the y axis. (b) Impulse response analysis of spread of recession. The world GDP change is plotted as a function of the CCC. The reduction in the world GDP is greater when the CCC value is low. Insert figure: The GDP recovery from recession can be well fit by the relation $Y(t) \sim Y(\infty) - a\exp(-\lambda t)$. Yearly recovery rates, λ, are shown versus the CCC. In accord with theory, the recovery rate is positively correlated with the CCC. 68

5.1 The evolutionary path of 2009 A(H1N1) influenza. 82

5.2 (a), Kernel density estimation for the protein distance map of 2009 A(H1N1) influenza as of December 5, 2009. (b), The protein distance map of 2009 A(H1N1) influenza. 84

5.3 (a) The protein distance map and (b) corresponding Kernel density estimation of influenza from 1968 to 2007. 90

5.4 (a) Kernel density estimation and (c) protein distance map for H3N2 viruses between October 1, 2002 and February 1, 2003. (b) Kernel density estimation and (d) protein distance map for H3N2 viruses between October 1, 2001, and September 9, 2002. A 0.0030 unit of protein distance equals one substitution of the HA1 protein sequence of H3N2. 99

5.5 (a) Kernel density estimation and (c) protein distance map for H3N2 viruses from October 1, 2008, to June 14, 2009. (b) Kernel density estimation and (d) protein distance map for H3N2 viruses between October 1, 2008, and March 30, 2009. A 0.0030 unit of protein distance equals one substitution of the HA1 protein sequence of H3N2. 102

5.6 (a), Kernel density estimation for the protein distance map for H3N2 viruses between 10/01/2003 and 09/30/2004. (b), Kernel density estimation for the protein distance map for H3N2 viruses between 10/01/2003 and 02/01/2004. (c), Protein distance map for H3N2 viruses between 10/01/2003 and 09/30/2004. We plot a dotted line to separate the two clusters. (d), Protein distance map for H3N2 viruses between 10/01/2003 and 02/01/2004. The vertical and horizontal axes of all figures represent protein distance. A 0.0030 unit of protein distance equals one mutation of the HA1 protein sequence of H3N2. 109

5.7 Plot of Euclidean distances of proteins as in Fig. 5.4(d) on x-axis and plot of distance of corresponding proteins in y-axis. Closeness to the diagonal measures fidelity of the low dimensional projection. A 0.0030 unit of protein distance equals one mutation of the HA1 protein sequence of H3N2. 118

6.1 A schematic representation to describe CRISPR-phage dynamics. 127

6.2 Diversity of two spacers of CRISPR with time. 132

6.3 Diversity of spacers at different positions of CRISPR, when the system reaches steady state. 136

6.4 Diversity of spacers of CRISPR loci 1 of *S. thermophilus* strains[71]. The positions with a small number in the x-axis are leader-proximal. 138

6.5 Diversity of spacers of CRISPR loci of *Leptospirillum* species. The data are noisy because the CRISPR loci sequence data of *Leptospirillum* are fragmented. 139

7.1 VDJ repertoire correlation analysis for all 14 fish. (a) Correlation of VDJ repertoire with highly expressed VDJ combinations removed. (b) Histogram of correlation of VDJ repertoire with the top 2% of most expressed VDJ combinations removed. In both figures, the simulation is conducted by randomly swapping the sequence reads between VDJ combinations. . 146

7.2 Naive and activated VDJ repertoire correlation analysis for all 14 fish. . 148

7.3 Correlation matrix of the activated VDJ repertoire. Only fish 4 and 6, fish 12 and 13 have strong correlations. 149

7.4 (a) Correlation D^n between naive VDJ repertoire $T^n_{ijk}(N)$ and model estimation P^n_{ijk}. (b) Fitting quality as illustrated by fish 5. 154

7.5 $P^n_i(V)$ as estimated from the model for all 39 V segments in 14 fish. . . 155

7.6 Average $P_i(V)$ over 14 fish. 156

7.7 $P_k^n(J)$ as estimated from the model for all 5 J segments in 14 fish. Fish 10 and 14 have distinct probability distribution. Recall in Figure 7.4(a) that the model prediction fits the data of fish 10 and 14 less well than the other fish. 157

7.8 Average $P_k(J)$ over 14 fish. 158

List of Tables

3.1 Evotionary rate of regularoty genes in pairs of organisms (Org.1 and Org. 2). Group 1 are kernels, group 2 are plug-ins, group 3 are I/O, group 4 are batteries. STRPU: *Strongylocentrotus purpuratus*, PATVU: *Patella vulgata*, ASTM: *Asterina miniata*, HEMPU: *Hemicentrotus pulcherrimus*, PARLI: *Paracentrotus lividus*, HELER: *Heliocidaris erythrogramma*, HELTB: *Heliocidaris tuberculata*, LYTVA: *Lytechinus variegatus.* 52

3.2 P-value of Wilcoxon test for different hierarchical levels. The hypothesis for the Wilcoxon test is that two independent samples come from distributions with the same median. 53

5.1 Summary of results 105

5.2 The geographical spread pattern of 2009 A(H1N1). "Others" refers to other countries except USA and Mexico. 116

5.3 Consensus strain 1 is the calculated from all strains in the cluster on the right side of Figure 5.5(c). Consensus strain 2 is the calculated from all strains in the cluster on the left side of Figure 5.5(c). 119

Chapter 1
Introduction

The main goal of the research topic in this Doctoral Thesis is the statistical investigation of the modularity, diversity, and stochasticity in evolving systems. This thesis contains three parts covering modularity and hierarchy, influenza virus evolution, and immune systems [63, 60, 61, 62, 59].

1.1 The big pictures

The main theme of biology in twentieth-century is an attempt to reduce biological phenomena to the behavior of molecules [57]. Enormous success has been achieved with this approach. However, a discrete biological function can only rarely be attributed to an individual molecule. In most cases, biological behavior and functions arises from the complex interactions of proteins, genes and many other components. For example, the signal transduction systems involves signal receptors, chemical messagers, molecules for amplifying signals, proteins related to gene expression and so on. To understand the organization of complex interaction networks of molecules, we need new concepts to describe the "design principles" of biological systems, profoundly shaped by evolution. We argue here that "hierarchy" is a critical level of biological organization.

Hierarchy and modularity are prevalent in biology. Generally, a "module" is a sub-network that has more internal edges than external edges, or is composed of features that act together in performing some discrete function that is semi-autonomous in relation to others. For example, some proteins are composed of several domains with independent function. Each domain in these proteins is a module. The gene regulatory network that control the animal body plan development is also organized into modules. Different modules controls development of different patterns. Modules are often hierarchically organized and hierarchy is a multi-level organization of modularity. Hierarchy is also observed in other fields. In the world trade, economies of individual countries organize into groups, or modules, that trade more with other countries in the group than with countries outside the group. For example, the US, Mexico, and Canada trade more with each other than with other countries. These trading groups organize into higher-level groups and so on. In this way the world trade network organizes into a hierarchical structure.

Modularity arises in the evolution. Recent studies have proposed the theory of evolution of modularity in rugged landscape [113, 120, 118, 119, 19, 156, 128]. This theory can be summarized in the formula:

$$P_E = \frac{M'}{R} \tag{1.1}$$

Here, P_E is environmental pressure, R is the resistance to evolving, and M' is the rate of change of modularity. This theory state that in the rugged fitness landscape, where evolution is relatively slow, the rate of change in modularity is proportional to environmental pressure. According to this theory, the modularity is inevitable in evolution under three conditions: 1, the evolution is in a rugged fitness landscape and therefore relatively slow; 2, the environment is changing; and 3, horizontal gene transfer is present. We will show in Chapter 2 that the growth of modularity in the evolution of protein interaction and domain interaction networks.

Hierarchy can increase the evolvability of biological systems [107, 126, 31, 91]. The space of all genotype is exponentially large. For example, random searching for fitness maxima in the landscape seems costly and nearly impossible even on the evolutionary time scales. For example, a single nucleotide mutation in the genome rarely increase the fitness. But a system that can be decomposed into modules can evolve one module at a time. Since the subspace of modules is much smaller than the whole space of all genotype, it is much easier and faster to find the local fitness maxima in modules. Also, a modular structure to the molecules of life allows for biological information to be stored in pieces. Evolution can proceed not just by changing one base of the genetic code or movement of one atom or amino acid at a time, but rather by exchange of these functional pieces among living organisms. In addition, embedding particular functions in the discrete modules allows the core function of a module to be robust to change, and allows for changing of cell's function and properties by alerting the connections between modules. We will show in Chapter 3 that the evolution of animal body plan occurs in a hierarchical way, in which the the core modules are resistant to change and the periphery modules evolve much faster.

Hierarchy can increase the robustness of the systems [77]. Robustness enables the system to main functionalities again external and internal perturbations. Because the weak interactions between modules can buffer the effect of perturbation, hierarchy retains the impacts of a perturbation within a single module, while minimizing the effects on the whole systems. We will show in Chapter 4 that hierarchy in the world trade network can help countries to recover more quickly from recessions and to reduce the loss in recessions.

1.2 Organization of the thesis

This thesis is organized into 3 parts. The first part discusses the modularity and hierarchy in evolving systems. The second part discusses influenza virus evolution and influenza vaccine design. The third part discuss the bacterial immune system CRISPR and VDJ recombination in animal immune systems. This thesis contains 6 different chapters. Each chapter documents findings from a different research projects and is self-contained.

Chapter 2 documents the spontaneous of modularity in biological systems [63]. Theories of protein structure postulate a universal, primordial diversity of folds, from which all proteins are constructed. It has been further argued that an increasingly diverse array of selective pressures upon proteins as evolution has proceeded may have lead to compartmentalization of protein functions into discrete modules. Scant evidence exists to date as to whether modularity has increased or decreased with evolutionary progress. Here we show that protein interactions became increasingly modular as evolution proceeded over the last four billion years. We also introduce a new method to determine the divergence time of a protein. We suggest that the modularity of protein interactions arose as the mechanism by which the increasingly large and complex protein interaction network maintained the ability to evolve. As evolution proceeded, and the diversity of species increased and the environment changed, proteins became more modular and specialized in their interactions.

Chapter 3 documents the hierarchical evolution of animal body plan development [60]. An open question in animal evolution is why the phylum- and superphylum-level body plans have changed so little, while the class- and family-level body plans have changed so greatly since the early Cambrian. Davidson and Erwin [32] proposed that the hierarchical structure of gene regulatory networks leads to different observed evolutionary rates for terminal properties of the body plan versus major aspects of body plan morphology. In this chapter, the speed of evolution of genes in these gene regulatory networks is calculated. We found that the genes which determine the phylum and superphylum characters evolve slowly, while those genes which determine the classes, families, and speciation evolve more rapidly. This result furnishes support to the hypothesis that the hierarchical structure of developmental regulatory networks provides an organizing structure which guides the evolution of aspects of the body plan.

Chapter 4 documents the structure and response in world trade network [62]. We examine how the structure of the world trade network has been shaped by globalization and recessions over the last 40 years. We show that by treating the world trade network as an evolving system, theory predicts the trade network is more sensitive to recessionary shocks and recovers more slowly from them now than it did 40 years ago, due to structural changes in the world trade network induced by globalization. We also show that recession-induced change to the world trade network leads to an increased hierarchical structure of the global trade network for a few years after the recession.

Chapter 5 documents the evolution of influenza viruses [61]. Influenza has been circulating in the human population and has caused three pandemics in the last century (1918 H1N1, 1957 H2N2 and 1968 H3N2). The 2009 A(H1N1) was classified by World Health Organization as the fourth pandemic. Influenza has a high evolution rate, which makes vaccine design challenging. We here consider an approach for early detection of new dominant strains. By clustering the 2009 A(H1N1) sequence data, we found two main clusters. We then define a metric to detect the emergence of dominant strains. We show on historical H3N2 data that this method is able to identify a cluster around an incipient dominant strain before it becomes dominant. For example, for H3N2 as of 30 March 2009, the method detects the cluster for the new A/British Columbia/RV1222/2009 strain. This strain detection tool would appear to be useful for annual influenza vaccine selection.

Chapter 6 documents the bacterial immune systems [59]. Clustered regularly interspaced short palindromic repeats (CRISPR) in bacterial and archaeal DNA have recently been shown to be a new type of antiviral immune system in these organisms. We here study the diversity of spacers in CRISPR under selective pressure. We propose a population dynamics model that explains the biological observation that the leader-proximal end of CRISPR is more diversified and the leader-distal end of CRISPR is more conserved. This result is shown to be in agreement with recent experiments. Our results show that the CRISPR spacer structure is influenced by and provides a record of the viral challenges that bacteria face.

Chapter 7 documents VDJ recombination in animal immune systems. It is commonly thought that the VDJ recombination is a random process and therefore there should be no correlation in the antibody repertoire. We developed computational methods to construct the naive antibody repertoire from high throughput zebrafish sequence data. We found that the naive VDJ repertoire are strongly correlated between individual fish, which suggest VDJ recombination involves regulated mechanisms. We further propose a model that the frequency of a particular VDJ combination is determined by the product of frequency of its V, D and J gene segments. This model can produce the original data and it provides insight on how VDJ recombination is regulated. This study allow us to understand the creation of the diversity of immune response and direct experiments to uncover the mechanism of VDJ recombination.

Part I

Hierarchy in evolving systems

Chapter 2

Spontaneous emergence of modularity

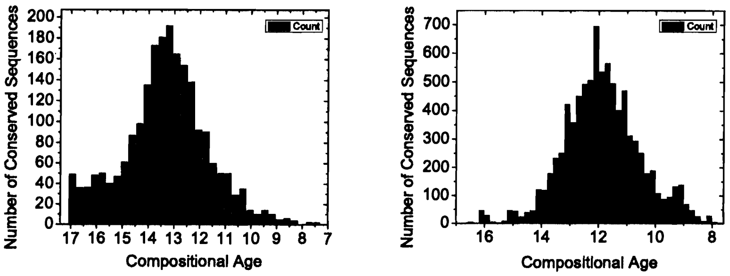
2.1 Introduction

Modularity and hierarchy are ubiquitous in biology, compartmentalizing information of and interactions among genes and proteins [104, 57, 17, 23]. Levels of hierarchy span atoms, amino acids, secondary structures, proteins, pathways, cells, tissues, organs and organisms [84, 116, 105]. Physical methods have been used to characterize modularity in network systems [81, 115]. For example, selection for stability of function has been shown to lead to modular networks [124]. Network motifs have been identified for the transcriptional regulation network of *E. coli* [106]. Once modularity has arisen, so that the environment a species faces is modular, these modularly varying goals were shown to select for modular structure [76, 9]. A number of other theories have been put forward that suggest modularity may grow with evolutionary progress [120, 118, 119, 19, 156, 128]. There are suggestions that by being modular, for example, a system will be more robust to perturbations and more evolvable [107, 126, 31, 77, 91]. On the other hand, there is a selective pressure for evolvability in a population evolving in a changing environment [38]. It has been hypothesized, therefore, that modularity will arise spontaneously in a population of individuals evolving in a changing environment [34]. Support for this hypothesis had been elusive [51], until the recent theoretical evidence that environmental change coupled with horizontal gene transfer inevitably and generically leads to the evolution of modular structures [113]. To date, experimental evidence in support of these theories has been difficult to come by, since we can not go back in evolutionary time to observe growth of modularity. By introducing a method to date the divergence time of proteins and a quantitative definition of modularity, we here show that modularity in protein-protein and domain-domain interaction networks has grown as evolution proceeded over the last 3.5 billion years.

2.2 A definition of compositional age

To study modularity in biology, both a quantitative definition of modularity and a calibration of evolutionary time for the biological objects of interest are needed. In this study, the compositional age approach is used to quantify the divergence time of a protein [112]. In this method the order of appearance of the amino acids over evolutionary time is identified. Proteins that contain a greater fraction of the oldest amino acids are then identified as arising earlier than those proteins that contain a greater fraction of the newer amino acids. By averaging the compositional age of each of the proteins in a species, the average evolutionary time of that species is determined. In this chapter, we make this method quantitative, calibrating it upon evolutionary time points over the last 3.5 billion years.

To find the time of divergence of the earliest proteins, 9 bacteria, 3 archaea, and 4 eukaryotic organisms are selected to find the conserved sequences presumed to have



(a) (b)

Figure 2.1 Distribution of conserved sequences with compositional age to find (a) age of LUCA, and (b) divergence time of fungi.

arisen from LUCA (Last Universal Common Ancestor). The bacterial species are *A. aeolicus, T. maritima, D. radiodurans, F. nucleatum, T. pallidum, C. glutamicum, C. acetobutylicum, S. aureus*, and *E. coli.* The archaea species are *A. fulgidus, S. solfataricus*, and *P. aerophilum.* The eukaryote species are *C. elegans, S. cerevisiae, S. pombe*, and *D. melanogaster.* All the sequence data come from EMBL-EBI. Using the software CONSERV (http://www.gen-info.osaka-u.ac.jp/ ngoto/CONSERV/), 2163 conserved sequences are found with greater than 7 amino acids that appear in all the three kingdoms and in at least 8 proteins. The compositional age for these sequences is calculated. A histogram is shown in Fig. 2.1(a). The distribution of compositional age peaks at 13.32. There is some debate about the age of LUCA, with estimates ranging from 3.5 to 4.0 billion years ago [65]. In this study, we set LUCA at 3.8 billion years ago. Therefore a compositional age of 13.32 corresponds to a real age of 3.8 billion years.

To find the divergence times of fungal proteins, 10 species of fungi are investigated. In the group Dikarya/Ascomycota/Saccharomycotina, we choose *S. cerevisiae, C. glabrata, K. lactis, Y. lipolytica*, and *P. stipitis.* In the group Dikarya/Ascomycota/Pezizomycotina, we choose *N. crassa, M. grisea*, and *A. fumigatus.* 8535 sequences are found with greater than 15 amino acids that appear in both branches and in at least 4 proteins. The histogram of compositional age of these sequences is shown in Fig. 2.1(b). The compositional age peaks at 12.1. 1.1 billion years ago is therefore chosen as the real age of divergence time of these two branches of fungi [65]. So, the compositional age of 12.1 corresponds to an evolutionary age of 1.1 billion years.

To find the compositional age of recent proteins, we search for the youngest proteins in *E. coli.* Only proteins in the COG (Clusters of Orthologous Groups of proteins) database are considered, to exclude those protein fragment without function in the FASTA file. We compare the proteins in two strains of *E. coli*: K12 and o157 :H7 EDL 933. The 0157 strain of *E. coli* diverged from K12 strain about 4 million years ago [97]. We take the strains of *E. coli* from the COG database that exclude the orthologous proteins that are shared by K12 and O157, which should be quite young, probably less than 4 million years. The youngest new protein of O157 has compositional age 9.607. The youngest new protein of K12 has compositional age 9.652. We, therefore, set the compositional age of the present as 9.6.

2.3 Compositional age and evolutionary rate

To test the definition of composition age, the relationship between the compositional age of proteins and the evolutionary rate of the corresponding genes is determined. As the measure of evolutionary rate, we use the commonly accepted ratio of nonsynonymous to synonymous substitutions per site (dN/dS). Hirsh et al. provided dN/dS for 3392 genes from orthologous open reading frames (ORFs) in four species of yeast [68]. We bin the proteins by compositional age and calculate the average dN/dS. Results are shown in Fig. 2.2. Newer genes are evolving more rapidly than older genes.

2.4 Growth of modularity in the protein-protein interaction network

We now turn to modularity. Modularity of both protein domain structure and of the protein-protein interaction network are quantified [12, 88, 87]. The protein-protein interaction network data come from DIP. 1846 proteins are obtained with 6971 interaction edges in *E. coli* and 3211 proteins with 17535 interaction edges in *S. cerevisiae.*

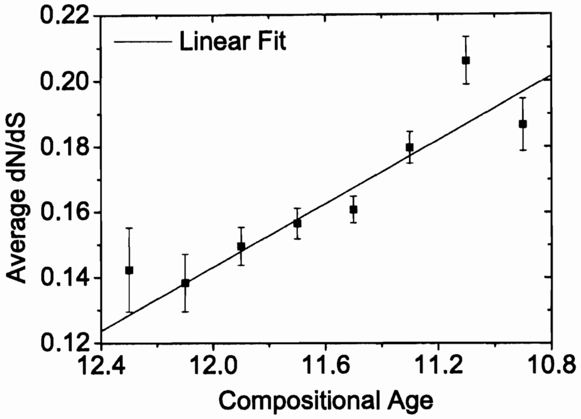


Figure 2.2 The dN/dS and compositional age of proteins in *S. cerevisiae*. As measured by the average dN/dS, newer genes are evolving more rapidly than older genes. The correlation coefficient is $R^2 = 0.82$

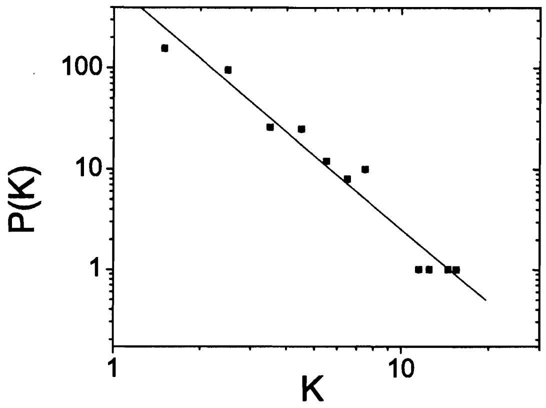
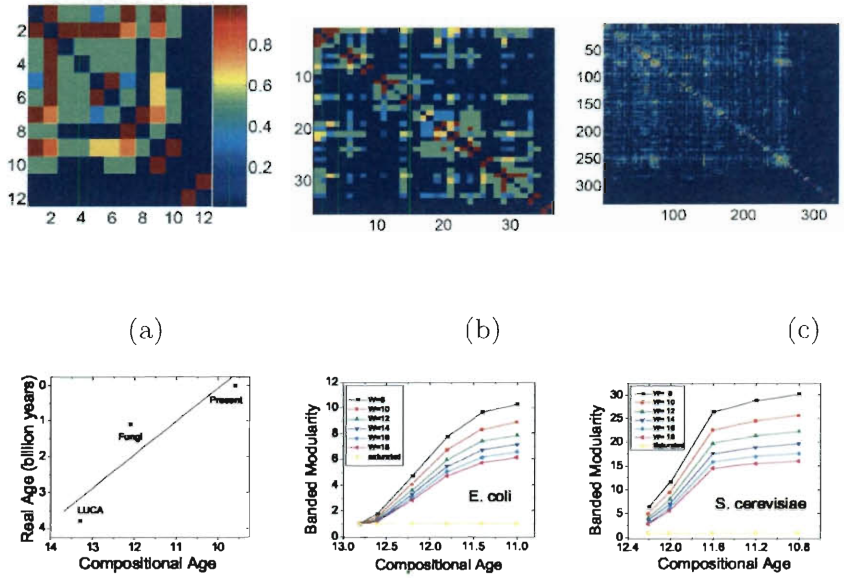


Figure 2.3 The degree distribution of the *S. cerevisiae* domain-domain interaction network.

The domain-domain interaction data come from InterDom. We only consider domain interactions based on the DIP database and take only these domain interactions with a score in the top 75%, to eliminate the noisy data. 276 proteins are obtained in *E. coli* and 427 proteins in *S. cerevisiae*, from which we extract the protein domains for study. Interestingly, the domain-domain interaction network is scale free with $\gamma = 2.4$, see Fig. 2.3.

To quantify modularity in the interaction networks, we construct the topological overlap matrix [96] from the interaction network, reorder it with the average linkage clustering method [39], and normalize the number of interactions within modules according to network size. The topological overlap matrix element, T_{ij}, is the ratio of common nearest neighbors of the interacting proteins i and j to their respective degrees. The topological overlap matrix reflects the topological overlap of the nearest neighbors of two nodes. For any two nodes i and j, the topological overlap is defined as [96]: $T_{ij} = \frac{\sum_u a_{iu}a_{uj}+a_{ij}}{\min(k_i,k_j)+1-a_{ij}}$. Here a_{ij} is the elements of the interaction network matrix with value 0 (not interacting) or 1 (interacting). The average-linkage hierarchical clustering algorithm is then used [96] to reorder the topological overlap matrix so that the more tightly linked and clustered nodes are moved close to each other. In this way, the modules and hierarchical structure of the network are identified.

The reordered topological overlap matrix of *E. coli* at different times is shown in Fig. 2.4. The color reflects the strength of the topological overlap of two nodes (from 0.0 to 1.0), as shown in the color bar in Fig. 2.4(a). The protein-protein interaction



(a)

(b)

(c)

(d)

(c)

(f)

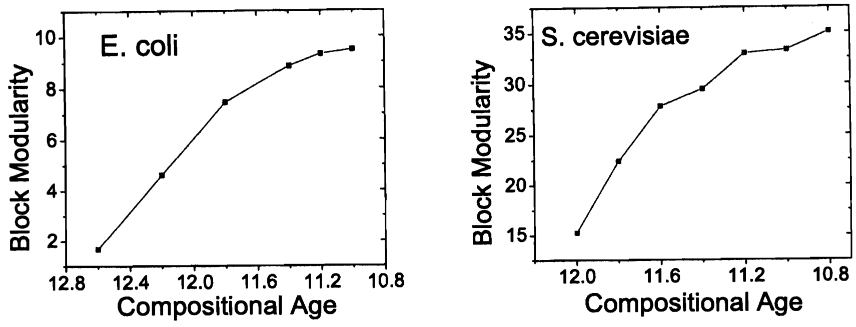
Figure 2.4 The reordered topological overlap matrix of the *E. coli* protein interaction network constructed from proteins whose compositional age are larger than 12.8 (a), 12.6 (b), and 12.2 (c). (d) The linear relationship between compositional age and real age. (e) and (f), The banded modularity evolution of *E. coli* and *S. cerevisiae*, respectively. The lines of different color in (e) and (f) correspond to different band sizes (W).

network evolves from an almost saturated, unstructured network in Fig. 2.4(a) to a mildly modular network with four modules in Fig. 2.4(b) and then to a highly modular network in Fig. 2.4(c). To compare the modularity quantitatively, we define banded modularity as the ratio of interaction within a diagonal band to the total interactions, normalized by the ratio of the area of the band to the area of the matrix: $M_{\text{banded}} = \frac{\sum^{D}_{0<|i-j|<W} T_{ij}}{\sum^{D}_{i\neq j} T_{ij}} * (\frac{\sum^{D}_{0<|i-j|<W} 1}{\sum^{D}_{i\neq j} 1})^{-1}$. Here, W is the width of the band, D is the dimension of matrix and T_{ij} is the element of reordered topological overlap matrix. Since the network size grows in time, modularity of network of different sizes is compared. The factor $(\frac{\sum^{D}_{0<|i-j|<W} 1}{\sum^{D}_{i\neq j} 1})^{-1}$ normalizes for the size effect.

Modularity grows with evolutionary time. In Fig. 2.4(e), the banded modularity grows with compositional age in *E. coli.* The similar result is observed in *S. cerevisiae* in Fig. 2.4(f). Banded modularity of a saturated matrix, *i.e.*, a matrix with all elements being 1 except the diagonal ones being 0, is shown in Fig. 2.4 (e) and (f) for comparison. The banded modularity of a saturated network is 1. This result holds true for different band widths and different organisms; this phenomenon is robustly observed. In a modular structure, there are more interactions within a module than between modules. Banded modularity is a concise definition of modularity, but may also be interpreted as simply locality, in which true modules may not be identifiable.

To measure modularity in a more detailed way, we search along the diagonal of the reordered topological overlap matrix to find the explicit modules, and the ratio of interactions in the modules to the total interactions is calculate, normalized by the ratio of the area of modules to the area of the whole matrix. These modules are defined quantitatively. First, we suppose the protein i and $i+1$ form a module, and we ask whether another another protein $i+2$, should be added to the module. The protein is added if the average interaction between $i+2$ and the existing module is larger than a cutoff, which is set as 0.2 in this study. This procedure continues. When it comes to a protein with average interaction less than the cutoff, this protein forms the first member of a new module, and we begin the search to add further proteins to this new module. The modules so identified depend on the cutoff. In this study, the *E. coli* and *S. cerevisiae* networks are highly modular. We tried several cutoff and found the results are quite stable, with results in accord with visual observation of the clustered matrix. The result is defined as as module modularity: $M_{\text{module}} = \frac{\sum'^{D}_{j,k\neq j=1} T_{jk}}{\sum^{D}_{j,k\neq j=1} T_{jk}} * (\frac{\sum'^{D}_{j,k\neq j=1} 1}{D(D-1)})^{-1}$, where in the upper sum with the prime, k is over those proteins in the same module as j, D is the dimension of the matrix.

This definition is applied to the reordered topological overlap matrix to obtain the result for *E. coli* and *S. cerevisiae* in Fig. 2.5. The growth of module modularity in both organisms is observed. There is a positive correlation between banded and modular modularity. The growth of modularity is robust to the precise definition of modularity. The average size of module at different compositional age network is stable, see Fig. 2.6(a). The relationship between the size of the network and compositional age is show in Fig. 2.6(b). The average module size does not change much in evolution, and the number of proteins in each module in of *S. cerevisiae* is fewer than that in *E. coli*,



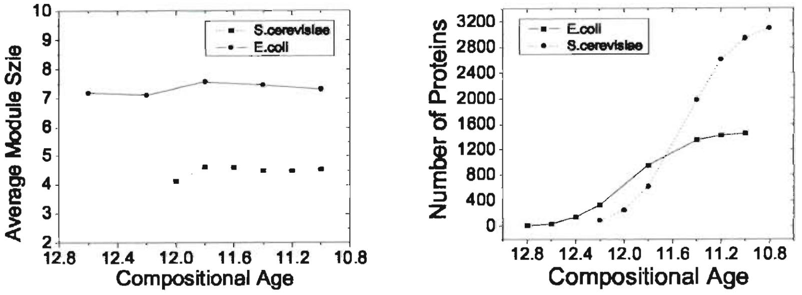
(a) (b)

Figure 2.5 Evolution of module modularity of protein interaction network in *E. coli* (a) and *S. cerevisiae* (b).

perhaps reflecting that *S. cerevisiae* is more modular.

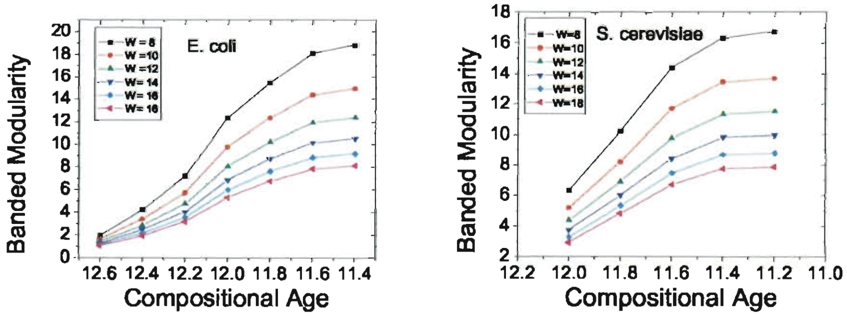
2.5 Growth of modularity in the domain-domain interaction network

We observed modularity not only in the protein-protein interaction network, but also in the domain-domain interaction network. The result of the banded modularity of the domain-domain interaction network of *E. coli* and *S. cerevisiae* is shown in Fig. 2.7. The growth of banded modularity is pronounced in both cases.



(a) (b)

Figure 2.6 (a) Average number of proteins in a module at different compositional ages, (b) Size of network in different compositional age network.



(a) (b)

Figure 2.7 Evolution of banded modularity of domain-domain interaction network in *E. coli* (a) and *S. cerevisiae* (b).

Our definitions of modularity allows the comparison of modularity of matrices of different sizes. The saturated interaction matrix does not have any modular structure, regardless of the band size, as shown in Fig. 2.4(e),(f). A network generated by randomly selected proteins in *E. coli* is of constant low modularity, independent of the number of proteins used. The network constructed based on its compositional age, however, shows a clear growth of its modularity. This result shows that the validity of organizing proteins by their compositional age.

Modularity of the unweighted domain-domain interaction network is also measured directly, without construction of topological overlap matrix. We determine the fraction of a protein to which other proteins interact. To the extent that interactions become more localized within proteins, the protein is defined to be more modular. If protein B interacts with protein A, and the interaction is with only a few of the domains of protein A, then this interaction is more modular than if protein B interacts with a greater number of the domains of protein A. Averaging this measurement over all proteins B, this procedure gives a measure of the modularity of protein A. So, we calculate the ratio of interacting domains to the number of domains in a protein, which gives the inverse of modularity. we define a Score, which is the inverse of modularity, as: $\frac{1}{2N}\sum_{l=1}^{N}(\frac{I_l^A}{D_l^A L_B^{2/3}}+\frac{I_l^B}{D_l^B L_A^{2/3}})$. Here l represents a protein-protein interaction or a link. To distinguish the two proteins in a link, one protein is marked as A, the other one as B. The number of links is N. The term L_A (L_B) is the number of amino acids of protein A (B). The number of interacting domains is I_A, and the number of total domains is D^A in protein A. The ratio of $\frac{I^A}{D_A}$ is normalized by the surface area of the target protein $L_B^{2/3}$, and so the Score should measure only the modularity and normalize out the size effect of target proteins.

In Fig. 2.8, we compare the Scores of different domain-domain interaction network at different compositional age. The inverse of the Score increases monotonically with evolutionary progress. Because the inverse of the Score is modularity, we again observe that modularity has increased through evolutionary time. This observation is robust under different definitions of the Score.

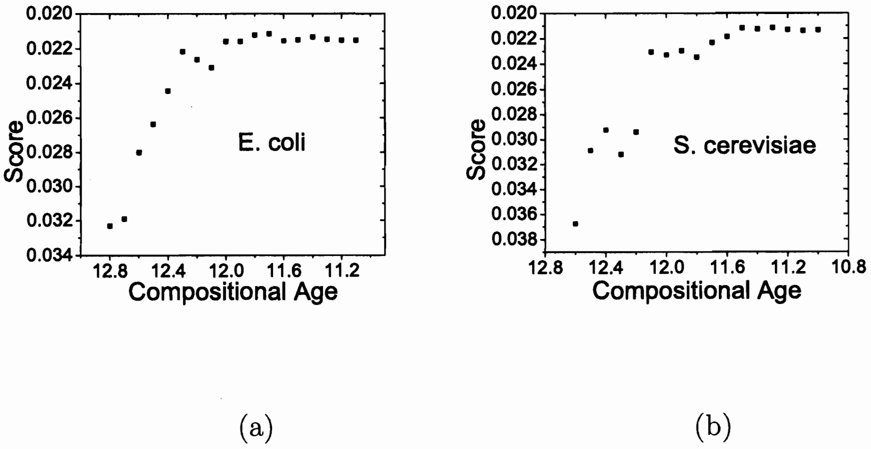


Figure 2.8 Domain interaction network modularity evolution in *E. coli* (a) and *S. cerevisiae* (b). Score is the inverse of modularity.

2.6 Conclusion

We have introduced several quantitative definitions of modularity for interacting networks. We use them to measure the modularity of the protein-protein interaction network and domain-domain interaction network in *S. cerevisiae* and *E. coli.* We have also introduced a method to quantify the evolutionary divergence time of proteins. We consistently find that modularity, by all definitions and in both organisms, has grown through evolutionary time. This observation is in agreement with the theory that environmental change coupled with horizontal gene transfer naturally and inevitably leads to evolution of increased modularity [113]. In this sense, early life was a generalist, being less modular. As evolution proceeded, and diversity of species increased and the environment changed, proteins became more modular and specialized in their interactions.

Chapter 3

Hierarchical evolution of animal body plans

3.1 Introduction

The Cambrian explosion has been an extensively debated topic in animal evolution for more than one century [127, 154]. Biological organisms were composed of individual cells, occasionally organized into colonies, before the Cambrian explosion [154, 64]. Subsequent to the Cambrian explosion, evolution greatly speed up, and the major phyla appeared. For example, the bilateral, anterior-posterior organization of body plan appears in fossil records from the early Cambrian [25]. These results are the basis for the open question in animal evolution of why the phylum- and superphylum-level body plans have changed so little and no more new phylum- and superphylum- body plans appeared, while the class- and family-level body plans have changed so greatly with so many class, family, and species appearing since the early Cambrian [123]. Since the development of the animal body plan is precisely controlled by gene regulatory networks, the mechanism to explain the different rates of change of the phylum- and superphylum-level body plans versus the class- and family-level body plans may lie in the structure and evolution of gene regulatory networks.

If the gene regulatory network were an unstructured or nearly random network, any change to the network such as deleting one gene would result in drastic difference in the body plan, because each gene may regulate or be regulated by several other genes, and the effects of deletion will spread out to the whole network quickly[15]. To resolve this problem, Davidson and Erwin [32, 42] proposed that the classic evolution theory based on selection of changes upon an unstructured genetic framework does not provide a satisfactory answer for the mechanism. Instead, they constructed the gene regulatory networks that control the early development of animal embryos (see Fig. 3.1) [80] and proposed a hierarchical modular structure of the gene regulatory network. The gene regulatory network of sea urchin endomesoderm specification up to 30 hours is composed of about sixty genes. This network is relatively modular. For example, as measured by the commonly used Newman modularity measure [86], defined as the fraction of edges that lie within modules rather than between modules relative to that expected by chance, the modularity of this gene regulatory network is 0.49. This modularity value greater than zero indicates that this network is quite modular.

Davidson and Erwin found that the gene regulatory network can be described by a hierarchy with four types of modules. The first type is named "kernel." For example, the endoderm specification kernel is composed of five genes in sea urchins, see Fig. 3.1. The heart-field specification kernel [103, 30] is used in both *Drosophila* and vertebrate development. The other three types are named as "plug-ins," "I/O switches," and "batteries." Each type of module functions differently in the development of embryo. The kernels might relate to the phylum- and superphylum- level characteristics; the plug-ins and I/Os might relate to the class, order, and family characteristics; and the batteries might relate to the speciation characteristics (see Fig. 3.2). This proposal stimulated debate [29, 32]. For example, the diverse kinds of changes in the hierarchy of gene regulatory networks and their evolutionary consequences are thought to be imperfect, and yet all essential major phylum-level body plans appeared at the early Cambrian. Davidson and Erwin stated that "Critically, these kernels would have formed through the same processes of evolution as affect the other components, but once formed and operating to specify particular body parts, they would have become refractory to subsequent change."

If this theory is correct, we would expect the evolution of the gene regulatory network to be heterogeneous. The "kernels" module should evolve more slowly than other parts of the gene regulatory network, since the phylum- and superphylum-level body plan characteristics have not changed substantially since the early Cambrian. The gene regulatory networks are primarily composed of two elements: transcription factors and cis-regulatory modules. Transcription factors are proteins that can either activate or repress transcription by binding to cis-regulatory elements. Transcription factor binding sites are often organized into clusters named cis-regulatory modules, which typically span a few hundred nucleotides and can contain dozens of binding sites for several transcription factors [26]. A full understanding of the evolution of the gene regulatory network would consider both transcription factors and cis-regulatory modules. Cis-regulatory modules are poorly conserved during evolution, and even in closely related species may differ drastically [26, 153]. Because experimental identification of cis-regulatory elements is still not well developed, and because computational prediction of cis-regulatory elements is still difficult [41], we will here consider only evolution of the transcription factors. Transcription factors are more conserved and evolve more slowly than cis-regulatory elements. On the timescale of hundreds of millions of years that we consider here, it is important to consider the evolution of transcription factor networks. For example, acquisition of an extra repressive regulatory domain in the insect protein *Ubx* results in the prevention of the development of abdominal legs [99]. Although transcription factors change slowly, their effect on the development body plan is of equal importance to that of cis-regulatory elements, since the variation of transcription factors directly changes the topology of the gene regulatory network. To test the theory of Davidson and Erwin, we calculated the speed of evolution of regulatory genes. We found that those genes which determine the phylum and superphylum characters evolve slowly, while those genes which determine the classes, families, and speciation evolve more rapidly. These observations provide support for Davidson and Erwin's theory.

3.2 Materials and Methods

3.2.1 Sea urchin gene regulatory network

The sea urchin is a traditional model organism in developmental biology. The sea urchin (*Strongylocentrotus purpuratus*) and sea star (*Asterina miniata*) are in the same phylum Echinodermata. The sea urchin is in the class Echinoidea, and the sea star is in the class Asteroidea. The last common ancestor of echinoid and asteroid existed about 0.5 billion years ago in the late Cambrian [109, 21, 125]. Other sea urchins used in this study are *Hemicentrotus pulcherrimus, Paracentrotus lividus, Heliocidaris erythrogramma, Heliocidaris tuberculata, Lytechinus variegatus.* They are in the same superorder Echinacea with *Strongylocentrotus purpuratus.* The genome of the sea urchin *Strongylocentrotus* *purpuratus* was sequenced in 2006. The gene sequences used in this paper were all downloaded from the EMBL-EBI database in July 2008. The experimentally determined sea urchin gene regulatory network was downloaded from http://sugp.caltech.edu/endomes/.

3.2.2 Ratio of nonsynonymous substitution to synonymous substitution of genes

We used the widely-accepted ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions (dN/dS) as a measure of the rate of evolution. Generally, dS is a measure of evolutionary divergence between two genes due to neutral substitution, and the dN/dS is the departure from the neutral substitution caused by functional constraints and selection. The larger the dN/dS value, the faster the is gene evolving due to selection. We used a standard method to calculate dN/dS [69]. First, we obtain the genes in the gene regulatory networks of sea urchin endomesoderm, see Fig. 3.1. We applied Wu-BLAST2 to search the orthologous genes in the 7 genomes mentioned above from EMBL-EBI. All the protein pairs are required to be reciprocal

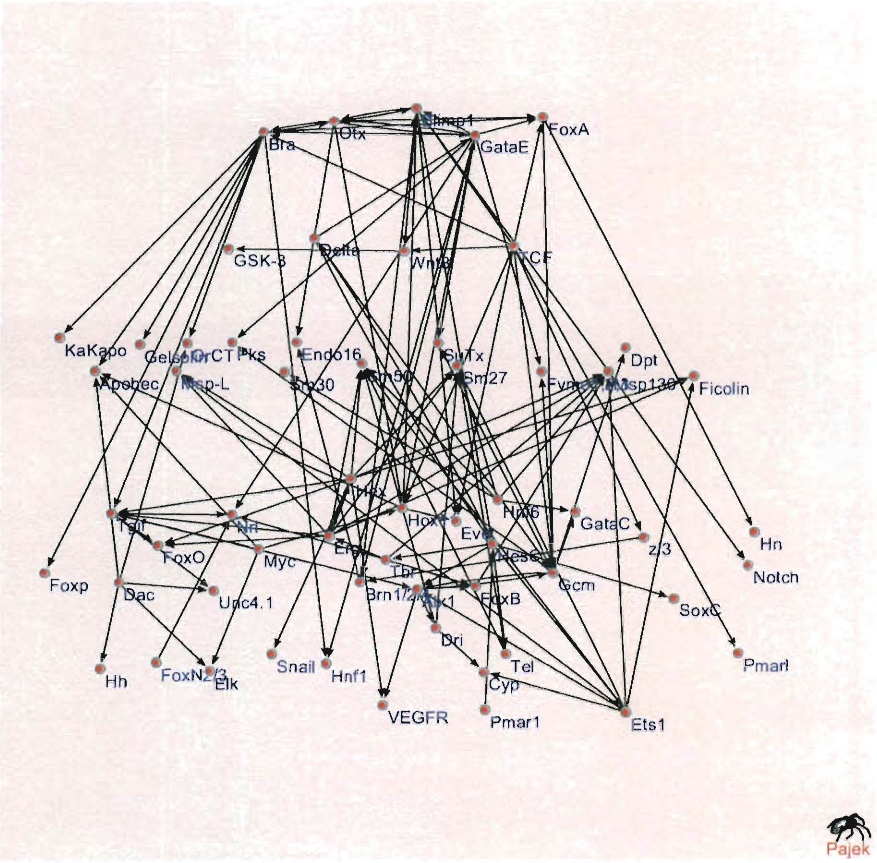


Figure 3.1 The gene regulatory network of sea urchin endomesoderm specification up to 30 hours. The top five genes form the kernel.

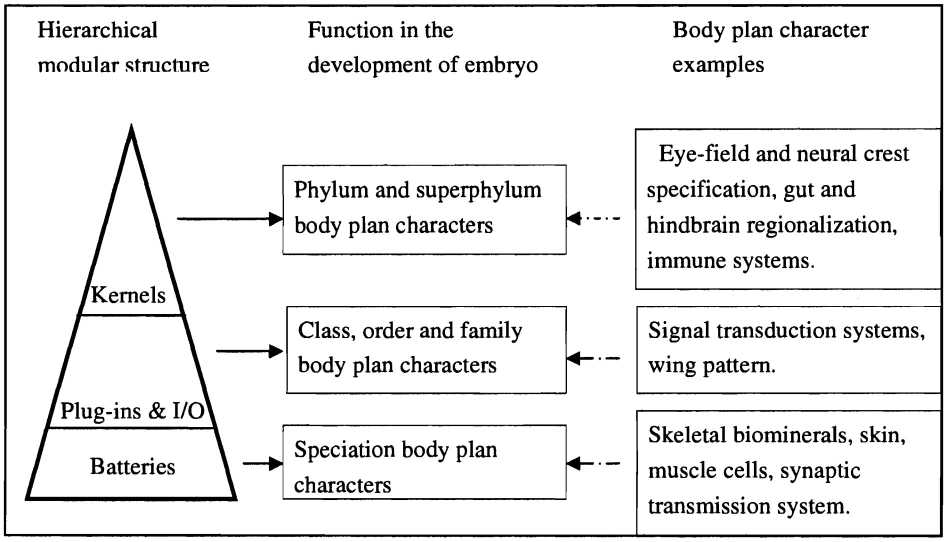

Hierarchical modular structure
Function in the development of embryo
Body plan character examples
Kernels
Plug-ins & I/O
Batteries
Phylum and superphylum body plan characters
Class, order and family body plan characters
Speciation body plan characters
Eye-field and neural crest specification, gut and hindbrain regionalization, immune systems.
Signal transduction systems, wing pattern.
Skeletal biominerals, skin, muscle cells, synaptic transmission system.

Figure 3.2 The hierarchy of the gene regulatory network and functions at different levels of development of the body plan.

best hits. Since these organisms are closely related, all orthologous protein have the same name and likely perform similar functions. Then, the orthologous protein pairs are aligned in ClustalW [117] and calculated the dN/dS in PAL2NAL [114]. For each gene, the dN/dS of all the protein pairs of that gene are averaged. Take gene *Bra* as an example, *Bra* genes are found in sea urchins *Hemicentrotus pulcherrimus*, *Paracentrotus lividus*, *Lytechinus variegatus* and *Strongylocentrotus purpuratus. Bra* should also appear in other sea urchins not yet sequenced. The *Bra* protein sequences are aligned in each two sea urchins, 6 pairs in total. For each pair, use PAL2NAL to calculate dN/dS, see Table 3.1.

3.3 Results

3.3.1 Evolutionary rate in hierarchy

The dN/dS for genes in gene regulatory networks are listed in Table 3.1. The gene regulatory network is composed of transcription factor (TF) and non-TF proteins. Most TF genes are utilized for diverse interactions, and the DNA binding domains of all of them are highly conserved across Bilateria. We average the dN/dS of all proteins in each hierarchical level, and the result is shown in Fig. 3.3. The value of dN/dS of kernels is significantly lower than plug-ins, I/Os and batteries, see Table 3.2 for P-value. Also, the regulatory gene group of kernels and plug-ins has a lower dN/dS value than group of I/Os and batteries (relative difference $= -0.055$, P-value $= 0.0157$ for Wilcoxon test). From the probability distribution of dN/dS in Figure 3.5, the distribution of kernels is narrow width, and the peak probability appears at a low dN/dS.

If only TF genes are considered, kernels ($dN/dS = 0.045$) still evolve more slowly than other components. Slight increase of dN/dS from plug-ins to I/Os and to batteries is also observed. ($dN/dS = 0.138$). Interestingly, the number of organisms for which an ortholog was detected varies from genes to genes. For example, *A. miniata* is the least close organism to *S. purpuratus* compared to other sea urchins, *S. purpuratus* and *Asterina miniata* are in the same phylum but different orders. Four orthologous genes between *A. miniata* and *S. purpuratus.* Three of them are kernel genes. The orthologs

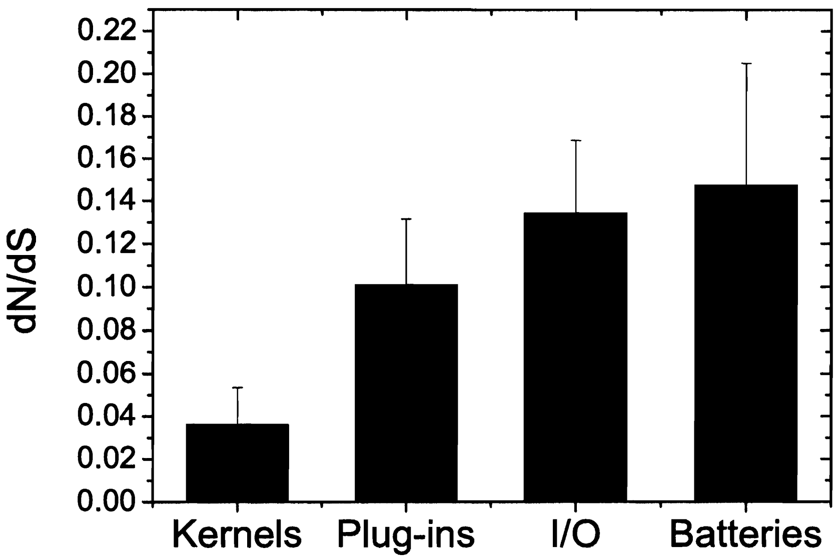


Figure 3.3 The ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions for different components of the gene regulatory networks that control the development of animal embryos.

of kernel genes are more likely detected than other genes in far related organism, which is a support of slower evolution of kernels.

The results show that if two organisms are in the same phylum, their kernel modules which determine the phylum level body plan are conserved. If two organisms are in the same class or order, their plug-ins and I/O modules are conserved since they determine the class and order level body plan.

3.3.2 Generative entrenchment

Another supporting evidence comes from the "generative entrenchment" theory by William Wimsatt [131, 132]. In the this model, the phenotype is considered as a generative structure. The generative structure of the system has a characteristic set of causal interactions which can be represented by the directed graph, see Fig. 3.1. In this model, nodes with more downstream connections should have slower evolution rates, since changes to them affect many epistatic interactions that must be accommodated during the evolution. Quantitatively, we account the downstream connections for each gene in gene regulatory networks. For example, if gene A regulates the expression of gene B, we say gene A has a downstream connection. We observe that “kernels” genes in Fig. 3.1 have an average of 5.4 downstream connections, while the other regulatory genes have an average of 1.6 downstream connections. All else being equal, nodes with more downstream connections such as kernels here should be more conservative, because their activity can bring more consequences. If they are changed, it is more likely that something will go wrong. One model to explain this idea was proposed by Rupert Riedl [98]. In this model, Riedl raised the idea of “burden” which states that the evolvability of a character change during evolution depends on the importance of the functions and structures depending on it. The “kernels” of the gene regulatory networks which determine the phylo-level body plan are thought to have “heavier” burden than other parts of gene regulatory works, since it is the base of animal body plan. So “kernels” are likely to be more conserved.

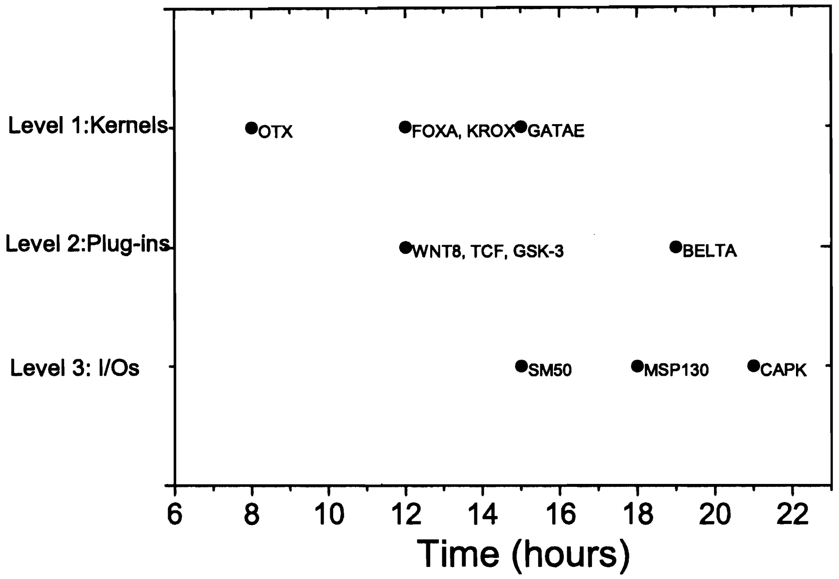


Figure 3.4 The earliest appearance time of regulatory genes for kernels, plug-ins and I/Os as experimental data available(Davidson et al., 2002).

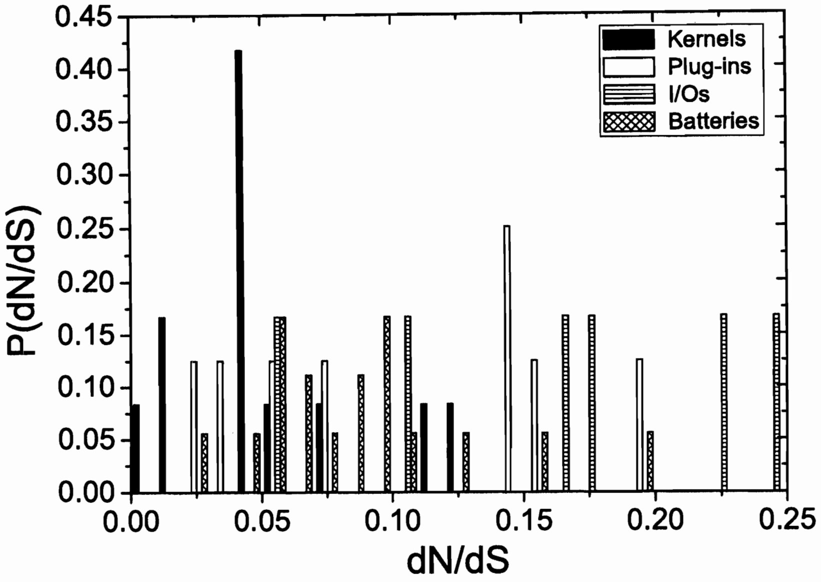


Figure 3.5 The distribution of dN/dS for each hierarchical level. P(dN/dS) is the probability of a gene with the dN/dS in specific hierarchical level.

3.3.3 Time of appearance of regulatory genes

As additional supporting evidence, the time of appearance of regulatory genes is considered during embryo development. From available experiment data, the earliest appearance time of regulatory genes is shown, defined as the time when a given gene is expressed and starts to regulate the expression of other genes. Figure 3.4 shows the endomesoderm specification up to 22 hours for genes listed in Table 3.1. The x-axis is appearance time of genes in embryo development, and the y-axis indicates in what hierarchical level genes belong to. The kernel genes generally express earlier in endomesoderm than other regulatory genes, and most plug-ins genes appear earlier than I/Os genes. The Karl Ernst von Baer's law states that "General characteristics of the group to which an embryo belongs develop before special characteristics. General structural relations are likewise formed before the most specific appear." That is: differentiation proceeds from the general to particular, with taxonomically more general parts expressed earlier in development. In this case, we can interpret as the kernels which expressed earliest in development are more related to the higher hierarchical level of taxon such as phylum- and superphylum-level body plan, while others are more likely related to lower hierarchical level body plan. Genes which are expressed earlier in development are, mostly likely, older and more likely to be conserved during evolution, because mutations of proteins expressed earlier in embryo development are more likely to have larger, more pervasive, and more deleterious effects on subsequent development [132].

3.4 Discussion

Recently, it has been found that biological networks are not random and unstructured networks, instead, many are modular networks [86]. Modular network can be decomposed into several highly interacting modules and are particularly interesting. Perturbations or errors in a modular network are typically restricted to one module, and the effect on the whole network is limited. Modular networks can evolve by rewiring the modules. This rewiring capability property tends to make modular networks more evolvable [58, 113]. A hierarchical network is an advanced modular network. In hierarchical networks, some modules are key modules that may relate to the core function and be resistant to mutation. Other modules are periphery modules that may be more likely affected by the environmental change [96]. Peripheral modules evolve rapidly and allow the organism to survive in a changing environment.

The origin of animal body plan is one of the central questions in developmental biology [13]. A long studied subject, it seems established that evolutionary rates of different characters and lineages are different. The results in Figs. 3.3 support Davidson and Erwin's theory that the hierarchical structure of the gene regulatory network has imposed constraints on the rate of further evolution of the most basic, and earliest-evolved features. The slow speed of evolution of the kernels that control the development of animal phylum- and superphylum-level body plan characteristics is why no new phylum-level body plans appeared after the pre-Cambrian period. The number of types of classes, orders, families and species is increasing, and The results show that this observation is surprisingly consistent with the increasing evolutionary speed from kernels to plug-ins to I/Os to batteries. We propose that the slow evolution of the top components and fast evolution of the bottom components of the hierarchy is a universal phenomenon in evolution, not only in the gene regulatory networks, but also in protein interaction networks, cell signaling networks, and metabolic networks [34].

GROUP	GENE	dN/dS	Org.1	Org.2	dS	dN
Kernels	FOXA	0.0083	PATVU	STRPU	52.38	0.4328
Kernels	KROX	0.0114	ASTM	STRPU	57.1692	0.6491
Kernels	OTX	0.0703	ASTM	STRPU	5.1028	0.3587
Kernels	OTX	0.11	HEMPU	STRPU	0.109	0.012
Kernels	OTX	0.122	HEMPU	ASTM	4.2	0.51
Kernels	GATAE	0.015	ASTM	STRPU	31.6839	0.4776
Kernels	BRA	0.0435	HEMPU	PARLI	0.87	0.03
Kernels	BRA	0.0408	HEMPU	LYTVA	0.997	0.407
Kernels	BRA	0.057	PARLI	LYTVA	0.897	0.051
Kernels	BRA	0.0413	HEMPU	STRPU	0.19	0.0079
Kernels	BRA	0.0456	PARLI	STRPU	0.85	0.038
Kernels	BRA	0.0493	LYTVA	STRPU	0.82	0.04
Plug-ins	WNT8	0.147	HELER	STRPU	0.7104	0.1051
Plug-ins	TCF	0.07	PARLI	STRPU	0.59	0.04
Plug-ins	TCF	0.0324	HELER	STRPU	0.4	0.013
Plug-ins	TCF	0.0224	PARLI	HELER	0.429	0.009
Plug-ins	DELTA	0.15	STRPU	PARLI	0.45	0.066
Plug-ins	DELTA	0.14	PARLI	LYTVA	0.64	0.09
Plug-ins	DELTA	0.19	STRPU	LYTVA	0.47	0.09
Plug-ins	GSK-3	0.0557	PARLI	LYTVA	0.4002	0.0223
I/Os	SM30	0.053	STRPU	HEMPU	0.106	0.005
I/Os	MSP130	0.226	HELTB	HELER	0.218	0.049
I/Os	MSP130	0.174	HELER	STRPU	0.741	0.129
I/Os	MSP130	0.167	HELTB	STRPU	0.736	0.123
I/Os	SM50	0.2436	HEMPU	STRPU	0.1891	0.046
I/Os	CAPK	0.101	STRPU	HEMPU	0.0637	0.0064
Batteries	HNF6	0.0637	ASTM	STRPU	4.5405	0.2894
Batteries	GSC	0.129	STRPU	HELTB	0.59	0.076
Batteries	GSC	0.08	LYTVA	HELER	1.12	0.1
Batteries	GSC	0.09	STRPU	LYTVA	0.903	0.082
Batteries	ALX1	0.028	STRPU	LYTVA	0.79	0.02
Batteries	ALX1	0.048	LYTVA	PARLI	1.15	0.056
Batteries	ALX1	0.068	STRPU	PARLI	0.85	0.058
Batteries	ETS	0.072	STRPU	PARLI	0.866	0.062
Batteries	KRL	0.601	STRPU	HEMPU	0.128	0.0776
Batteries	SOXB1	0.055	STRPU	HELER	0.45	0.02
Batteries	SOXB1	0.0527	HELER	HELTB	0.078	0.004
Batteries	SOXB1	0.0568	STRPU	HELTB	0.4423	0.251
Batteries	GCM	0.0911	LYTVA	PARLI	2.61	0.23
Batteries	GCM	0.0951	PARLI	STRPU	0.963	0.0916
Batteries	GCM	0.1952	LYTVA	STRPU	1.1161	0.2179
Batteries	HOX11	0.1513	HELTB	HELER	0.0639	0.009
Batteries	HOX11	0.1082	HELER	STRPU	0.35	0.038
Batteries	HOX11	0.0888	HELTB	STRPU	0.386	0.0343

Table 3.1 Evotionary rate of regularoty genes in pairs of organisms (Org.1 and Org. 2). Group 1 are kernels, group 2 are plug-ins, group 3 are I/O, group 4 are batteries. STRPU: *Strongylocentrotus purpuratus*, PATVU: *Patella vulgata*, ASTM: *Asterina miniata*, HEMPU: *Hemicentrotus pulcherrimus*, PARLI: *Paracentrotus lividus*, HELER: *Heliocidaris erythrogramma*, HELTB: *Heliocidaris tuberculata*, LYTVA: *Lytechinus variegatus*.

Hierarchy	Hierarchy	P-value
Kernels	Plug-ins	0.0826
Kernels	I/Os	0.0032
Kernels	Batteries	0.0092
Kernels	Plug-ins + I/Os + Batteries	0.0026
Kernels + Plug-ins	I/Os+Batteries	0.0157

Table 3.2 P-value of Wilcoxon test for different hierarchical levels. The hypothesis for the Wilcoxon test is that two independent samples come from distributions with the same median.

Chapter 4

Structure and Response in the World Trade Network

4.1 Introduction

Physical theory of evolution predicts that under certain conditions, a changing environment leads to development of modular structure [113, 63, 76]. The prediction depends only on 1) the dynamics of the response to change being "slow" due to a glassy landscape, 2) presence of change, and 3) exchange of information between evolving agents. Since the trade network is an evolving system, this physics of evolution may be applied to the world trade system, previously studied by network analysis [67, 16]. We assume that condition 1 is satisfied for the world trade network due to the complexities of inter-country relationships. Condition 2 is satisfied by viewing recessions as causing a change of the environment for the dynamics of the world trade system. Condition 3 is satisfied because information flow naturally results from transfer of business practices or material between countries. Thus, the theory of [113, 63] allows us to make three predictions: decreased modular structure in the world trade network increases the sensitivity to recessionary shocks, decreased modular structure decreases the rate of recovery, and recessions themselves spontaneously increase modular structure of the world trade network. All three predictions will be borne out by data. These results are general predictions about how the detailed structural parameters of the evolving economic system will organize. Theory shows that the modular and hierarchical structure formed in response to environmental fluctuation increases the resistance to and rate of recovery from perturbations. The theory predicts that globalization, which reduces hierarchical structure, should lead to increasingly large recessions and decreased rate of recovery, in contrast to standard economic understanding [8].

To apply the physical theory of evolution that describes the spontaneous emergence of modularity in fluctuating environments [113, 63] to world trade, we seek a mathematical representation of hierarchy in the world trade network. Identification of network motifs or modules is an active research field in the physics of networks [113, 86, 9], with the study of structure at multiple scales, i.e. hierarchy, somewhat more recent [96, 27]. In this chapter, we treat the world trade data as defining a geometry in trade space. We project the trade topology onto the best tree-like topology representing the data. The success of this projection in representing the original geometry is used to define the hierarchy of the original data.

We apply hierarchical clustering to construct the best tree-like representation of the world trade network. Correlation between the distances implied by the tree construction and the distances defined by the original trade data is calculated. This quantity is termed the cophenetic correlation coefficient (CCC) [43]. We will display the general trend of the CCC since 1969, noting especially the increase of the CCC after each recession. The magnitude of the CCC will be shown to correlate with the ability of total world GDP to resist a recessionary shock. Theory shows that this result is, in fact, causal, not simply a correlation, which is a major result here.

We focus on how global recessions, such as the 2008–2009 recession, have affected the structure of the world trade network. Modular structures arises in the trade network, for example, because countries in a trade group trade among themselves to a greater extent than with others. These trade groups may interact with each other to form higher level groupings. The detailed reasons for an increase of hierarchy in the world trade network are many: perhaps protectionism for the domestic economy [10], or because long-distance trade seems costly during a recession. Standard arguments in economic theory suggest a decreased rate of recovery from recession for trade networks with more modular structure [8]. We will see, however, that theory predicts that greater trade network structure increases both the resistance to recessionary shocks and the rate of recovery from recessions.

4.2 Results

4.2.1 Measure of hierarchy by CCC

A hierarchical trade network occurs when countries with strong trade connections group into trade modules or regional trade clusters. A flat or non-hierarchical structure

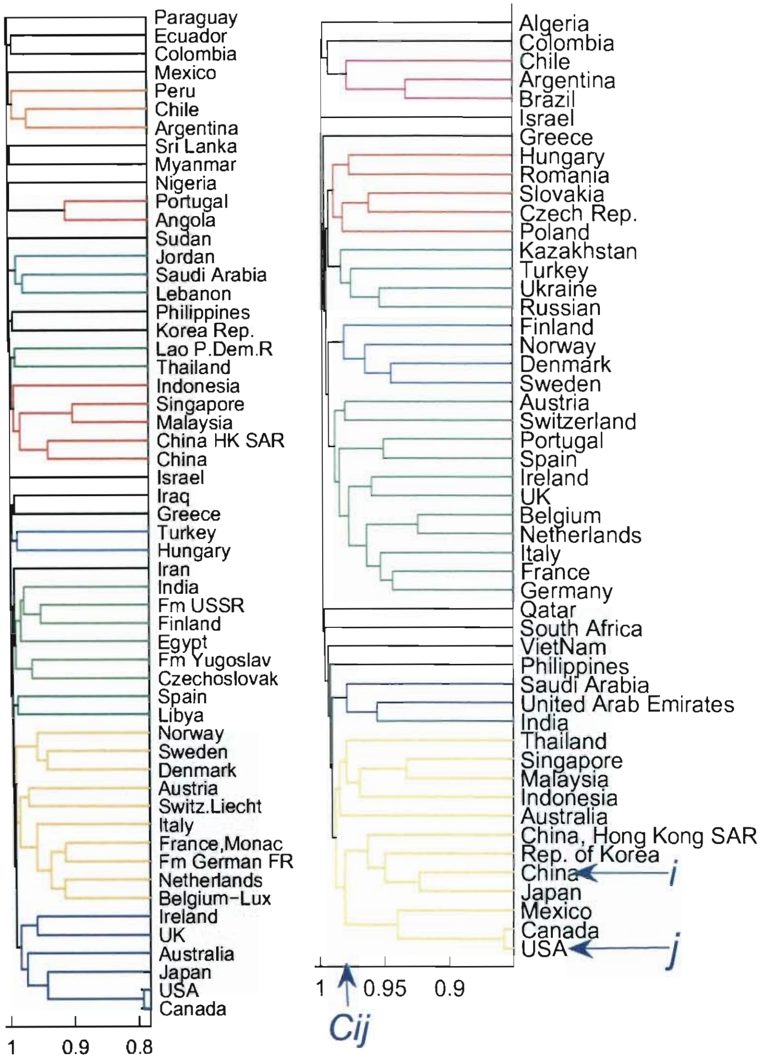


Figure 4.1 Dendrogram representation of trade networks for selected countries at 1969 (top figure) and 2007 (bottom figure).

occurs when countries trade evenly with all other countries, and there are no regional trade modules in the trade network. We use the historical trade data from United Nation database (Comtrade) from 1962 to 2007. We build the world trade network with nodes representing countries and links representing the trade value. We do not scale the trade volume by the GDP, because small economic units should not have the same weight as large economic units. First, a distance matrix is calculated from the trade network matrix by $d_{ij} = M^* - M_{ij}$, where $M^* = \max(M_{ij})$. Here, M_{ij} is trade value between two countries. The average linkage hierarchical clustering algorithm is applied to the distance matrix to produce the tree-like dendrogram [43], see Fig. 4.1. In this figure, trade modules are marked by different colors. Only selected countries are plotted, because the figure becomes crowded if all countries are plotted. We define the tree-like structure to have the most hierarchy. Therefore, the amount of hierarchy can be measured by the likeness between the original data and the best tree that is produced from original data by hierarchical clustering. The CCC quantifies this likeness. The cophenetic matrix is generated from the dendrogram. Its elements are the branch distance where two objects become members of the same cluster in the dendrogram: for two nodes, ij, the nearest common bifurcation point is located, and the branch length for this point is the cophenetic element of these two nodes, c_{ij}, see Fig. 4.1 for an example. The CCC is defined as $CCC = [\sum_{i<j}(d_{ij} - \overline{d})(c_{ij} - \overline{c})]/\sqrt{[\sum_{i<j}(d_{ij} - \overline{d})^2][\sum_{i<j}(c_{ij} - \overline{c})^2]}$, where d_{ij} and $\overline{d}$ are the element and average of elements of the distance matrix, and c_{ij} and $\overline{c}$ are the elements and average of elements of cophenetic matrix, respectively. Hierarchical datasets have a high CCC value, and nonhierarchical datasets have a low CCC value [74].

4.2.2 Evolution of structure in world trade

A major factor affecting the world trade network over the last 40 years has been the process of globalization. Qualitatively, this globalization has been expressed as a "flattening" of the world [50]. Here, we use the CCC to measure how the hierarchical structure

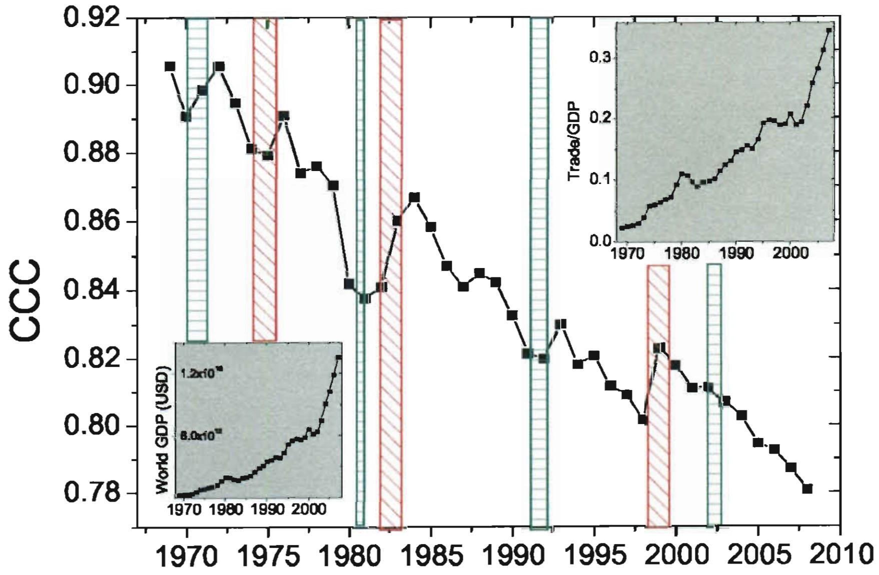


Figure 4.2 The CCC from 1969 to 2007. The upper right insert is the ratio of total world trade to world GDP. The lower left insert is the total world trade in units of US dollar.

of the world trade network has changed over time. Large CCC values indicate higher hierarchy. The major trend of CCC with time in Fig. 4.2 is a reduction of hierarchy as the “flattening” has taken place [50]. In Fig. 4.2, shaded rectangles marked the seven recessions. Left and right borders are positioned at the start and end of a recession, respectively, according to US National Bureau of Economic Research. We notice, however, that the CCC does not always decrease year by year. We notice that during and after each recession, marked on the figure, the CCC value increases. The CCC values at the year after recession are larger than that at the year before the recession (p-value $= 0.003$ of Kolmogorov-Smirnov test for null hypothesis that they are from the same distribution with the same mean, and p-value $= 0.0006$ for null hypothesis that CCC value before recession is larger than that after recession). This trend is true both for the past 3 major recessions and for the past 4 minor recessions. The scale of increase of hierarchical structure depends on the severity of recession. One possible reason for this CCC trend during recessions is the increase of trade protectionism during recessions. Also, regional integrations are greatly enhanced during recessions, leading to increased regional imports [40], which strengthens trade modules. Free trade promotes globalization and decreases the hierarchy of the trade networks. But trade protectionism and regional integration, which is common during recessions to protect domestic or regional economies by restraining trade between countries, tends to reduce trade between countries in different trade modules. Thus, recessions may promote the regionalization that enhances the modularity of the trade network. One example is the Asian currency crisis of 1997, which lead to the development of independent Asian monetary systems.

The CCC is a characterization of the world trade network that is independent from the total amount of world trade. In the process of globalization, a country tends not only to increase its total trade value, but also to trade with more partners. The upper right insert of Fig. 4.2 shows the typically increasing ratio of world trade to GDP. Only the recessions of 1981, 1991, 1997 and 2001 lead to a decrease in the trade to GDP ratio, whereas the CCC increased in all seven recessions. The increased hierarchical structure appearing

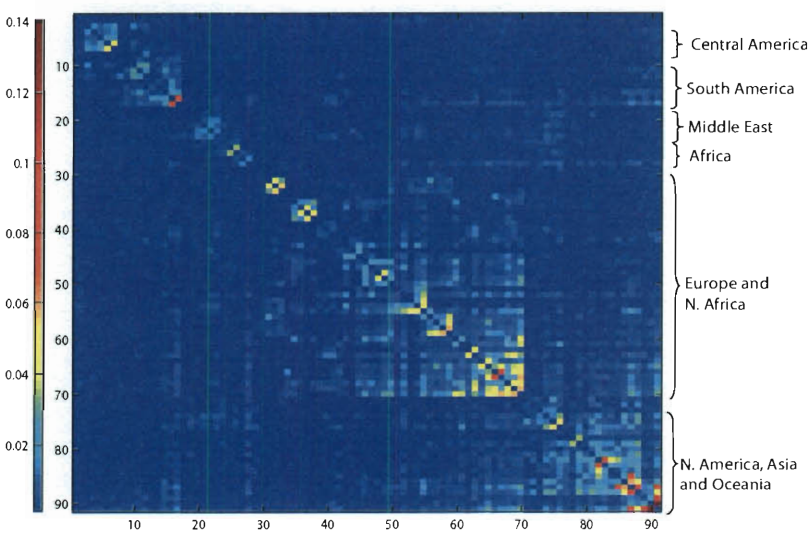


Figure 4.3 The trade share matrix $S_{ij} = M_{ij} / \left(\sum_{m=1}^{N} M_{im} + \sum_{n=1}^{N} M_{jn} \right)$ after hierarchical clustering between countries in 2007.

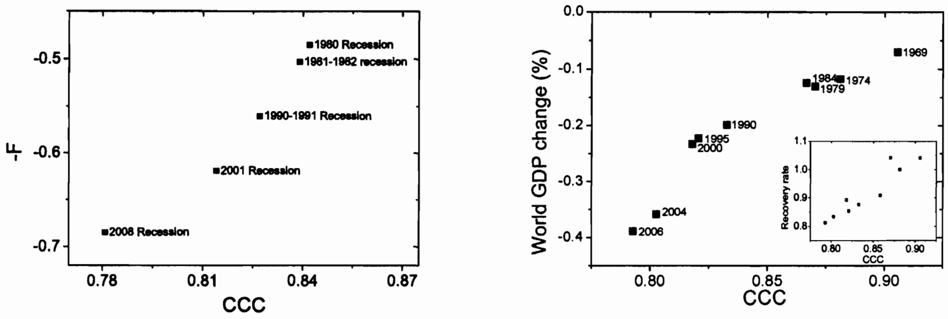
after all seven recessions in Fig. 4.2, is therefore, a sensitive correlate of recessions, and independent of the trade to GDP ratio shown in the upper right insert of Fig. 4.2 and total trade volume shown in the lower left insert of Fig. 4.2. Measurement of globalization by both hierarchical structure (CCC) and total trade provides complementary information.

The CCC quantifies the development of hierarchical structure in the trade network at multiple scales in an integrated way. The clustering of the world trade network shows the modularity of global trade, see Fig. 4.3. Several modules can be observed: North American+Asian+Oceanic countries, European+ North African countries, Middle and South Africa, Middle East, South America, and Central America. The development of regional trading partners occurs simultaneously with globalization. By comparing the structure of trade network in 1969 and in 2007, we found that the increased trade among Canada, United States, and Mexico as a result of NAFTA is one example of a regional trading group. Regional trade pacts among the Middle East countries are other examples of regionalization. In general, free trade markets will develop modular structure at multiple geographical scales.

4.2.3 Response of world trade to recessions

The ability of the trade system to respond to recessionary perturbations is proportional to the hierarchical structure present, i.e. increases with the CCC value, according to the evolutionary theory of modular structure development [113, 63]. That is, the modular structure that exists at multiple scales affects how recessions propagate in the trade network, just as modular structure of person-to-person contacts affects how diseases spread in a population. We examine how the network structure affects the propagation of a recession throughout the world. For example, if there is a one percentage decrease of the GDP of the USA, by how much does the total GDP of world excluding the USA decrease due to the spread of recession from the USA? We investigate the five most recent global recessions including the 2007-2009 crisis. We calculate the ratio of GDP change (percentage) of world excluding the USA to the GDP change (percentage) of USA in each recession as a function of the CCC value in each recession, see Fig. 4.4(a). In Fig. 4.4(a), the GDP change in one recession is defined as the GDP decline from peak to trough. Quarterly GDP data are used to find the peak of recession. The quarterly GDP data for 2001 and 2008 recessions are from *OECD Stat. Extracts* (http://stats.oecd.org/). The quarterly GDP data for earlier recessions are estimated from annual GDP data. We observe that in more recent recessions with less hierarchical structure of trade network, a recession in the USA has a stronger impact on the rest of the world. This result indicates a strong positive correlation between lack of hierarchical structure and severity of recession impact.

We also perform an impulse response analysis of the vector autoregression (VAR) model to analyze the time evolution of recession [24, 78, 7]. We explore the possible underlining causal links between lack of hierarchical structure and severity of recession. A recession is assumed to start in the USA. The US GDP is initially reduced by the maximal GDP decline during the recession, e.g. the maximal quarterly US GDP decline



(a)

(b)

Figure 4.4 (a) The ratio of the total world excluding the USA GDP change (percentage) to the change of the USA GDP (percentage) in 5 recessions, F in the y axis. (b) Impulse response analysis of spread of recession. The world GDP change is plotted as a function of the CCC. The reduction in the world GDP is greater when the CCC value is low. Insert figure: The GDP recovery from recession can be well fit by the relation $Y(t) \sim Y(\infty) - a\exp(-\lambda t)$. Yearly recovery rates, λ, are shown versus the CCC. In accord with theory, the recovery rate is positively correlated with the CCC.

was 5.4% S in the 2008–2009 recession [6]. The export from country i to country j, X_{ij}, is updated by a factor of ratio of GDP of country j at time t, $Y_j(t)$, and $t-1$, $Y_j(t-1)$) [7]. Thus, $X_{ij}(t) = X_{ij}(t-1)Y_j(t)/Y_j(t-1)$. Then the GDP of country i is updated by $Y_i(t+1) = Y_i(t) + P_i(X_i(t)/X_i(t-1) - 1)$. Where $P_i = X_i/Y_i$ is the ratio of export to GDP for country i. The GDP of each country decreases until steady state is reached, at which point the simulation is terminated. We calculate the world GDP change as $(\sum_i Y_i^{\text{steady}} - \sum_i Y_i^{\text{initial}})/\sum_i Y_i^{\text{initial}}$. We observe how the crisis spreads globally and measured the GDP loss during crisis.

The impulse response analysis results support that the severity of the 2008-2009 recession may be due to loss of hierarchical structure in the global trade network. Lack of hierarchical structure makes the world trade network less resistant to recession, as observed from Fig. 4.4(b). In the simulation as shown in Fig. 4.4(b), initial values are set to historical trade and GDP data in each year. A recession is assumed to begin in the USA and spread to the rest of the world. We believe this increased sensitivity is due to a loss of modular or hierarchical structure in the world trade network, see Fig. 4.2. As an example, the impact of a recession on the GDP is more severe in 2006 than in 1968, by a factor of 5.7. Interestingly, after this calculation was carried out, an estimate of the ratio of the reduction of GDP in 2009 to the average reduction over past recessions equaling 6 was reported [6].

Evolutionary theory has shown that systems under environmental perturbation not only increase their modularity, but also increase their response function to perturbations [113, 63]. In the present context, this would imply that as trade has been globalized, and the CCC reduced, the rate of recovery from recession should decrease. We consider this phenomenon in the world trade network, using the VAR model. After the system reaches steady state following the reduction to the USA GDP, we impose a positive impulse to restore the USA GDP to its initial value. The world GDP recovers, at a rate that depends on the hierarchical structure of the trade network. We observe that when the trade network has greater hierarchical structure, indicated by a larger CCC value, the trade network recovers more quickly from recession, as shown in the insert figure of Fig. 4.4(b).

4.3 Conclusion

We have used the concept of viewing the world trade network as defining a geometry in trade space and the idea of projecting this geometry to the best tree-like topology to define the hierarchy in the world trade network. With that necessary mathematical prolegomena, we introduced the world trade network as an evolving system. Physics of evolution in changing environments was then used to predict that the world trade network is more sensitive and recovers more slowly from evolutionary shocks now than it did 40 years ago, because globalization has reduced hierarchical structure in the world trade network. We also predict that recession-induced change to the world trade network should lead to a temporarily *increased* hierarchical structure of the global trade network. These predictions, contrary to standard economic thinking, were born out by our study of the world trade data since 1969.

Part II

Influenza virus evolution

Chapter 5

Prediction of incipient dominant influenza strain by clustering

5.1 Introduction

The recent outbreak of 2009 A(H1N1) caused immediate international attention[33, 52, 49, 111]. This new 2009 A(H1N1) virus contains a combination of gene segments from swine and human influenza viruses[52, 49]. Confirmed infections reached 270,000 globally as of September 2009[149]. The novel 2009 A(H1N1) strain was defined as a pandemic strain by the World health Organization(WHO) in 2009[152], and was the epidemic strain in the 2009 Northern winter.

Influenza viruses are hyper-mutating viruses. It has been estimated that the nucleotide mutation rate per genome per replication is approximately 0.76[37]. Influenza viruses escape the human immune system by continual antigenic drift and shift[48, 129, 54, 47, 53, 85]. The quasispecies nature of influenza viruses makes the strain structure complex[36]. Usually, there is one or a few dominant influenza strains circulating in the population for each flu season. The flu vaccine is most effective when it matches this dominant circulating strain[56, 55]. The degree to which immunity induced by a vaccine protects against a different viral strain is determined by the antigenic distance between the vaccine and the virus. Due to evolution of the antigenic regions of the influenza virus, the composition of the flu vaccine is typically modified annually[100]. However, since the influenza strains used in the flu vaccine are decided 6 months before the flu season, a mismatch between the vaccine strain and dominant circulating strain may occur if the virus evolves significantly. Such a situation arose for the H3N2 virus in the 2009–2010 flu season, when A/British Columbia/RV1222/2009 emerged in the early spring[108, 2]. Accurate early prediction of the dominant circulating strain is an essential and important task in influenza research.

There are several ways to estimate the flu vaccine effectiveness. Gupta *et al.*[55] proposed p_{epitope} as a measure of antigenic distance between influenza A vaccine and circulating strains. The hemagglutinin protein has five epitopes. The dominant epitope for a particular circulating strain in a particular season was taken as that which had the largest fractional change in amino acid sequence relative to the vaccine strain. The value of p_{epitope} is defined as the fraction of number of amino acid differences in the dominant epitope to total number of amino acids in the dominant epitope. The antigenic distance between the vaccine strain and the circulating strain is quantified by p_{epitope}. By a metaanalysis of historical vaccine efficacy data from over 50 publications, Gupta *et al.* showed in a metaanalysis that the p_{epitope} between vaccine strain and circulating strain correlates well with the vaccine efficacy, with $R^2 > 0.8$[55].

Understanding the evolution of influenza viruses has benefited from phylogenetic reconstructions of the hemagglutinin protein evolution[101, 47]. In an alternative approach, Lapedes and Farber[79], followed by Smith *et al.*[110], applied a technique called multidimensional scaling to study antigenic evolution of influenza. Plotkin *et al.* clustered hemagglutinin protein sequences using the single-linkage clustering algorithm and found that influenza viruses group into clusters[94].

Here, we present a low-dimensional clustering method that can detect the cluster containing an incipient dominant strain for an upcoming flu season before the strain becomes dominant. The method builds upon the dimensional projection technique used by Lapedes and Farber[79] and Smith *et al.*[110] to characterize hemagglutination inhibition data. Importantly, the present method requires only sequence data, unlike the approach of Lapedes and Farber[79] and Smith *et al.*[110], which require ferret hemagglutination inhibition assay data. In this paper, we first study the evolution of 2009 A(H1N1) by an evolutionary path map which leads to a suggestion for the H1N1 vaccine strain. Then, we introduce the low-dimensional protein sequence clustering method. We propose an influenza vaccine selection procedure based on this sequence clustering. The procedure is demonstrated and tested in detail using historical data. We show the performance of the method to predict the dominant H3N2 strain in an upcoming flu season using data solely from before the flu season, on data since 1996. We compare the results to those from existing methods since 1996. In the discussion section, we discuss the relationship between the protein sequence clustering method and previous approaches. We discuss the false positive rate, as well as other challenges.

5.2 Results

5.2.1 Evolutionary path of 2009 A(H1N1) influenza

We first construct the directional evolutionary path for the 2009 A(H1N1) influenza. We use high resolution data in sequence, time, and world spatial coordinate to construct this evolutionary relationship. Since its first detection, the 2009 A(H1N1) virus has been extensively sequenced[52, 49]. By May 1, 2009, the number of confirmed cases reported by WHO was 333[149]. At the same time, the sequenced hemagglutinin protein (HA) available in NCBI Influenza Resources Database were 312[14]; that is to say most of the confirmed cases at that time were sequenced. At July 1, 2009, the ratio of sequenced HA protein to confirmed cases by WHO was 1039/77201[149], a number which is still much larger than that for seasonal flu. In addition, the Influenza Resources Database contains the date of collection of each 2009 A(H1N1) virus strain. We reconstruct the evolutionary history of swine flu viruses with the following procedure. If strain B is mutated from strain A, we term strain A "founder" and strain B "F1" We align the HA proteins of all 2009 A(H1N1) strains. Then, for each strain, we find its founder strain based on the following four criteria: 1, the founder strain should appear earlier than the strain, as judged by collection date; 2, the founder strain should have only one amino acid difference in the HA1 protein relative to the F1 strain; 3, the founder should also have the most similar nucleotide sequence relative to F1; and 4, the founder strain should have a large number of identical copies circulating in human population, as approximated by the number of different strains with identical HA sequences in the Influenza Resources Database. By applying these four criteria to 2009 A(H1N1) influenza, we construct the directional evolutionary path map, as shown in Fig. 5.1. In this figure, several strains are notated (Strain #1: A/California/05/2009; Strain #2: A/California/04/2009; Strain #7: A/California/07/2009; Strain #12: A/Texas/05/2009; Strain #28: A/New

York/19/2009). Strains from the Northern and Southern hemisphere are shown as red dots and blue dots respectively. One branch represents one substitution in the amino acid sequence. Two clusters are observed in Fig. 5.1 : one around A/New York/19/2009 (#28), and another one around A/Texas/05/2009 (#12). Most new strains are from the Northern hemisphere, and strains from the Southern hemisphere are mainly located at the edge of the map, such as strain #96, #120, and #126. That the Southern hemisphere strains appear at the boundary of the figure provides a self-consistency check of the validity of the assumptions entering the construction of this figure. Geographically, we see many founder to F1 links are from US and Mexico to other countries, but we rarely see founder to F1 links that are from other countries to US and Mexico, or from other countries to other countries except US and Mexico (see Materials and Methods). We also found that strains with more F1 in Fig. 5.1 are more frequently seen in the human population. For example, in the Influenza Resources Database, we found 153 strains to be identical with A/New York/19/2009, which has 29 F1 strains, and 120 strains to be identical with A/Texas/05/2009, which has 24 F1 strains. We can see in Fig. 5.1 that A/Texas/05/2009 is at the very upstream of the map, with downward connections to most of the other strains by direct or two-step links. This result agrees with the US Food and Drug Administration[44] recommendation of A/Texas/05/2009 as a vaccination strain. The alternative vaccine strain A/California/7/2009 (#7) has fewer F1 strains and it is not located at the center of the network.

5.2.2 Low-dimensional clustering

We use a low-dimensional clustering method to visualize the antigenic distance matrix of the viruses. We use a statistical tool called "multidimensional scaling"[43]. This method was used by Lapedes and Farber[79] and Smith *et al.*[110] to project ferret hemagglutination inhibition assay data to low dimensions. The influenza viral surface glycoprotein hemagglutinin is a primary target of the protective immune response. Here we project the hemagglutinin protein sequence data, rather than animal model data, to

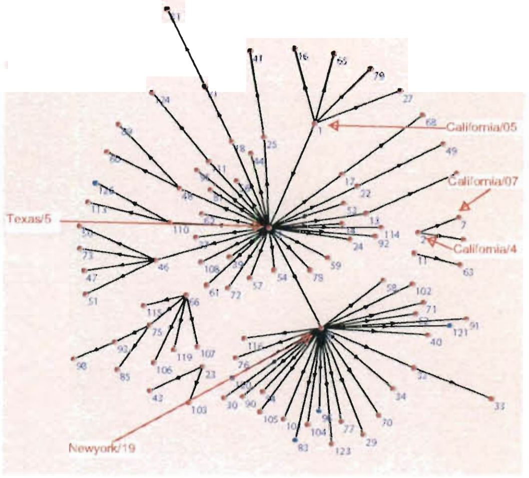


Figure 5.1 The evolutionary path of 2009 A(H1N1) influenza.

low dimensions. The HA1 protein of influenza with 329 residues can be considered as a 329-dimension space. The multidimensional scaling method is applied to rescale the 329-dimension space to a 2-dimensional space, so that we can plot and visualize it. First, we do a multialignment of the HA1 proteins. Then, the distance between any two proteins is calculated as

$$d_{ij} = \frac{1}{N} \sum_{m=1}^{N} (1 - \delta_{s_{i,m}, s_{j,m}}) \tag{5.1}$$

where $s_{i,m}$ is the amino acid of protein i at position m. The term $\delta_{s_{i,m}, s_{j,m}}$ is 1 if amino acids of protein i and j at position m are the same. Otherwise, it is 0. For the 2009 H1N1 viruses, we consider the entire HA protein, and $N = 566$. For H3N2 viruses, we consider only the HA1 protein, and $N = 329$, because the entire HA proteins are not completely sequenced in many cases. Thus, d_{ij} is the number of amino acid differences between HA proteins normalized by length. The multidimensional scaling produces a protein distance map, for example, Fig. 5.2(b). In this map, each data point represents a flu strain isolate.

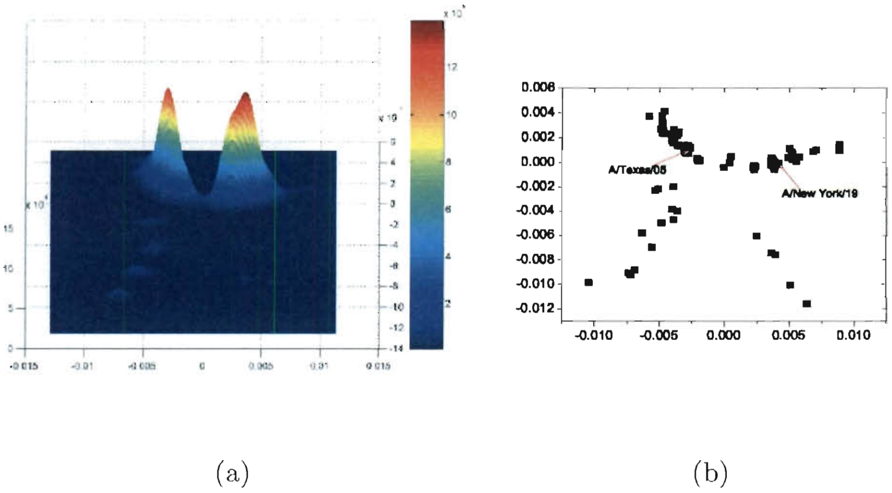


Figure 5.2 (a), Kernel density estimation for the protein distance map of 2009 A(H1N1) influenza as of December 5, 2009. (b), The protein distance map of 2009 A(H1N1) influenza.

The Euclidean distance between two points in the map approximates the protein distance in Equation 5.1 between these two flu strains (see Materials and Methods for details of this distance approximation procedure). Two closely located points imply two strains with similar HA protein sequences.

We apply the low-dimensional clustering method to study 2009 A(H1N1). In Fig. 5.2, the vertical and horizontal axes of both figures represent protein distance as defined in Equation 5.1. A 0.0018 unit of protein distance equals one substitution in the HA protein sequence of H1N1. The height and colors in Fig. 5.2(a) both represent the density of isolates. We plot the protein distance map in Fig. 5.2(b). Both A/Texas/05/2009 and A/New York/19/2009 are located near the center of the cluster, in good agreement with the observation from Fig. 5.1 that they are the founder strains for many F1 strains. To detect the clusters in the protein distance map, we use a statistical method known as kernel density estimation[43]. Kernel density estimation is a non-parametric method to estimate the probability density function from which data come. The kernel density figure is produced from the protein distance map, and it shows the density of influenza strains in sequence space. We plot the kernel density as the three dimensional shaded surface. For example, the kernel density surface Fig. 5.2(a) is produced from Fig. 5.2(b). The x and y axes in Fig. 5.2(a) are the same as that in Fig. 5.2(b) and are protein distance coordinates. The z dimension measures the density of flu strains around point (x, y). We use the surface height and the colors to represent z values, and the color is proportional to surface height. A peak in kernel density Fig. 5.2(a) indicates a cluster of related flu strains in the protein distance map Fig. 5.2(b)

There are two significant clusters in the Fig. 5.2(a), as two peaks are observed. The cluster on the left side contains A/Texas/05/2009. Another cluster on the right side contains A/New York/19/2009. The 2009 A(H1N1) virus has evolved slowly to date. The greatest p_{epitope} antigenic distance between A/Texas/05/2009 and all sequenced strains is measured to be < 0.08. Values of p_{epitope} less than 0.45 for H1N1 indicate positive expected vaccine efficacy[92], and so a vaccine is expected to be efficacious. All of the amino acids in all five epitopes of a strain of A/Texas/05/2009 and a strain of A/New York/19/2009 are the same. Multidimensional scaling predicts that A/Texas/05/2009 will be the dominant strain in the 2009–2010 season, and that A/Texas/05/2009 is a suitable strain for vaccination. Our focus is on the expected vaccine effectiveness, as it can be judged from antisera hemagglutination inhibition (HI) assay or sequence data alone. We do not consider other aspects such as growth in hen's eggs or other manufacturing constraints. Laboratory growth and passage data are needed to address these aspects.

5.2.3 H3N2 virus evolution for 38 years

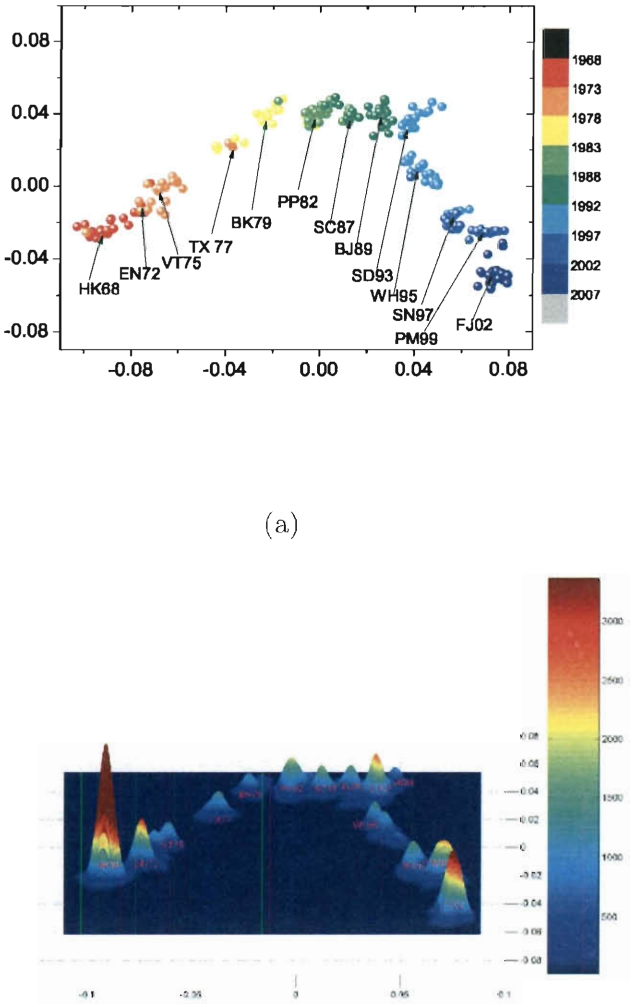
We construct the protein distance map to determine the evolution of influenza A(H3N2) virus from 1969 to 2007. Sequences of HA1 proteins were downloaded from the Influenza Virus Resources database[14]. We use the multidimensional clustering method[79] to generate the protein distance map and corresponding kernel density estimation in Fig. 5.3. Smith *et al.*[110] produced a similar graph using ferret antisera HI assay data. The figure presented here has a higher resolution, and more clusters are observed, because protein sequences data are more abundant and accurate than antisera HI assay data. The evolution of influenza tends to group strain into clusters. In Fig. 5.3, the vertical and horizontal axes of both figures represent protein distance as defined in Equation 5.1. A 0.0030 unit of protein distance equals one substitution in the HA1 protein sequence of H3N2. The colors in Fig. 5.3(a) represent the time of collection of the isolates. The colors and height in Fig. 5.3(b) represent the density of isolates.

we identified 14 major clusters by setting a cutoff value of kernel density for the past 38 years from 1969 to 2007. Each cluster is named after the first vaccine strain in the cluster. For example, HK68: Hongkong/1/68, EN72: England/42/72, VT75: Victoria/3/75, TX77: Texas/1/77, BK79: Bangkok/1/79, PP82: Philippines/2/82, SC87: Sichuan/2/87, BJ89: Beijing/32/92, SD93: Shandong/9/93, JB94: Johannesburg/33/94, WH95: Wuhan359/95, SN97: Sydney/5/97, PM99: Panama/2007/99, FJ02: Fujian/411/2002; The average duration time for a cluster is therefore 2.7 years, which is also the approximate duration of a vaccine. We marked each cluster by the first vaccine strain in the cluster. There are apparent gaps between clusters. The antigenic distance between two strains in two separate clusters is larger than the distances within the same cluster. The influenza virus evolves within one cluster before jumping from one cluster to another cluster. This dynamics occurs because small antigenic drift by one or a few sequential mutations does not lead the virus to completely escape from cross immunity induced by vaccine protection or prior exposure.

For vaccine design, when the viruses evolve as a quasispecies in the same cluster, the vaccine that is targeted to the cluster provides protection. This protection decreases with antigenic distance. When the viruses jump to a new cluster by antigenic drift or shift, one would want to update the vaccine to provide protection against strains in the new cluster. In Fig. 5.3(a), the arrows point to the exact position of vaccine strains. It can be seen that the positions of vaccine strains are near the center of clusters. It can be shown mathematically that choosing the consensus strain of a cluster as vaccine strain minimizes the p_{epitope} antigenic distance between vaccine strain and cluster strains, and thus maximizes expected vaccine efficacy[55].

5.2.4 Influenza vaccine strain selection

We now use the low-dimensional sequence clustering method in an effort to detect a new flu strain before it becomes dominant. A question of interest in the influenza research is whether we can predict which strain will be dominant in the next flu season based on



(a)

(b)

Figure 5.3 (a) The protein distance map and (b) corresponding Kernel density estimation of influenza from 1968 to 2007.

the information we have at present. The WHO gathers together every February to make a recommendation for influenza strains to be used in vaccine for next flu season in the Northern hemisphere. The vaccine is expected to have high efficacy if the chosen strain is dominant in the next flu season. The recommendation is especially challenging to make when the dominant strain in next flu season has not been dominant before February of that year. For example, in mid-March 2009, a new H3N2 strain appeared[108, 2], which infected a significant fraction of the population in the Southern hemisphere.

The current accepted influenza vaccine strain selection procedure is as follows [100]. Isolates samples are collected by WHO GISN and are characterized antigenically using the hemagglutination inhibition(HI) assay. About 10% of samples are also sequenced in HA1 domain of HA gene. Antigenic maps are constructed from the HI assay data using dimensional projection technique. Examination of HI data is not dependent on analysis using dimensional projection, but rather, the primary HI data may carry the most weight. If the vaccine does not match the current circulating strains, the vaccine is updated to contain one representative of the circulating strains. The emerging variant strains are identified. If the antigenically distinct emerging variants are judged to be the dominant strains in the upcoming season, the vaccine is updated to include one representative of emerging variants. The key issue and major difficulty is how to judge whether emerging variants will be the dominant variants in next season. If a fourfold difference in antisera HI titer between the vaccine strain and the emerging strains is observed, the emerging strain is to be determined to be dominant strains in upcoming season, and an updated vaccine is recommended to include the emerging strains[100].

Here, we propose a modified vaccine selection process based on clustering detection. First, we apply the multidimensional scaling to make a protein distance map from HA1 sequences, instead of constructing an antigenic map from HI assay data. Then, we use kernel density estimation to determine the clusters of strains. If the vaccine does not match the current circulating cluster, the vaccine is updated to contain the current circulating strain. If the vaccine matches the current circulating cluster, but an emerging cluster is judged likely to be the major cluster in the upcoming season, the vaccine is updated to contain the consensus strain of the emerging cluster. We judge whether a cluster is an emerging dominant cluster by two criteria. The first criterion is that this cluster can be detected by kernel density estimation, and is separate from the cluster that contains the current circulating strain or vaccine strain. A cluster that can be detected by kernel density estimation usually contains a central strain that has multiple identical copies and some F1 strains that are closely related to the central strain. An example is the cluster of A/Texas/05/2009(H1N1) in Fig. 5.1. A/Texas/05/2009(H1N1) is the central strain, which has 120 strains with identical HA protein sequences in the Influenza Virus Resource database[14]. A/Texas/05/2009(H1N1) also has 29 F1 strains with one amino acid different. So, A/Texas/05/2009(H1N1) and the surrounding strains form a cluster as we detected in Fig. 5.2 by kernel density estimation.

The second criterion is that the current vaccine strain does not match the consensus strain of the cluster and is estimated to provide low protection against strains in the cluster. That is, an immune response stimulated by a vaccine cannot effectively protect against infection by sufficiently distant by new strains. The consensus strain is a protein sequence that shows which residues are most abundant in the multialignment at each position. The efficacy of current vaccine to the new cluster can be estimated from ferret antisera HI assay data. However, the antisera data has low resolution and has an imperfect correlation to vaccine effectiveness in humans[157, 55]. Instead, we use p_{epitope}, which is calculated as the fraction of mutations in dominant epitope, to estimate vaccine efficacy and which has a more robust correlation to vaccine effectiveness in human than do ferret HI data[55]. When the p_{epitope} between the current vaccine strain and consensus strain of the new cluster is larger than 0.19, expected vaccine efficacy decreases to 0 for H3N2 influenza, and the current vaccine cannot be expected to provide protection from new strains. As the examples shown below, our method can detect an incipient dominant strain at its very early stage, and the method appears to require about 10 sequences in the new cluster for detection.

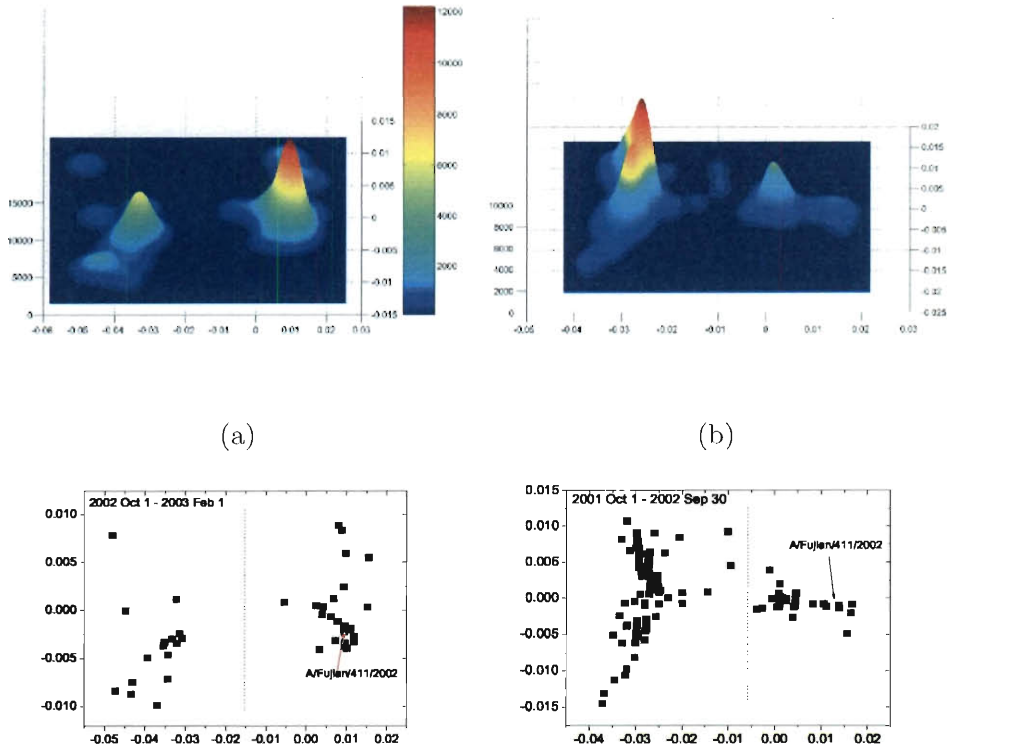
5.2.5 Demonstration of low-dimensional sequence clustering method.

We demonstrate the method of detecting the A/Fujian/411/2002(H3N2) strain. The A/Panama/2007/1999 had been the vaccination strain for four flu seasons between 2000 and 2004 in the Northern hemisphere. The vaccine strain was replaced by A/Fujian/411/2002 in the 2004-2005 flu season, as described in Table 5.1. The vaccine strain in the 2003-2004 season was A/Panama/2007/1999, while the dominant circulating strain became A/Fujian/411/2002(H3N2). This mismatch resulted in a large decrease in vaccine efficacy in the 2003-2004 flu season[55]. The vaccine efficacy is estimated to be only 12%[3]. We test whether our method can detect A/Fujian/411/2002(H3N2) as an incipient dominant strain before it actually became dominant. We use only virus sequence data before October 1, 2003. We did not use any virus data collected in 2003-2004 season. Therefore, our prediction and results are made without any knowledge from what happened in the 2003-2004 season. We plot the protein distance map of the 2001-2002 flu season in Fig.

5.4(d). To detect the clusters, we plot the kernel density in Fig. 5.4(b) for the data in Fig. 5.4(d). There are two separate significant clusters. The one with the largest kernel density on the left contains the current dominant strain A/Panama/2007/1999 and the widespread A/Moscow/10/1999 strain. The smaller one on the right is a new cluster, which contains A/Fujian/411/2002. Using the data as of September 30, 2002, we seek to determine whether the new cluster on the right in Figs. 5.4(b) and (d) will be the next dominant strain after A/Panama/2007/1999. We determine whether this cluster fulfills the two criteria above. First, this new cluster can be significantly detected by kernel density estimation. This cluster is separate from the current dominant strain, as we can see in figure. Second, we calculated the average p_{epitope} of the new cluster on the right with regard to A/Moscow/10/1999, A/Panama/2007/1999 and A/Fujian/411/2002 to be 0.214, 0.1214, and 0.083, respectively. This means the current vaccine contains A/Moscow/10/1999 is expected to provide little protection against viruses in the new cluster. This result makes the new cluster fulfill the second criterion. Thus, we predict based on the data as of September 30, 2002, that the cluster on the right in Fig. 5.4(d) will be the next dominant cluster. This prediction was made on data collected one year earlier than when the A/Fujian/411/2002 became dominant in the 2003-2004 season. To further support our prediction, in Fig. 5.4(c), we plot the protein distance map from October 1, 2002, to February 1, 2003, right before the WHO selected the vaccine strain for 2003-2004 season. To detect the clusters, we plot the kernel in Fig. 5.4(a) for the data in Fig. 5.4(c). There are two separate major clusters observed in the kernel density estimation in Fig. 5.4(a). The left cluster has the current dominant strain of A/Panama/2007/1999 and also A/Moscow/10/1999. The right cluster has the A/Fujian/411/2002. We calculated the average p_{epitope} of the right new cluster with regards to A/Moscow/10/1999, A/Panama/2007/1999, and A/Fujian/411/2002 to be 0.2725, 0.1811, and 0.0367 respectively. This result further supports the prediction that the new cluster will become dominant, and A/Fujian/411/2002, which is the most frequent strain in the new cluster, will be or is very close to the next dominant strain. This suggestion proceeds the vaccine component switch by 1-2 years, as shown in Table 5.1.

5.2.6 Prediction for H3N2 influenza in 2009–2010.

By applying our method to the 2008-2009 flu season, we predict that the dominant H3N2 strain in the 2009–2010 flu season may switch. Based on the flu activity in the 2008-2009 flu season, the WHO made the recommendation in February 2009 that A/Brisbane/10/2007(H3N2) should be used as the vaccine[150]. However, a new strain evolved just after the recommendation was published. The British Columbia Center for Disease Control detected a new virus strain[108, 2] with 3 mutations in antigenic sites (two in epitope B and one in epitope D). Since this new strain is relatively far from the vaccine strain, with $p_{\text{epitope}} = 0.095$, vaccine efficacy is expected to decrease to 20%[55, 33]. However, since the mutations in this new strain "do not fulfill the criteria proposed by Cox as corresponding to meaningful antigenic drift"[108, 28], and this strain still remained the minority of H3N2 viruses in July 2009, health authorities were not certain that this



(a)

(b)

(c)

(d)

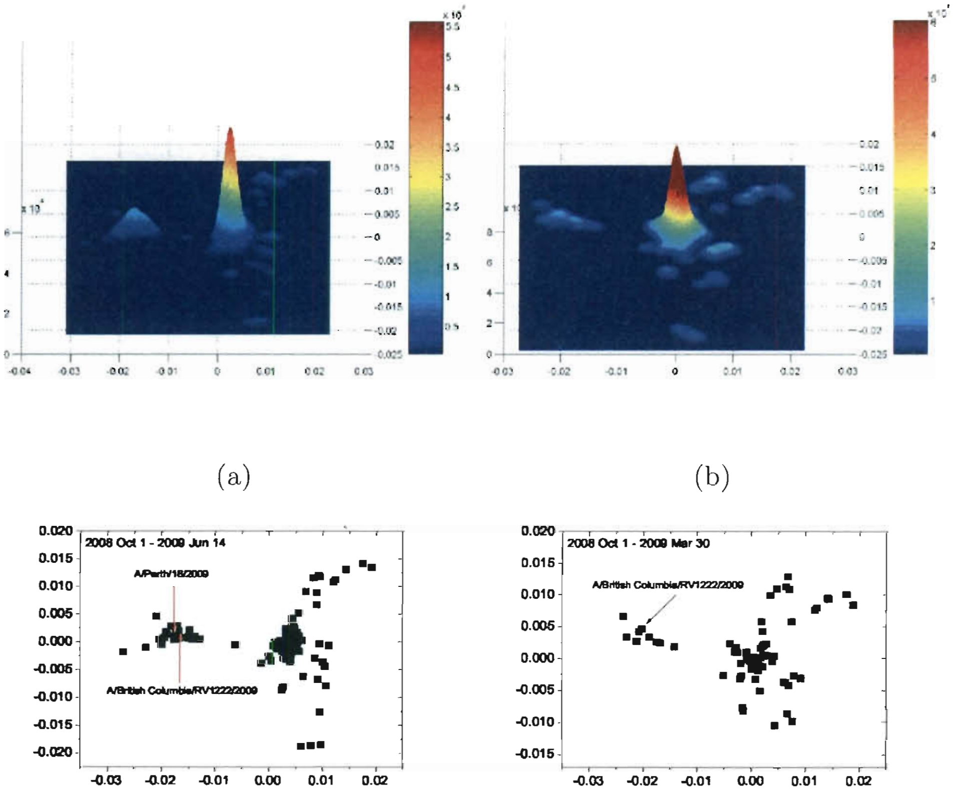
Figure 5.4 (a) Kernel density estimation and (c) protein distance map for H3N2 viruses between October 1, 2002 and February 1, 2003. (b) Kernel density estimation and (d) protein distance map for H3N2 viruses between October 1, 2001, and September 9, 2002. A 0.0030 unit of protein distance equals one substitution of the HA1 protein sequence of H3N2.

new strain would replace the current dominant strain in 2009–2010 flu season. We use our method to investigate whether this new strain will be the next dominant strain. We construct the protein distance map as shown in Fig. 5.5(c). We plot the kernel density estimation in Fig. 5.5(a) for data in Fig. 5.5(c). By the data up to June 14, 2009, we see two major clusters in Fig. 5.5(a). The larger one on the right contains the current dominant strain A/Brisbane/10/2007, and the left one is a new cluster which contains A/British Columbia/RV1222/2009. It is apparent that this new cluster is separate from the current dominant cluster. Thus, this cluster fulfills the first criterion. We calculated the average of p_{epitope} of strains in the left new cluster with regards to A/Brisbane/10/2007 and A/British Columbia/RV1222/2009 to be 0.103 and 0.042 respectively. The vaccine that contains A/Brisbane/10/2007 has an expected efficacy of 20% to the virus strains in the new cluster. Thus, this new cluster satisfies both two criteria, and so we predict that this cluster which contains A/British Columbia/RV1222/2009 will be the dominant cluster in the 2009–2010 season. The earliest time for us to make this prediction is March 30, 2009.

In Fig. 5.5(d) and (b), we already see this new cluster on the left side of figure, though since there are only about 10 sequences in the new cluster, the kernel density of this new cluster is smaller than that in the dominant cluster. This strain was mentioned as a concern on 5 May 2009, although by conventional methods the strain was not considered a potentially new dominant strain in July 2009[108]. With the method of the present paper, this new cluster is suggested earlier using the data as of March 30, 2009.

5.2.7 Comparison with previous results.

Here we present a historical test of the method. For each flu season in the North Hemisphere from 1996, we use only the H3N2 sequences data until February 1, before WHO published the recommendation for vaccine. We use the low dimensional clustering to made the prediction for the dominant strain. The conventional method as used by WHO is phylogenetic analysis combined with ferret antisera HI assay. In Table 5.1, we compare the method with the conventional method. This table includes the H3N2



(a)

(b)

(c)

(d)

Figure 5.5 (a) Kernel density estimation and (c) protein distance map for H3N2 viruses from October 1, 2008, to June 14, 2009. (b) Kernel density estimation and (d) protein distance map for H3N2 viruses between October 1, 2008, and March 30, 2009. A 0.0030 unit of protein distance equals one substitution of the HA1 protein sequence of H3N2.

vaccine strains, our prediction of dominant strains, the reported dominant circulating H3N2 strains[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 145, 146, 147, 148, 151], and the circulating subtypes in the northern hemisphere[133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 145, 146, 147, 148, 151]. Circulating H3N2 strains are absent if the dominant subtype is H1 or influenza B. The reported dominant H3N2 strains and circulating subtypes data are from WHO Weekly Epidemiological Record (http://www.who.int/wer/en/). In the most recent 14 flu seasons, influenza subtype H3 was dominant in 10. The WHO H3N2 vaccine component matches the circulating strains in 8 seasons. Our predictions match the circulating strains in 9 seasons. In 1997-1998 season, a novel flu strain Sydney/5/97 was found in June 1997. Because no similar strains were collected before February 1, neither of the two methods can predict it. In 2003-2004 season, our method predicts Fujian/441/2002 as the dominant strain, while phylogenetic analysis combined with ferret antisera HI assay did not. For all other 8 seasons dominated by influenza subtype H3, the predictions of both methods matched the dominant circulating strain. The 2009–2010 influenza season was dominated by H1N1. But data from local outbreaks of H3N2 infections[108, 2] showed that the dominant H3N2 strain was A/British Columbia/RV1222/2009, as predicted in Table 5.1, rather than the vaccine strain A/Brisbane/10/2007. For the 2010-2011 season, we recommend A/British Columbia/RV1222/2009 as a vaccine strain, and the WHO recommended A/Perth/16/2009. These two strains are in the same cluster and antigenically similar with a small $p_{\text{epitope}} = 0.048$. Although these two strains are slightly different, the vaccine is expected to be effective.

Flu season	Vaccine strain	Our prediction	Circulating	Circulating
	from WHO[150]		H3N2 strain	subtype
1996-1997	Wuhan/359/95	Wuhan/359/95	Wuhan/359/95	H3
1997-1998	Wuhan/359/95	Wuhan/359/95	Sydney/5/97	H3
1998-1999	Sydney/5/97	Sydney/5/97	Sydney/5/97	H3
1999-2000	Sydney/5/97	Sydney/5/97	Sydney/5/97	H3
2000-2001	Panama/2007/1999	Panama/2007/1999	N/A	H1
2001-2002	Panama/2007/1999	Panama/2007/1999	Panama/2007/1999	H3
2002-2003	Panama/2007/1999	Fujian/411/2002	N/A	H1
2003-2004	Panama/2007/1999	Fujian/411/2002	Fujian/411/2002	H3
2004-2005	Fujian/411/2002	Fujian/411/2002	Fujian/411/2002	H3
2005-2006	California/7/2004	California/7/2004	California/7/2004	H3
2006-2007	Wisconsin/67/2005	Wisconsin/67/2005	Wisconsin/67/2005	H3
2007-2008	Wisconsin/67/2005	Wisconsin/67/2005	N/A	H1
2008-2009	Brisbane/10/2007	Brisbane/10/2007	Brisbane/10/2007	H3
2009-2010	Brisbane/10/2007	BritishColumbia/RV1222/09	BritishColumbia/RV1222/09	H1
2010-2011	Perth/16/2009	BritishColumbia/RV1222/09	N/A	N/A

Table 5.1 Summary of results

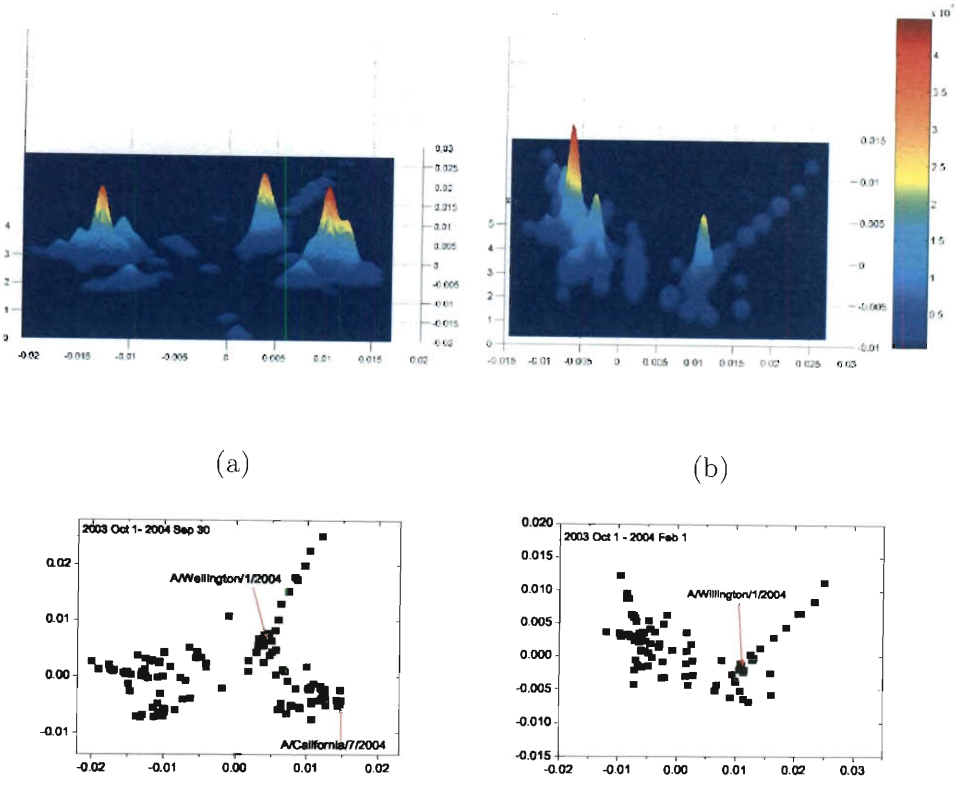
5.2.8 Detecting A/Wellington/1/2004 in the 2004 flu season in the Southern hemisphere

The low-dimensional clustering can also be applied to influenza in the Southern hemisphere. As an example, we test our method on the 2004 flu season. The recommended H3N2 vaccine strain by WHO used in the 2004 flu season in the Southern hemisphere was A/Fujian/411/2002. Data from the surveillance network suggested that the circulating dominant flu strain in the 2004 season in Southern hemisphere was A/Fujian/411/2002, and a late surge of A/Wellington/1/2004 was also observed. For example, in Argentina, a study showed that about 50% of infections were closely related to A/Fujian/411/2002 and another 50% were closely related to A/Wellington/1/2004[102]. In New Zealand, the dominant flu strain was A/Fujian/411/2002 which caused 78% of flu infections[4], and a late season surge of A/Wellington/1/2004 was also reported[1]. Therefore, the vaccine recommended by WHO matches the dominant strain and would be expected to have vaccine efficacy in the 2004 season in Southern hemisphere.

We here use the low-dimensional clustering method to detect the A/Wellington/1/2004 strain, which is not the major dominant strain but caused significant infections in the 2004 flu season. We plot the protein distance and kernel density estimation for the H3N2 viruses in Fig. 5.6(d) and (b). We use the data only as of February 1, 2004, 3 months prior to the 2004 flu Southern hemisphere season, which is usually from May to September. We observed two clusters. The major cluster on the left side of Fig. 5.6(d) is A/Fujian/411/2002-like, which was the vaccine strain in 2004 season. There is a new cluster in the right side of Fig. 5.6(d) which contains A/Wellington/1/2004. The p_{epitope} of A/Wellington/1/2004 with regards to A/Fujian/411/2002 is 0.118. Therefore, we predict that A/Wellington/1/2004 will infect a large fraction of the population, and the A/Fujian/411/2002 vaccine is expected to provide only partial protection against the A/Wellington/1/2004 virus. However, since the appearance of A/Wellington/1/2004 was just before the the 2004 flu season, it did not have sufficient time to spread out and become the dominant strain in the 2004 flu season. From our observation, it usually takes about 8 months or longer for a new strain to become dominant after its appearance in a new cluster. Therefore, the predominant flu strain in 2004 season is expected to be A/Fujian/411/2002 based on the data as of February 1, 2004. This result agrees with the dominant flu strain in the 2004 flu season.

5.2.9 Detecting A/California/4/2004 as a future dominant strain

As a further example of applying the low-dimensional clustering method to influenza in Southern hemisphere, we test the method on the 2005 flu season. The H3N2 vaccine strain in the 2005 flu season in the Southern hemisphere was A/Wellington/1/2004. Data from HI assay tests and surveillance suggest that the dominant H3N2 strain in the 2005 season was A/California/7/2004. In HI tests with postinfection ferret sera the majority of influenza A(H3N2) viruses from February 2005 to October 2005 were closely related to A/California/7/2004, as reported by WHO on October 7, 2005[144]. Surveillance data from Victoria, Australia, show that 45% of influenza A infections



(a)

(b)

(c)

(d)

Figure 5.6 (a), Kernel density estimation for the protein distance map for H3N2 viruses between 10/01/2003 and 09/30/2004. (b), Kernel density estimation for the protein distance map for H3N2 viruses between 10/01/2003 and 02/01/2004. (c), Protein distance map for H3N2 viruses between 10/01/2003 and 09/30/2004. We plot a dotted line to separate the two clusters. (d), Protein distance map for H3N2 viruses between 10/01/2003 and 02/01/2004. The vertical and horizontal axes of all figures represent protein distance. A 0.0030 unit of protein distance equals one mutation of the HA1 protein sequence of H3N2.

were A/California/7/2004-like (H3), 11% were A/Wellington/1/2004 (H3) and 44% were A/New Caledonia /20/99-like (H1), as collected in the 2005 flu season[121]. Surveillance data from New Zealand also show that the dominant H3N2 strain in the 2005 flu season was A/California/7/2004[5].

We plot the protein distance for the H3N2 viruses in the 2003-2004 flu season in Fig. 5.6(c). We only use the data as of September 30, 2004, earlier than the October 2004 date when the WHO published the influenza vaccine recommendation for Southern hemisphere. We plot the kernel density estimation in Fig. 5.6(a) for the data in Fig. 5.6(c). There are three major clusters in Fig. 5.6(a). The one on the left is the current dominant cluster which are mostly A/Fujian/422/2002-like viruses. There is a middle cluster centered on A/Wellington/1/2004. The one on the right contains A/California/7/2004. Both the A/California/7/2004 cluster and the A/Wellington/1/2004 cluster are antigenically novel from A/Fujian/411/2002.

When the protein distance map and kernel estimation as of February 1, 2004, is plotted in Fig. 5.6(d) and (b), we still see the A/Wellington/1/2004 cluster. With these data, the A/California/7/2004 cluster is no longer observed. Thus, A/California/7/2004 cluster is a newly appearing cluster and we consider it to be the emerging strain. The new cluster which contains A/California/7/2004 is separate from the current dominant cluster. We calculated the average p_{epitope} of the new cluster that contains A/California/7/2004 with regard to A/Fujian/411/2002 to be 0.112. This makes the new cluster fulfill both criteria for an incipient dominant strain cluster. So we predict based on the information as of September 30, 2004, that A/California/7/2004 will be the next dominant strain after A/Fujian/411/2002 in Southern hemisphere. We further predict from these data that A/California/7/2004 will be the dominant strain in the following flu season in the Northern hemisphere. These predictions agree with the observed dominant strain in the 2005 flu season.

5.3 Discussion

The evolution of influenza virus is driven by cell receptor distributions, non-specific innate host defense mechanisms, cross immunity[54, 47], and other contributions to viral fitness. In this paper, we focused on HA protein evolution under antibody selection pressure. The degree to which the immunity induced by one strain protects against another strain depends on their antigenic distance[55]. Because the human immune response to viral infection is not completely cross protective, natural selection favors amino-acid variants of the HA protein that allow the virus to evade immunity, infect more hosts, and proliferate. Mutant strains surround the dominant strain and group into a cluster rather than evolve in a defined direction[110, 94]. After the virus has circulated in population for one or more years, effective vaccines and cross immunity of the population drive the evolution of influenza by mutation and reassortment. This evolution increases the immune-escape component of the fitness of new strains, and eventually causes a new epidemic. These new immune-escape strains will form a new cluster, and the old clusters will die out, thus starting a new cycle. This process of creating of new clusters is what our method detects.

The low dimensional clustering can be used not only in genetic sequences but also on distances calculated from inhibition assays of antibody and antigens, as first shown by Lapedes and Farber[79] and Smith *et al.*[110]. The inhibition assay provides an approximation of antigenic distance and is broadly used as a marker for vaccine efficacy. The inhibition assay suffers from low resolution of data, which multidimensional scaling improves, and is less able to predict the vaccine efficacy than the p_{epitope} method[55]. The genetic sequences used here are a direct description of the evolution of pathogen and antigenic distance of influenza. To aid vaccine selection, the low dimensional clustering on genetic sequences appears informative.

Challenges may arise in application of the method described here. If two or more new clusters appear in one season, additional information is needed to decide which cluster should be chosen for vaccine. Fortunately, it has been shown that the evolution of influenza is typically in one direction[47, 110]. It is rare to have two or more new clusters in the protein distance map in one season. As experience with the low dimensional sequence clustering is gained, it may be that cluster structure will allow more precise prediction of vaccine efficacy. Despite these issues, the method described here can assist the design of vaccines, and it provides a new tool to analyze influenza viral dynamics. We did not see any false positive results in Table 5.1.

The current WHO method works quite well in many years. The method discussed here appears to offer an additional tool which may provide additional utility.

5.4 Materials and Methods

5.4.1 Data sources

Influenza hemagglutinin A(H3N2) sequences before October 1, 2008, and A(H1N1) sequences as of December 5, 2009, were downloaded from NCBI Influenza Virus Resources[14]. All hemagglutinin sequences used in our study are filtered by removing identical sequences. Thus, all groups of identical sequences in the dataset are be represented by the oldest sequence in each group. This approach reduces the number of sequences by keeping only the unique sequences in the dataset. Influenza A(H3N2) sequences after October 1, 2008, were downloaded from GISAID database.

5.4.2 Geographical spread pattern of 2009 A(H1N1)

It is believe that the 2009 A(H1N1) virus was most likely originated from Mexico[49]. It first spread to the neighboring country USA and then to other countries. We display this geographical spread pattern in Fig.5.1. We take the founder-F1 relationship from Fig. 5.1, and assume the virus spreads from location of founder to the location of F1. We consider three regions: USA, Mexico and other countries except USA and Mexico. Then we count the cases of spreading from one region to another region. In Table 5.2, we show that we observed many more paths of spreading from the USA to other countries than from other countries to the USA. The major path of spreading is from USA to other

Spreading path	Number of cases
USA to Others	32
Others to USA	1
Mexico to Others	1
Others to Mexico	0
Others to others	6

Table 5.2 The geographical spread pattern of 2009 A(H1N1). "Others" refers to other countries except USA and Mexico.

countries. This result indicates our directional evolutionary map of Fig. 5.1 is in good agreement with the pattern of geographical spread.

5.4.3 Multidimensional scaling

The goal of multidimensional scaling is to represent the distance of proteins by a Euclidean distance in coordinate space. We calculate the distance between proteins i and j, d_{ij}, by the number of amino acid residue differences divided by the total number of amino acid residues, as defined by Equation 5.1. To do multidimensional scaling, we start with the distance of the proteins. The object of multidimensional scaling is to find the two, or p in general, directions that best preserve the distances d_{ij} between the N

proteins

$$F = \sum_{i,j=1}^{N} (d_{ij} - D_{ij})^2 \tag{5.2}$$

Here, $D_{ij} = \|x_i - x_j\|$ is the Euclidean distance between proteins i and j in the projected space, and $\|\bullet\|$ is the vector norm. The algorithm is as follows. Let the matrix $A = [(a_{ij})]$, where $a_{ij} = -\frac{1}{2}d_{ij}^2$. The eigenvalues of A are $\gamma_1, \gamma_2, ..., \gamma_N$ and $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_N$. Let $V^{(1)} = (v_1^{(1)}, v_2^{(1)}, ..., v_N^{(1)})$ be the eigenvector of γ_1 and $V^{(2)} = (v_1^{(2)}, v_2^{(2)}, ..., v_N^{(2)})$ be the eigenvector of γ_2. Let $x = \sqrt{\gamma_1}V^{(1)}$ and $y = \sqrt{\gamma_2}V^{(2)}$. The two coordinates in protein distance maps are x and y. The x-axis in the protein distance map is the largest eigenvector. We take H3N2 2008–2009 season as an example. In Fig. 5.5(c), we observe two clusters. One cluster is on the right side of figures with x value positive and another one has negative x values. We define the consensus sequence of a group of flu strains by taking the most frequent amino acid at each position. We calculate the consensus sequences both for the strains in the cluster on the right and on the left of figure. We found

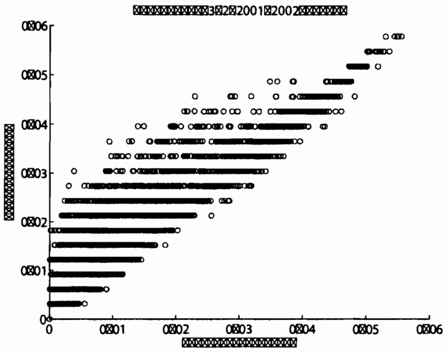


Figure 5.7 Plot of Euclidean distances of proteins as in Fig. 5.4(d) on x-axis and plot of distance of corresponding proteins in y-axis. Closeness to the diagonal measures fidelity of the low dimensional projection. A 0.0030 unit of protein distance equals one mutation of the HA1 protein sequence of H3N2.

amino acids at four positions (76, 160, 172, 203) are different for these two consensus H3N2 strains, see Table 5.3. Interestingly, the Shannon entropy calculated from all 2008–2009 season sequences at these four positions (0.43, 0.67, 0.59, 0.50) are the largest, which means the diversity at these four position are the largest.

There is software available to run the multidimensional scaling. We use the Matlab function "CMDSCALE" to generate an $N \times p$ configuration matrix Y. Rows of Y are the coordinates of N points in p-dimensional space. The "CMDSCALE" also returns a vector

Position in HA1 protein of H3N2	76	160	172	203
Amino acid in consensus strain 1	Glu	Asn	Lys	Asn
Amino acid in consensus strain 2	Lys	Lys	Asn	Lys
Shannon Entropy	0.43	0.67	0.59	0.50

Table 5.3 Consensus strain 1 is the calculated from all strains in the cluster on the right side of Figure 5.5(c). Consensus strain 2 is the calculated from all strains in the cluster on the left side of Figure 5.5(c).

E containing the sorted eigenvalues of what is often referred to as the "scalar product matrix," which, in the simplest case, is equal to $\mathrm{YY}^{\top}$. If only two or three of the largest eigenvalues E are much larger than others, then the matrix D based on the corresponding columns of Y nearly reproduces the original distance matrix d. We used the influenza H3N2 in 2001–2002 season as an example. The five largest of all 180 eigenvalues are 0.0361, 0.0032, 0.0024, 0.0020, 0.0016. The first two largest eigenvalues contribute 70% to the sum of all 180 eigenvalues, which indicates $p = 2$. Then, we plot the the N points in a two-dimensional graph. Each point represents a protein. The Euclidean distance between any two points D_{ij} on the graph should be equal to or close to the distance of these two proteins. that is, $D_{ij} \approx d_{ij}$. As an example, in Fig. 5.7, we show that D_{ij} and d_{ij} have a strong linear relationship. A short MATLAB program of multidimensional scaling is as follow.

```
% Multidimensional scaling.

% alignment.aln is a sequence multialignment file

% generated by software ClustalW.

clear

Sequences = multialignread(’alignment.aln’);

distances = seqpdist(Sequences,’Method’,’p-distance’);

Y = cmdscale(distances);

scatter(Y(:,1), Y(:,2));
```

5.4.4 Biases in the data

There are two biases in the sequence data. First, more isolates are sequenced in recent years. Generally speaking, more sequences make the vaccine selection based on low-dimensional clustering methods more reliable. That is why we compared low-dimensional clustering methods with WHO results only since 1996 in Table 5.1. To avoid these biases in the generation of the figure of evolution history of influenza for the 40 years (Fig. 5.3), we choose 20 random isolates for each season, even though the database contains more sequences in recent years. Second, most isolates are collected in USA. We found that many isolates collected in USA are identical, because of the high sampling rate in USA. To reduce this bias, we collapse redundant strains, keeping only distinct strains.

Part III

Bacterial and animal immune systems

Chapter 6

Heterogeneous Diversity of Spacers within CRISPR

6.1 Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) has been recently suggested to provide adaptable immunity to bacteria and archaea[18, 22, 35]. A typical CRISPR system is composed of CRISPR-associated(*cas*) genes and a CRISPR-cassette[73, 20, 95]. A CRISPR-cassette is formed by nearly identical direct repeats of 24-47 bp long nucleotides separated by similar sized, unique spacers. Repeats usually show some dyad symmetry but are not truly palindromic, implying the presence of a conserved secondary structure. Arrays of the same CRISPR are commonly followed by a conserved AT-rich sequence known as the leader. The leader appears to promote transcriptions towards the repeats, generating the RNAs that constitute the molecular base of the interference action. Recent studies have proposed that CRISPRs and *cas* genes function in anti-viral defense. A considerable fraction of spacer sequences are found to be similar to known phage sequences, indicating that the spacer sequences may derive from viruses and phages[20]. Moreover, when bacteria that possess the CRISPR-Cas system are exposed to viruses, the surviving individuals appear to have new virus-derived sequences at the leader-proximal end of CRISPR loci[18, 35]. Further, the acquisition or loss of CRISPR elements or of Cas protein genes has been directly correlated with phage and plasmid resistance or sensitivity, respectively[18, 35, 22]. The CRISPR system has began to attract a large amount of attention due to its potential role in restricting horizontal gene transfer. Because CRISPR system interference targets external DNA directly, it may prevent conjugation and plasmid transformation[82]. CRISPR system can also be used in anti-phage bacteria selection[70].

Recent experiments demonstrated the heterogeneous distribution of diversity of the spacers in the CRISPR system[122, 11]. However, the mechanism by which the phage-bacteria interaction shapes the spacer structure is poorly understood. In this paper, we propose a model that describes how the newly added spacers are more diversified and the old spacers are more conserved due to selective pressure on the CRISPR system. This model explains the underlying mechanism that shapes the spacer structure. This model confirmed that the diversity of CRISPR spacers is essential for the survive of bacteria.

6.2 Results

We describe the CRISPR-phage dynamics in the schematic representation of Fig. 6.1. In this figure, spacers are shown in numbers, and repeats are shown in dark squares. Leader sequences are directly adjacent to the short spacer-repeat units and possibly involved in promoting transcription towards the repeats. The virus DNA that is recognized by CRISPR is represented by the letter "i." Only the CRISPR of the bacterial genome are shown; other parts of genome are assume to be identical in all bacteria strains. When bacteria are exposed to phage viruses, there are three possible scenarios: bacteria are infected, viruses are defended, or bacteria acquire new spacers. In Fig. 6.1, the bacteria incorporate a piece of the phage DNA represented by the letter 'i' into its own genome as a new spacer. New spacers are always added to the leader-proximal end[70]. To avoid infinite growth of CRISPR, an old spacer is dropped when CRISPR in excess of certain length[122]. The CRISPR system provides an immune response. After insertion of exogenous DNA from phages or plasmids, the CRISPR spacers are transcribed and processed to CRISPR RNA units. The CRISPR RNA units serve as templates to recognize foreign nucleotide acids. If any of the CRISPR RNA units match the phage-derived sequences, the phage genetic material is degraded by bacteria. If none of CRISPR RNA units matches the phage-derived sequences, the bacteria are likely to be infected by the phage, and the phages will reproduce. When bacteria divide, the CRISPR are copied to the daughter cells[75].

6.2.1 Differential equation model

We use a population dynamics model to describe the bacteria-virus community. We assume only one CRISPR locus for each bacteria individual. We first consider a simple

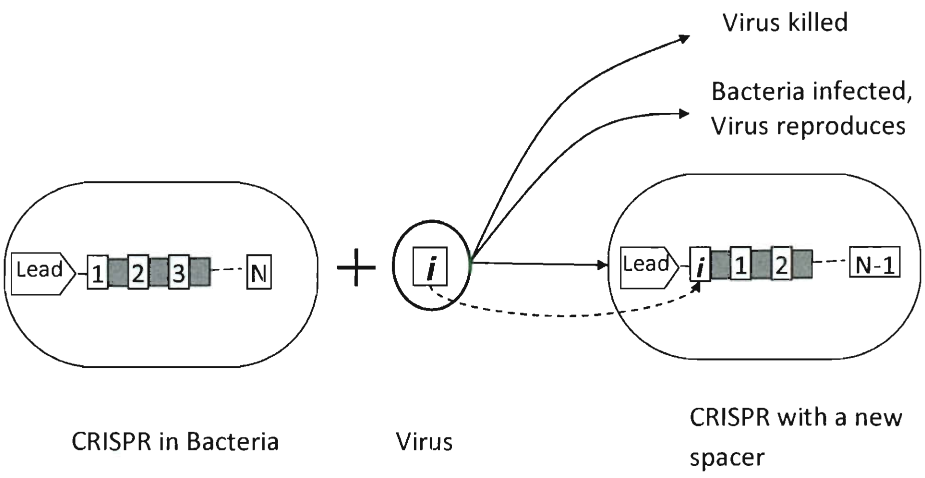


Figure 6.1 A schematic representation to describe CRISPR-phage dynamics.

case in which there are no more than two spacers for each CRISPR locus. By the first spacer, we mean the spacer that is nearest to the leader sequences. The second spacer is the spacer that is the next nearest to the leader sequences. We consider the following system of ordinary differential equations:

$$\frac{dx_{i,j}}{dt} = cx_{i,j} - \beta \sum_{k \neq i,j} v_k x_{i,j} + \beta\gamma \sum_m x_{j,m} v_i \tag{6.1}$$

$$\frac{dv_k}{dt} = rv_k - \beta \sum_{i,j} x_{i,j} v_k (\delta_{i,k} + \delta_{j,k}) \tag{6.2}$$

There are two variables in the above equations: v_k is the population of virus strain k, and $x_{i,j}$ is the population of bacteria with CRISPR with spacers i and j. The first spacer recognizes virus strain i and the second spacer recognizes virus strain j. In the absence of phage infection, the bacterial growth is exponential at rate c. The term $\beta \sum_{k \neq i,j} v_k x_{i,j}$ represents the bacteria with spacers of type i and j infected by viruses strains other than i or j. Bacteria can be infected or killed when they are exposed to viruses that bacteria do not recognize by CRISPR. The exposure rate of bacteria to virus is β. The term $\beta\gamma \sum_m x_{j,m} v_i$ represents the process of the converting other types of bacteria into bacteria of type i, j. When bacteria of type j, m incorporate virus of strain i into their own genome and add a new spacer, bacteria type j, m are converted to type i, j. The probability of adding a new spacer when a bacteria is exposed to a virus is γ. In the absence of resistance from CRISPR, viral growth is exponential at rate r. The term $\beta \sum_{i,j} x_{i,j} v_k (\delta_{i,k} + \delta_{j,k})$ represents the degradation of viruses by bacteria. If any spacers of bacteria of type i, j match viruses of strain k, the bacteria degrade the viruses. The

Kronecker delta function $\delta_{i,k}$ is 1 if spacer type i matches virus strain k; otherwise, it is 0. This model is modified from the classic immune response model with antigenic variation[89]. In this model, we take only the essential factors into consideration. We do not distinguish the lysis and lysogeny cycle. Horizontal gene transfer is not considered. Furthermore, because viruses usually have more than one type of host to infect, viral growth is not limited by one specific type of target bacteria abundance[89, 11].

Solution of the model shows that the diversity of old spacer decreases upon challenge by viruses. We solve the differential equations by Matlab software using Runge-Kutta method. The initial value for the differential equations are naive bacteria whose CRISPR provide no resistance to viruses because their spacers are empty. The population of bacteria drops fast at the beginning. Some bacteria acquire spacers from viruses and therefore develop resistance. The population of bacteria is steadily recovered. We measure the diversity of spacers by the Shannon entropy:

$$D_1 = -\sum_i (\sum_j P_{i,j}) \ln(\sum_j P_{i,j}) \tag{6.3}$$

$$D_2 = -\sum_j (\sum_i P_{i,j}) \ln(\sum_i P_{i,j}) \tag{6.4}$$

$$P_{i,j} = \frac{x_{i,j}}{\sum_{m,n} x_{m,n}} \tag{6.5}$$

Here, D_1 and D_2 are the diversity for the first and second spacers. Because new spacers are always added to the leader-proximal end, the first spacer is "younger" than the second spacer. If there is no selective pressure on CRISPR, or CRISPR do not provide resistance against viruses, the diversity of spacers along CRISPR should be homogeneous, $D_1 = D_2$, because adding and deleting spacers is completely random. With the selective pressure on CRISPR to evolve resistance to phage, we observe a decline of diversity of the second spacer, as shown in Fig. 6.2. In this figure, the differential equation solution and simulation are based on the parameter values $c = 0.15, \beta = 2 \times 10^{-6}, \gamma = 0.1$, and $r = 0.01$. The viruses have four strains (length of string $n = 2$) with an initial population ratio 6:2:1:1. The maximal population size allowed is 10^6 for virus and 10^5 for bacteria. Diversity is measured by Shannon entropy. Other measures of diversity such as Simpson's index of diversity give similar results. Error bars are one standard error. The insert figure is solutions of differential equations with 200 different parameter combinations using LHS. The up branches are the first spacer, and the down branches are the second spacers. At the beginning, both positions have high diversity of spacers. With the continuous challenge of viruses and selective pressure for the effective resistance against viruses, the diversity of spacers at the second position decreases with time. When steady state is reached after some time, we observe that the diversity of spacers at the second position is lower than that at the first spacer. Our observation is true for a broad choice of parameters. Parameter space was explored by using the statistical technique of Latin hypercube sampling (LHS). LHS selects combinations of parameter values from parameter value range and probability distribution function. The parameter ranges we used are: $c \in (0.01, 0.15), \beta \in (10^{-5}, 2 \times 10^{-5}), \gamma \in (0.01, 0.1), r \in (0.01, 0.1)$. We used

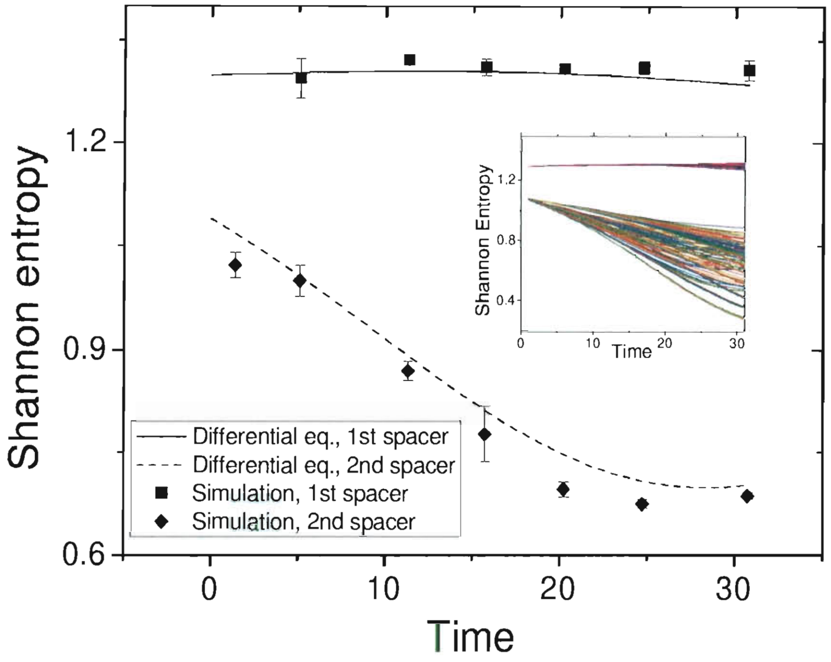


Figure 6.2 Diversity of two spacers of CRISPR with time.

200 samplings to sample parameter space. In the insert of Fig. 6.2, we observe that diversity of the old spacer is decreasing and the diversity of the young spacer is almost constant over time for all samplings.

Selection for bacteria that contain the most effective spacers decreases the diversity of the old spacer. The bacteria randomly take virus genomes from the environment and incorporate a corresponding spacer. Therefore, the diversity of the first spacer approaches the diversity of viruses in the environment. If the spacers match the dominant virus strain, bacteria containing these spacers are more likely to survive, and therefore spacers that match dominant viruses accumulate in the CRISPR. Bacteria that contain unused spacer elements that provide little protective potency are more likely to be infected by phage. The spacers corresponding to the dominant virus strain are enhanced and accumulate at the second spacer position. Therefore, the diversity of the second spacers decreases.

6.2.2 Stochastic simulation model

We seek to identify finite size effects by stochastic simulation. We use a stochastic individual-based model (IBM) for the large bacteria and virus population by Lebowitz-Gillespie algorithm. Each bacteria and each virus is an agent. Viruses are represented as bit-strings. Each bit has two alleles, designated as a “1” or “0”. In individual-based model, the length of virus strings are n, therefore 2^n genotypes are available for viruses. For bacteria, we ignore other parts of genome, but consider CRISPR locus only. Each spacer is n bit long[75], which is the same size as viruses. The simulation starts with a population of viruses of different genotypes and bacteria without spacers in CRISPR locus. Viruses infect bacteria with a contact rate β. If any spacer of bacteria matches the infecting virus, the virus is killed. Otherwise, the bacteria is infected and die. Bacteria and viruses reproduce at rate c and r respectively. Bacteria add a new spacer with a rate γ from contacting virus. In Fig. 6.2, we adopt the same parameters as used in differential equations. We observed the simulation results agree with the analytic results.

We further extend our individual-based simulation to allow the CRISPR to have more spacers, random loss of spacers and mutation. Most CRISPR contain fewer than 50 repeat-spacer units. For example, the average number of spacers of *Streptococcus thermophilus* is 23 per CRISPR locus in one study[71]. In our extended simulation, when the array of spacers of bacteria is more than 30, a spacer is randomly deleted with probability proportional to its distance to the leader sequence. When viruses replicate, the mutation rate of each strain is ε. We perform mutation by randomly flipping one bit of viruses bit-string from "1" to "0" or from "0" to "1". The extended simulation starts with a population of 150 virus genotypes and bacteria without spacers. The simulation runs until it reaches steady state. We run the simulation 100 times to average the results. After the simulation reaches steady state, we calculate the diversity of spacers for each position by Shannon entropy. In Fig. 6.3, the positions with a small number in the x-axis are leader-proximal. In this extended simulation, we use the parameters: $c = 0.15, \beta = \times 10^{-5}, \gamma = 0.1$, $r = 0.05$, mutation rate $\varepsilon = 0.01$, size of virus bit-string $n = 10$. Initially, there are 150 phage strains with a logarithm population distribution[122]. Other parameter settings give similar results. Error bars are one standard error. In Fig. 6.3, we observe that the "young" spacers which are leader-proximal are highly diversified and that the "old" spacers which are leader-distal are more conserved.

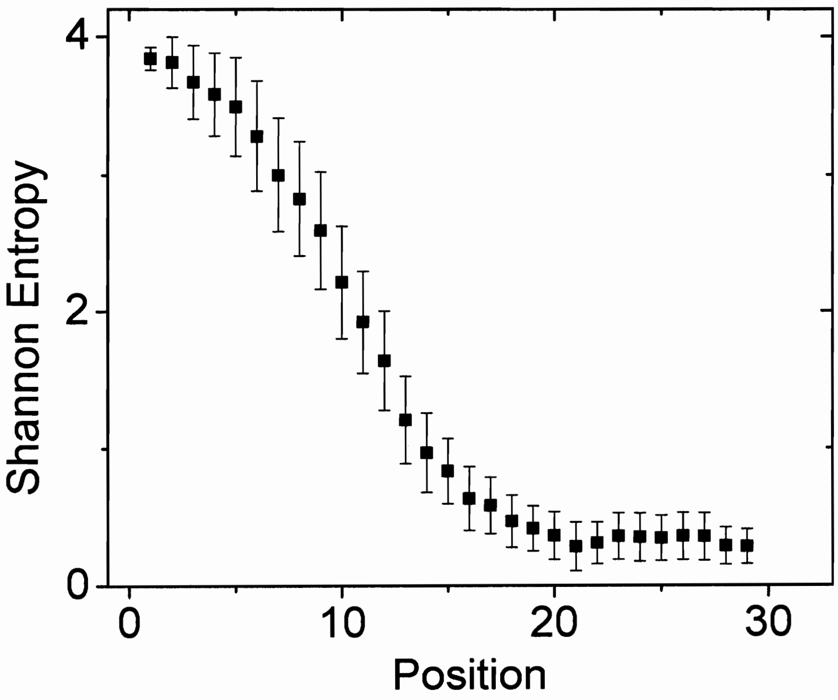


Figure 6.3 Diversity of spacers at different positions of CRISPR, when the system reaches steady state.

These results support the following scenario: Infection by a novel viral genotype results in the lysis or weakening of most individuals, except those that are able to capture and incorporate a corresponding spacer into their CRISPR locus. Resistant individuals rapidly gain a selective advantage, leading to the fixation of the resistant spacer. Increasing polymorphism toward the leader-proximal end provides support that the CRISPR are an actively evolving and functioning phage defense mechanism.

6.2.3 Experimental results

This model is in agreement with recent experiment results. Horvath *et al.*[71] sequenced the CRISPR regions of 124 *S. thermophilus* strains and analyzed 3626 spacers, 926 of which are unique. We aligned the spacers of CRISPR loci 1 for 124 strains. Shannon entropy was calculated for each aligned position, see Fig. 6.4. Spacers at leader-proximal positions are more diverse and spacers at leader-distal positions are highly conserved across strains. For example, at the most leader-distal position, 34 of 124 strains

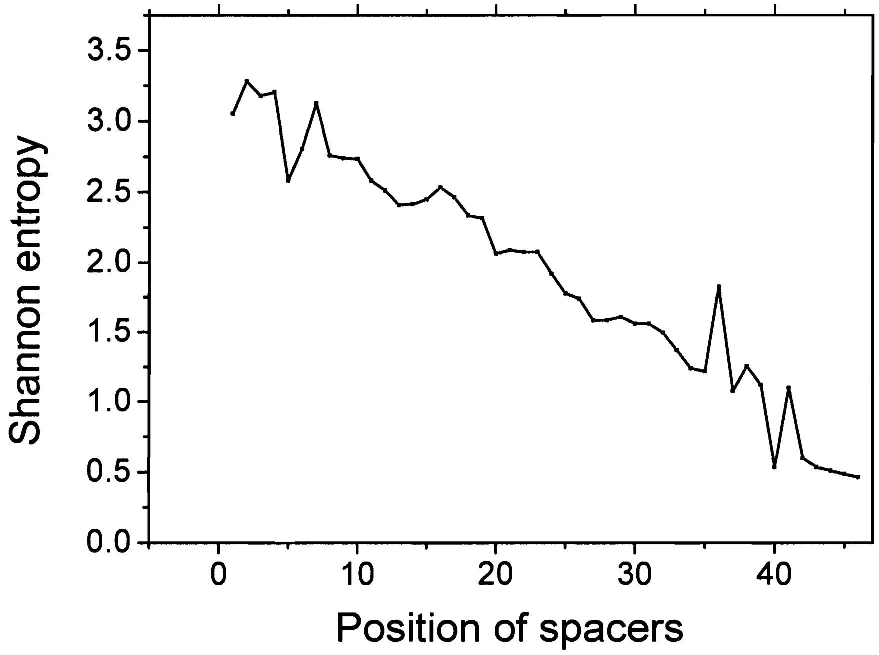


Figure 6.4 Diversity of spacers of CRISPR loci 1 of *S. thermophilus* strains[71]. The positions with a small number in the x-axis are leader-proximal.

share the identical spacer.

Recent metagenomic studies of environmental microbial samples provide a population-wide view of the dynamics between phage and CRISPR of the hosts[122, 11, 66]. In one study, sequence data were assembled from biofilm community samples[122, 11]. The CRISPR loci of the predominant *Leptospirillum* species display extensive polymorphism.

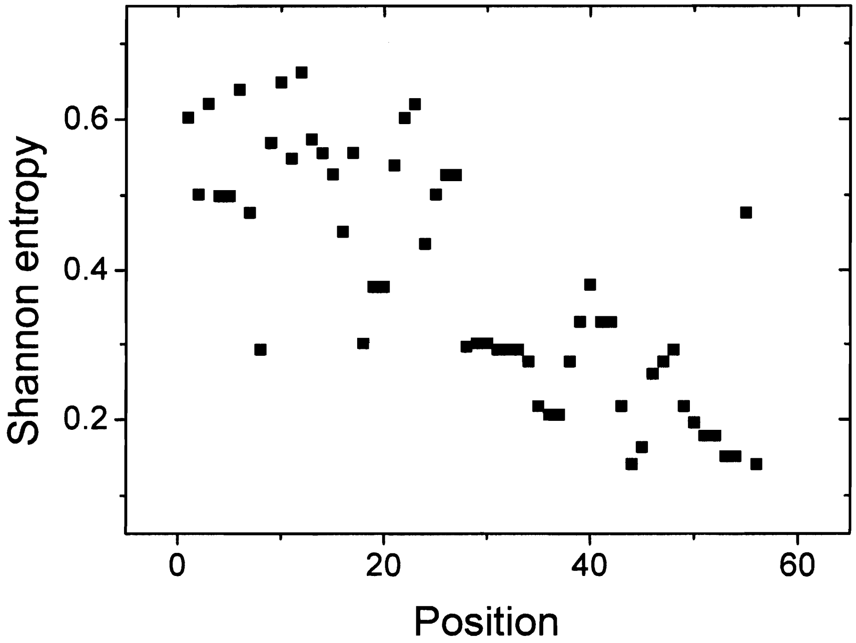


Figure 6.5 Diversity of spacers of CRISPR loci of *Leptospirillum* species. The data are noisy because the CRISPR loci sequence data of *Leptospirillum* are fragmented.

We calculate the Shannon entropy for each position of CRISPR, see Fig. 6.5. The bacteria community shared spacer sequences at the leader-distal end of their CRISPR loci, while the leader-proximal end of the loci contained spacers that were mostly unique to each individuals. The decrease of diversity of spacers from leader-proximal end to leader-distal end supports a model in which highly plastic CRISPR loci continuously respond to challenge from a rapidly evolving pool of phage.

6.3 Conclusion

To sum up, the CRISPR provide adaptable immunity to bacteria and archaea. Bacteria continuously incorporate nucleotide material from phage genomes into CRISPR to gain resistance against phage infection. Viruses continuously perform nucleotide mutation and horizontal gene transfer to avoid being recognized. The coevolution interaction between viruses and bacteria CRISPR system has shaped the spacer structure of CRISPR locus. Both our model and recent experiments support the declining diversity of spacers towards leader-distal end, implying that the CRISPR is an actively anti-viral system. Our model explored that the underlying mechanism of shaping spacer structure is the selection of bacteria CRISPR systems that match best with viruses in the environment, and the diversity of bacteria CRISPR is vital for survive. Further effort can extend our model to study the population dynamics of phage under the pressure of CRISPR.

Chapter 7
Regulated mechanism in antibody VDJ recombination

7.1 Introduction

The adaptive immune system is one of the most well characterized, yet complex biochemical systems in the animal body, but we still have much to learn about its design and how it functions [72]. One question is how the diversity of antibody repertoire is created. The diversity of the antibody repertoire is achieved by both combinatorial and junctional diversity, followed by somatic mutation. The large diversity of the antibody repertoire allows the immune system to recognize a wide variety of antigens. In each B cell precursor, one each of the many V, D, and J gene segments recombine to form a heavy chain. This process is called VDJ recombination. The B cell precursor becomes a mature naive B cell after negative selection and is released to the blood from the bone marrow. When the mature naive B cell binds to foreign antigen, it is activated and gives rise to plasma cells and memory cells. Therefore, the B cell antibody repertoire is mainly composed of two types of antibodies: the naive B cells that are not activated by antigen are called the naive antibody repertoire, and the plasma and memory B cells that have been under clonal selection are called activated antibody repertoire.

It is commonly believed that a heavy chain is generated by randomly combining V, D and J gene segments [72]. However, some studies have shown that V, D, and J genes may be not used equally in the pre-immune repertoire and that individual V gene segments rearrange at different frequencies [45, 155, 93]. One explanation for the unequal frequency of V segments is the natural variation in recombination signal sequences that are recognized by recombinase enzymes [46]. Previous work focuses on the individual gene segment usage, correlations of VDJ combinatorial usages between individuals have only recently been studied.

In this study, strong correlations in the zebrafish naive antibody repertoires are observed. The entire expressed VDJ repertoires from 14 zebrafish are used [130]. Previous study found limited correlations between fish antibody repertoires[130]. By classifying the whole antibody repertoire into naive antibody repertoire and activated antibody repertoire, we show that the naive antibody repertoire has a strong correlation, and the activated antibody repertoires has almost no correlation. We further propose a stochastic model of VDJ recombination in which each V, D and J segment are chosen at some frequency, which is conserved across individuals. Our results suggest that the VDJ recombination process is regulated.

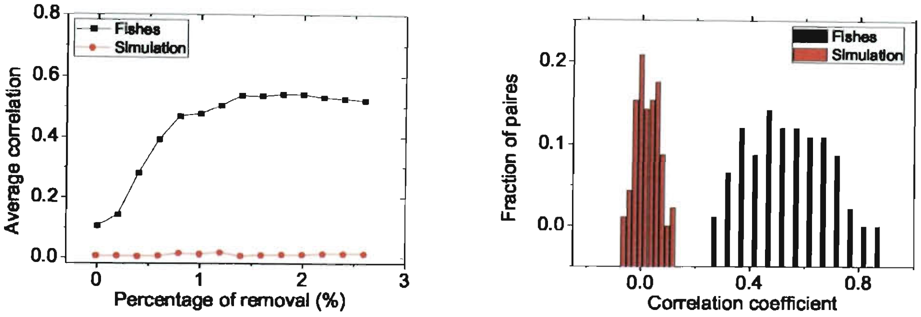
7.2 Results

7.2.1 Correlation in the naive VDJ repertoire

In zebrafish, there are 39 choices for the V segment, 5 for D and 5 for J, for a total of 975 possible VDJ combinations. Previous experiments have produced the complete antibody repertoire in each of 14 zebrafish [130]. The V, D and J segments of all the sequences are recognized by aligning the genomes. In this study, T^n_{ijk} is used to donate the sequence reads of VDJ combinations(type i V segement, type j D segment, and type

k J segment) of fish n. The naive antibody VDJ repertoire can be constructed from the entire antibody VDJ repertoire by removing the highly represented VDJ combinations. The inactive naive B cells have low copies. Once they are activated by antigen, they start to duplicate themselves and can increase in number by up to 1000 fold [72]. Considering that each B cell bears a single type of receptor with a unique VDJ combination, the highly expressed VDJ combinations in repertoire are very likely from the activated B cells.

Strong correlation between zebrafish antibody repertoire is observed when highly represented VDJ combinations are removed. T^n_{ijk} is ranked by their value and the top p percentage of VDJ combinations that have the most sequence reads are removed from the 975 possible VDJ combinations. The remaining VDJ repertoire has $975 * (1 - p)$ VDJ combinations and is a vector in which each element T^n_{ijk} records the number of reads that map to a particular VDJ combination. The Pearson correlation coefficient is calculated between the remaining VDJ repertoire vectors from different fish. The control experi-



(a) (b)

Figure 7.1 VDJ repertoire correlation analysis for all 14 fish. (a) Correlation of VDJ repertoire with highly expressed VDJ combinations removed. (b) Histogram of correlation of VDJ repertoire with the top 2% of most expressed VDJ combinations removed. In both figures, the simulation is conducted by randomly swapping the sequence reads between VDJ combinations.

ments are also constructed by randomizing the VDJ repertoire vectors. Fish were nearly uncorrelated in all VDJ repertoires when all VDJ combinations are considered in Figure 7.1(a). The correlation increases when the more highly expressed VDJ combinations are removed, saturating at maximal value of 0.52. The activated VDJ combinations are estimated to account for 1% $\sim$ 2% of entire 975 possible VDJ combinations. The sequence reads are added up for the top 2% mostly expressed VDJ combinations and find that the activated VDJ combinations account for 57% of all sequences reads. The activated VDJ combinations has a small diversity and a large amount of volume in the entire repertoire. Note that in the human repertoire, memory B cells have at a roughly 100 times higher copy number and a roughly 100 times lower diversity, compared with naive B cells. The results here suggest the same number may hold for zebrafish.

The correlation between fish is calcuated by removing the top 2% of VDJ combinations. The correlations are significantly larger than that for simulated control experiment. The correlations in the naive VDJ repertoire are unexpectedly high, which indicates the

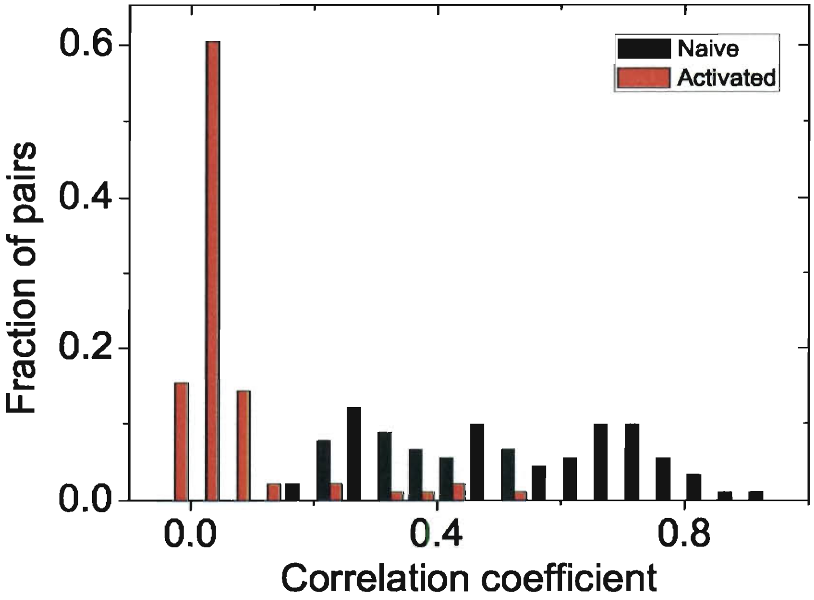


Figure 7.2 Naive and activated VDJ repertoire correlation analysis for all 14 fish.

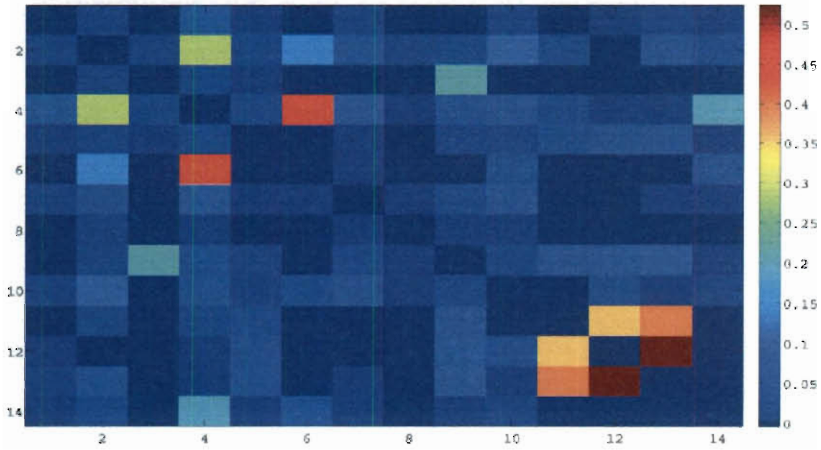


Figure 7.3 Correlation matrix of the activated VDJ repertoire. Only fish 4 and 6, fish 12 and 13 have strong correlations.

VDJ recombination is a regulated process. These results are in contrast with the common view that the VDJ repertoire is generated by a series of uniformly random molecular events within independent B cells [72].

Another method to distinguish naive antibody repertoire from activated repertoire is based on somatic hypermutation. When a B cell is activated by an antigen, it is stimulated to divide (or proliferate). During proliferation, the B cell receptor locus undergoes an extremely high rate of somatic mutation that is at least $10^5 - 10^6$ fold greater than the normal rate of mutation across the genome[90]. For each sequence read in ref.[130], we compare it with standard V, D, and J gene segments to find non-junctional mutations that are considered as somatic hypermutations. Sequence reads are grouped into a cluster if their sequences are identical. If a sequence cluster has no somatic hypermutation and there are fewer than 5 sequence reads in the cluster, this cluster is considered as naive antibody cluster. we impose a cutoff of 5 sequence reads in the cluster, so that activated B cells without somatic hypermuations are excluded. If sequence cluster has somatic hypermutations and there are more than one sequence reads in cluster, this cluster is considered as naive antibody cluster. we impose a cutoff of greater than one sequence reads to exclude the sequencing errors from somatic hypermuations. The naive VDJ repertoire $T^n_{ijk}(N)$ is calculated using sequence reads in naive antibody clusters. The activated VDJ repertoire $T^n_{ijk}(A)$ is calculated using sequence reads in activated antibody clusters. Further, correlation coefficients are calculated $C_{m,n}(N) =< (T^m(N), T^n(N) >$ and $C_{m,n}(A) =< (T^m(A), T^n(A) >$.

Strong correlations in the naive VDJ repertoires are observed in Figure 7.2. This is in agreement with Figure 7.1(b). Small correlations exist in activated VDJ repertoires. The activated antibody repertoire is developed under clonal selection that is correlated with the environment. Therefore the small correlations in activated VDJ repertoires can be explained that different fish see different environment and therefore their immune responses have different history [72]. Although most fish were uncorrelated in their activated VDJ repertoires, two pairs of fish are highly correlated in Figure 7.3. Detail examination of these two pairs of fish reveals that they are from the same family. It is possible that these two pairs of fish are living in the same environment and have a similar disease history, and so they developed a similar antibody repertoire.

7.2.2 Regulated model of VDJ recombination

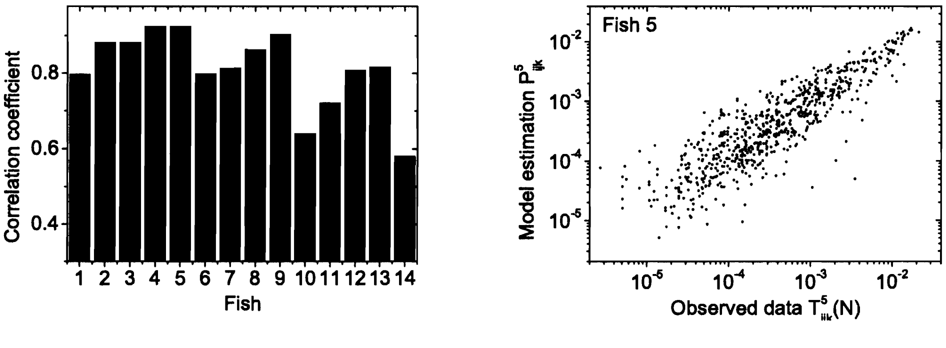
One question naturally arises is how the zebrafish immune system explores the large space of all possible VDJ combinations to find the tiny 1% $\sim$ 2% effective VDJ combinations to defect antigens, see Figure 7.1. A random search in the entire space seems costly and inefficiently. When an antigen invades, does the immune system have pre-designed pathes to search for VDJ combinations with the best affinity binding the antigen?

Here, we propose a statistical model to describe the mechanism of VDJ recombination and how the immune systems explore the large space of the antibody repertoire. We assume the frequency probability of a particular VDJ combination is the product of the frequency probability of its V, D and J segments.

$$P^n_{ijk} = P^n_i(V)P^n_j(D)P^n_k(J) \tag{7.1}$$

Here, P^n_{ijk} is the frequency of observing the VDJ combination of type "ijk" in fish n. $P^n_i(V), P^n_j(D)$ and $P^n_k(J)$ are the probability of selecting i type V, j type D and k type J during VDJ recombination in fish n, respectively. The value of $P^n_i(V), P^n_j(D)$ and $P^n_k(J)$ are estimated by fitting the model to the naive VDJ repertoire data by maximal likelihood Monte Carlo method. The initial value of probability of gene segments is chosen as $P^n_i(V)_{int} \propto \sum_{jk} T^n_{ijk}(N)$, $P^n_j(D)_{int} \propto \sum_{ik} T^n_{ijk}(N)$ and $P^n_k(J)_{int} \propto \sum_{ij} T^n_{ijk}(N)$. The value of $P^n_i(V), P^n_j(D)$ and $P^n_k(J)$ is perturbed to maximize the correlation between model prediction and real data as defined as $D^n =< P^n_{ijk}, T^n_{ijk}(N) >$

The model is calibrated with the zebrafish data. The model can produce the naive VDJ repertoire similar the original data with high fidelity. The model estimation fits well with the original data with high correlation in Figure 7.4(a). By directly comparing



(a) (b)

Figure 7.4 (a) Correlation D^n between naive VDJ repertoire $T^n_{ijk}(N)$ and model estimation P^n_{ijk}. (b) Fitting quality as illustrated by fish 5.

the model estimation and observed data in Figure 7.4(b), the simple model predicts the data. From these results, the diversity of naive VDJ repertoire is likely generated by choosing V, D and J segments with some probability to create a VDJ combination.

We find that the probability of choosing a particular V segment is the same across fish. In other words, all fish share the same probability distribution of V segments. The $P^n_i(V)$ with $n = 1, 2..., 14$, and $i = 1, 2, ..., 39$ is shown in Figure 7.5. Heterogenous usage of V segments are observed. Approximately half of V segments are heavily used and the

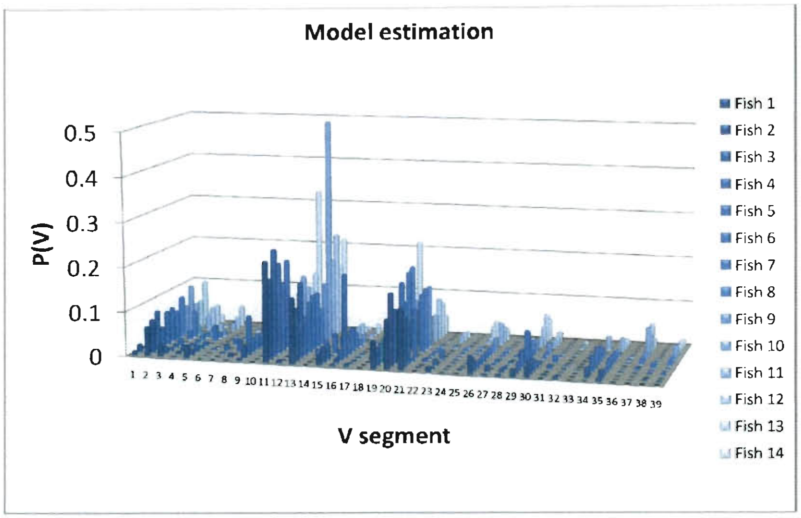


Figure 7.5 $P_i^n(V)$ as estimated from the model for all 39 V segments in 14 fish.

other half are infrequently observed. One interesting question is whether the heavily used V segments in one fish are also heavily used in other fish. Intuitively, this trend can be observed in Figure 7.5. For example, V segments of type 11, 13, 20, and 21 are the most frequently used segments in all fish. We determine that the probability distribution V segments is conserved across fish with a significant average correlation of $P^n(V)$ across fish ($R^2 = 0.57$). The average of $P_i^n(V)$ over 14 fish is shown in Figure 7.6.

The conserved probability frequency are also observed in J segments. The $P_k^n(J)$

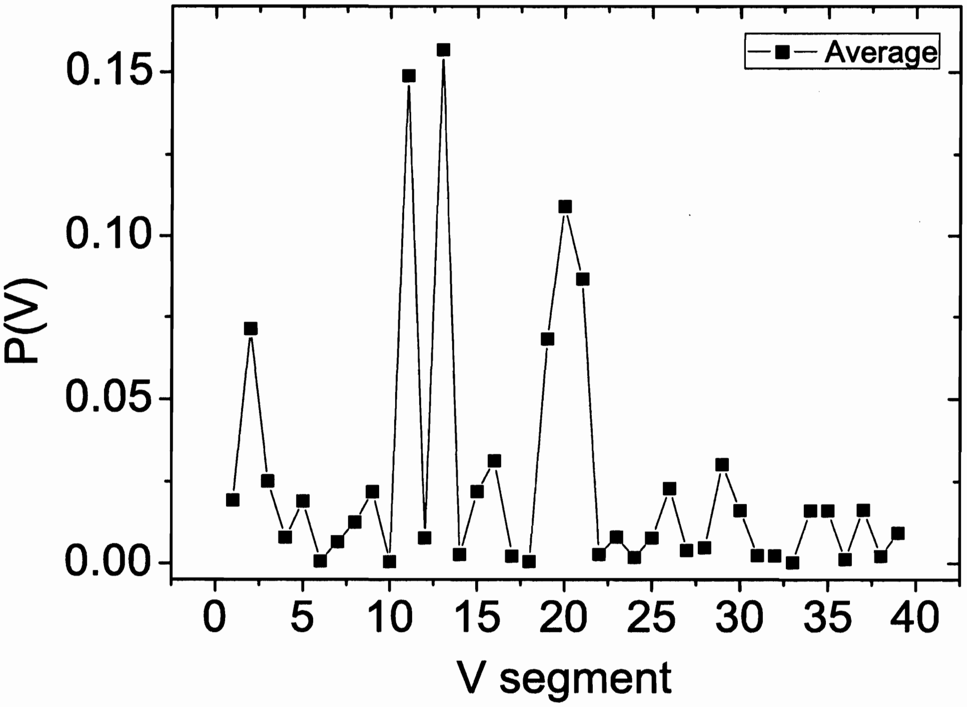


Figure 7.6 Average $P_i(V)$ over 14 fish.

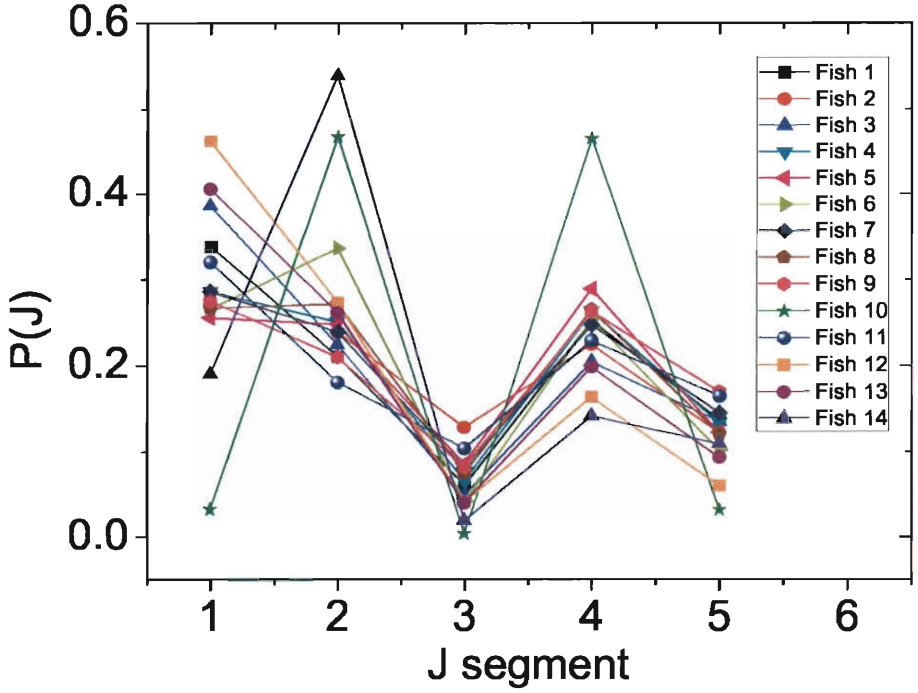


Figure 7.7 $P_k^n(J)$ as estimated from the model for all 5 J segments in 14 fish. Fish 10 and 14 have distinct probability distribution. Recall in Figure 7.4(a) that the model prediction fits the data of fish 10 and 14 less well than the other fish.

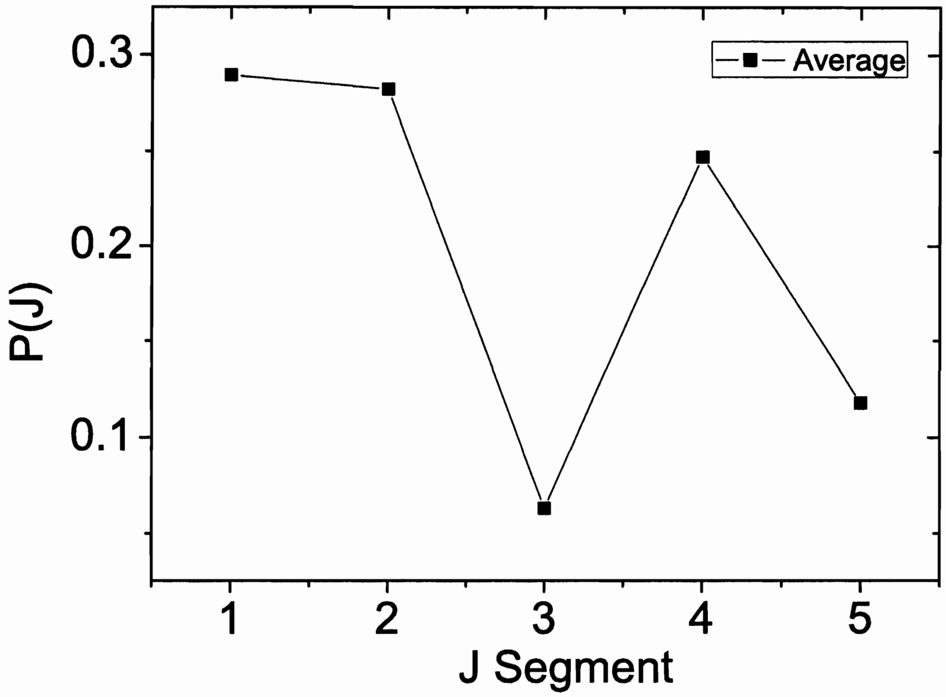


Figure 7.8 Average $P_k(J)$ over 14 fish.

is shown in Figure 7.7 and the average $P_k^n(J)$ over 14 fish in Figure 7.8. Except for fish 10 and 14, most of other fish share a common frequency distribution of J segments. These results signify that the VDJ recombination is a regulated process. Each V and J segment is chosen with at some frequency probability, and the probability is shared across individuals. The unequal probability distribution seems to be the result of evolution and it may help the immune system to quickly explore the large space of VDJ repertoire to find the best antibody for an external antigen.

7.3 Conclusion and discussion

In summary, we have observed a strong correlation in the naive VDJ repertoires. A simple statistical model is proposed to describe the mechanism of VDJ recombination. The diversity of the VDJ repertoire seems to be generated by a regulated VDJ recombination process, in which each V, D, and J segments is chosen with a regulated frequency.

The results here suggest several experiment to exam the diversity of the antibody repertoire. First, experiments can be performed to sequence bone marrow antibody repertoire. This would provide a window to study the mechanism of VDJ recombination. In our study, two computational methods are used to distinguish the naive antibody repertoire from the activated antibody repertoire. The naive antibody repertoire could be directly measured by sequencing antibodies in the bone marrow from mice or kidney from zebrafish [83]. Second, to understand how environment and disease affect the antibody repertoire, individuals from the same and different families and environments could be sequenced to investigate the possible feedback control from environment to VDJ recombination. Alternatively, an individual's antibody repertoire could be sequenced multiple times when its environment is changed. Third, we can sequence siblings' antibody repertoires at different ages to determine the genetic and epigenetic control of VDJ recombination.

Chapter 8
Conclusion

Evidence from protein interaction, protein domain interaction, animal body plan development, and world trade show that hierarchy will spontaneously emerge and grow in evolving systems in the changing environment. The theory of modularity is a general law in biology and will lead to new discoveries in biology. Hierarchy in evolving systems is shown to increase the evolvability and robustness of the systems.

Influenza has a high evolution rate, which makes vaccine design challenging. New dominant strains can be detected early by low-dimensional clustering. An influenza vaccine selection procedure is proposed based on this sequence clustering. The procedure is demonstrated and tested in detail using historical data. The performance of the method to predict the dominant H3N2 strain in an upcoming flu season is shown using data solely from before the flu season. The method was demonstrated on data since 1996. This strain detection tool would appear to be useful for annual influenza vaccine selection.

The CRISPR provide adaptable immunity to bacteria and archaea. The coevolution interaction between viruses and bacteria CRISPR system has shaped the spacer structure of CRISPR locus. Both the models and recent experiments support the declining diversity of spacers towards leader-distal end, implying that the CRISPR is an actively anti-viral system. The models explored that the underlying mechanism of shaping spacer structure is the selection of bacteria CRISPR systems that match best with viruses in the environment, and the diversity of bacteria CRISPR is vital for survive.

Naive VDJ repertoires is shown to have strong correlations in individuals. A simple statistical model is proposed to describe the mechanism of VDJ recombination. The diversity of the VDJ repertoire seems to be generated by a regulated VDJ recombination process in which each V, D, and J segments is choose with a regulated frequency.

References

1. Northern hemisphere: Risk of A/Wellington/1/2004(H3N2)-like virus. ProMed. 2004 October 24. Available from http://www.promedmail.org, archive no. 20041024.2879.

2. Seasonal influenza (H3N2) virus - potential vaccine mismatch. ProMed. 2009 July 24. Available from http://www.promedmail.org, archive no. 20090724.2623.

3. Preliminary assessment of the effectiveness of the 2003-04 inactivated influenza vaccine–colorado, december 2003. *MMWR Morb Mortal Wkly Rep*, 53:8–11, 2004.

4. Virology quarterly report Jul-Sep 2004. 2004. http://www.surv.esr.cri.nz/virology/virology_quarterly_report.php.

5. Influenza Weekly Update. 22-38, 2005. http://www.surv.esr.cri.nz/virology/influenza_weekly_update.php.

6. Trade, exchange rates, budget balances and interest rates. *Economic and financial indicators*, May 2nd, 2009.

7. T. Abeysinghe and K. Forbes. Trade linkages and output-multiplier effects: a structural var approach with a focus on asia. *NBER Working paper series*, 2001.

8. A. Alesina, E. Spolaore, and R. Wacziarg. *Trade, growth and the size of countries*, chapter 23, pages 1499–1542. Handbook of economic growth. Elsevier, first edition, 2005.

9. U. Alon. Network motifs: Theory and experimental approaches. *Nat. Rev. Genet.*, 8:450–461, 2007.

10. J. Anderson. *Measurement of protection.* 2009. Prepared for the Palgrave Handbook of International Trade.

11. A. F. Andersson and J. F. Banfield. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. *Science*, 320(5879):1047–1050, 2008.

12. G. Apic, J. Gough, and S. A. Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, 310:311–325, 2001.

13. W. Arthur. *The origin of animal body plans: a study in evolutionary development biology.* Cambridge University Press, Cambridge, 2000.

14. Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. The influenza virus resource at the national center for biotechnology information. *J. Virol.*, 82:596–601, 2008.

15. A. Barabsi and Z. N. Oltvai. Network biology: Understanding the cells functional organization. *Nat. Rev. Genet.*, 5:101–113, 2004.

16. M. Barigozzi, G. Fagiolo, and D. Garlaschelli. Multinetwork of international trade: A commodity-specific analysis. *Phys. Rev. E*, 81:046104, 2010.

17. M. Baron, D. G. Norman, and L. D. Campbell. Protein modules. *TIBS*, 16:13–17, 1991.

18. R. Barrangou et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*, 315(5819):1709–1712, 2007.

19. L. D. Bogarad and M. W. Deem. A hierarchical approach to protein molecular evolution. *Proc. Natl. Acad. Sci. USA*, 96:2591–2595, 1999.

20. A. Bolotin, B. Quinquis, A. Sorokin, and S. Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151:2551, 2005.

21. S. D. Bowring and D. H. Erwin. A new look at evolutionary rates in deep time: Uniting paleontology and high-precision geochronology. *GSA Today*, 8:1–8, 1998.

22. S. J. J. Brouns et al. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*, 321(5891):960–964, 2008.

23. I. D. Campbell and M. Baron. The structure and function of protein modules. *Phil. Trans, Biol. Sci.*, 332:165–170, 1991.

24. V. Carvalho. Aggregate fluctuation and the network structure of intersectoral trade. *CREI working paper*, 2009.

25. J. Chen. *The dawn of animal world.* Jiangsu Science and Technology Publishing House, Nanjing, 2004.

26. K. Chen and N. Rajewsky. The evolution of gene regulation by transcription factors and micrornas. *Nature Reviews Genetics*, 8:93–103, 2007.

27. A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

28. N. Cox and C. Bender. The molecular epidemiology of influenza viruses. *Seminars in virology*, 6:359–370, 1995.

29. J. A. Coyne. Comment on gene regulatory networks and the evolution of animal body plans. *Science*, 313:761b, 2006.

30. R. M. Cripps and E. N. Olson. Control of cardiac development by an evolutionarily conserved transcriptional network. *Development Biology*, 246:14–28, 2002.

31. M. E. Csete and J. C. Doyle. Reverse engineering of biological complexity. *Science*, 295:1664–1669, 2002.

32. E. H. Davidson and D. H. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311:796, 2006.

33. M. Deem and K. Pan. The epitope regions of H1-subtype influenza A, with application to vaccine efficacy. *Protein Engineering, Design & Selection*, 22:543–546, 2009.

34. M. W. Deem. Adventures in mathematical biology. *Physics Today*, 60:42–47, 2007.

35. H. Deveau et al. Phage response to CRISPR-encoded resistance in streptococcus thermophilus. *J. Bacteriol.*, 190(4):1390–1400, 2008.

36. E. Domingo, J. Holland, and C. Biebricher. *Quasispecies and RNA virus evolution: Principles and consequences.* Landes, Austin, TX, 2002.

37. J. W. Drake and J. J. Holland. Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. USA*, 96:13910–3, 1999.

38. D. J. Earl and M. W. Deem. Evolvability is a selectable trait. *Proc. Natl. Acad. Sci. USA*, 101:11531–11536, 2004.

39. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

40. R. J. R. Elliott and K. Ikemoto. AFTA and the Asian crisis: Help or hindrance to ASEAN intra-regional trade? *Asian Economic Journal*, 18(1):1–23, 03 2004.

41. L. Elnitski, V. Jin, P. Farnham, and S. Jones. Locating mammalian transcrip-

tion factor binding sites: A survey of computational and experimental techniques. *Genome Res*, 16:1455–1464, 2006.

42. D. H. Erwin and E. H. Davidson. The evolution of hierarchical gene regulatory networks. *Nat. Rev. Genetics*, 10:141–148, 2009.

43. B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis.* Oxford Univ. Press, 2001.

44. FDA. Regulatory considerations regarding the use of novel influenza A (H1N1) virus vaccines. July 23, 2009.

45. A. Feeney. Genetic and epigenetic control of v gene rearrangement frequency. *Adv. Exp. Med. Biol.*, 650:73–81, 2009.

46. A. Feeney, A. Tang, and K. Ogwaro. B-cell repertoire formation: Role of the recombination signal sequence in non-random V segment utilization. *Immunol. Rev.*, 175:59–69, 2000.

47. N. Ferguson, A. Galvani, and R. Bush. Ecological and immunological determinants of influenza evolution. *Nature*, 422:428, 2003.

48. W. M. Fitch, R. M. Bush, C. A. Bender, and N. J. Cox. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15):7712–7718, 1997.

49. C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. M. E. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, and T. W. R. P. A. Collaboration. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324(5934):1557–1561, 2009.

50. T. Friedman. *The world is flat: A brief history of the twenty-first century.* Farra,

Straus and Giroux, 2005.

51. A. Gardener and W. Zuidema. Is evolvability involved in the origin of modular variation? *Evolution*, 57:1448–1450, 2003.

52. R. J. Garten, C. T. Davis, C. A. Russell, B. Shu, S. Lindstrom, A. Balish, W. M. Sessions, X. Xu, E. Skepner, V. Deyde, M. Okomo-Adhiambo, L. Gubareva, J. Barnes, C. B. Smith, S. L. Emery, M. J. Hillman, P. Rivailler, J. Smagala, M. de Graaf, D. F. Burke, R. A. M. Fouchier, C. Pappas, C. M. Alpuche-Aranda, H. Lopez-Gatell, H. Olivera, I. Lopez, C. A. Myers, D. Faix, P. J. Blair, C. Yu, K. M. Keene, J. Dotson, P. David, D. Boxrud, A. R. Sambol, S. H. Abid, K. St. George, T. Bannerman, A. L. Moore, D. J. Stringer, P. Blevins, G. J. Demmler-Harrison, M. Ginsberg, P. Kriner, S. Waterman, S. Smole, H. F. Guevara, E. A. Belongia, P. A. Clark, S. T. Beatrice, R. Donis, J. Katz, L. Finelli, C. B. Bridges, M. Shaw, D. B. Jernigan, T. M. Uyeki, D. J. Smith, A. I. Klimov, and N. J. Cox. Anti-

genic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science*, 325(5937):197–201, 2009.

53. E. Ghedin, N. A. Sengamalay, M. Shumway, J. Zaborsky, T. Feldblyum, V. Subbu, D. J. Spiro, and J. Sitz. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, 437:1162–1166, 2005.

54. S. Gupta, N. Ferguson, and R. Anderson. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science*, 280:912, 1998.

55. V. Gupta, D. J. Earl, and M. Deem. Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, 24:3881–3888, 2006.

56. E. Hak, J. Nordin, F. Wei, J. Mullooly, S. Poblete, R. Strikas, and K. Nichol. Influence of high-risk medical conditions on the effectiveness of influenza vaccination among elderly members of 3 large managed health care organizations. *Clinical Infectious Diseases*, 35(4):370–377, 2002.

57. L. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:47–52, 1999.

58. L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:47–52, 1999.

59. J. He and M. W. Deem. Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Phys. Rev. Lett.*, 105:128102, 2010.

60. J. He and M. W. Deem. Hierarchical evolution of animal body plans. *Developmental Biology*, 337:157–161, 2010.

61. J. He and M. W. Deem. Low-dimensional clustering detects incipient dominant influenza strain clusters. *Protein Engineering, Design & Selection*, 23:935–946, 2010.

62. J. He and M. W. Deem. Structure and response in theworld trade network. *Phys. Rev. Lett.*, 105:198701, 2010.

63. J. He, J. Sun, and M. Deem. Spontaneous emergence of modularity in a model of evolving individuals and in real networks. *Phys. Rev. E*, 79:031907, 2009.

64. S. Hedges, J. Blair, M. Venturi, and J. Shoe. A molecular timescale of eukaryote evolution and the rise of complex multicelluar life. *BMC Evolutionary Biology*, 4:2, 2004.

65. S. B. Hedges. The origin and evolution of model organisms. *Nature*, 3:838–849, 2002.

66. J. F. Heidelberg, W. C. Nelson, T. Schoenfeld, and D. Bhaya. Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS ONE*, 4(1):e4169, 01 2009.

67. C. Hidalgo, B. Klinger, A. Barabasi, and R. Hausmann. The product space conditions the development of nations. *Science*, 317:482–487, 2007.

68. A. E. Hirsh, H. B. Fraser, and D. P. Wall. Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Molecular Biology and Evolution*, 22:174–177, 2005.

69. A. E. Hirsh, H. B. Fraser, and D. P. Wall. Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol. Biol. Evol.*, 22:174–177, 2005.

70. P. Horvath and R. Barrangou. CRISPRCas, the immune system of bacteria and archaea. *Science*, 327:167, 2010.

71. P. Horvath et al. Diversity, activity and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.*, pages JB.01415–07, 2007.

72. C. Janeway, P. Travers, M. Walport, and M. Shlomchik. *Immunobiology.* Garland Science, New York, 2005.

73. R. Jansen, J. Embden, W. Gaastra, and L. Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, 43:1565, 2002.

74. R. Kaesler and J. Cairns. Cluster analysis of data from limnological surveys of the upper potomac river. *American Midland Naturalist*, 88:56, 1972.

75. F. Karginov and G. Hannon. The CRISPR system: Small RNA-guided defense in bacteria and archaea. *Molecular Cell*, 37:7, 2010.

76. N. Kashtan and U. Alon. Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci. USA*, 102:13773–13778, 2005.

77. H. Kitano. Biological robustness. *Nature Reviews Genetics*, 5:826–837, 2004.

78. G. Koop, M. Pesaran, and S. Potter. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, 74:119–147, 1996.

79. A. Lapedes and R. Farber. The geometry of shape space: Application to influenza. *Journal of Theoretical Biology*, 212(1):57 – 69, 2001.

80. M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA*, 102:4936–4942, 2005.

81. H. Lipson, J. B. Pollack, and N. P. Suh. On the origin of modular variation. *Evolution*, 56:1549–1556, 2002.

82. L. A. Marraffini and E. J. Sontheimer. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science*, 322(5909):1843–1845, 2008.

83. N. Meeker and N. Trede. Immunology and zebrafish: Spawning new models of human disease. *Development and Comparative Immunology*, 32:745–757, 2008.

84. D. Misevic, C. Ofria, and R. E. Lenski. exual reproduction reshapes the genetic architecture of digital organisms. *Proc. Roy. Soc. B*, 273:457–464, 2006.

85. M. I. Nelson and E. C. Holmes. The evolution of epidemic influenza. *Nature Reviews Genetics*, 8:196–205, 2007.

86. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

87. S. Ng, Z. Zhang, and S. Tan. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19:923–929, 2003.

88. S. Ng, Z. Zhang, S. Tan, and K. Lin. InterDom: A database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucl. Acids Res.*, 31:251–254, 2003.

89. M. Nowak and R. May. *Virus Dynamics: The Mathematical Foundations of Immunology and Virology.* Oxford Univ. Press, 2000.

90. V. Odegard and D. Schatz. Targeting of somatic hypermutation. *Nature Reviews Immunology*, 6:573, 2006.

91. P. Oikonomou and P. Cluzel. Effects of topology on network evolution. *Nature Physics*, 2:532–536, 2006.

92. K. Pan, K. Subieta, and M. Deem. Quantify seasonal H1N1 influenza vaccine efficacy and antigenic distance. 2009. Submitted.

93. R. Perlmutter, J. Kearney, S. Chang, and L. Hood. Developmentally controlled expression of immunoglobulin VH genes. *Science*, 227:1597–1601, 1985.

94. J. B. Plotkin, J. Dushoff, and S. A. Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6263–6268, 2002.

95. C. Pourcel, G. Salvignol, and G. Vergnaud. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tool for evolutionary studies. *Microbiology*, 151:653, 2005.

96. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierar-

chical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

97. S. D. Reid, C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam. Parallel evolution of virulence in pathogenic escherichia coli. *Nature*, 406:64–67, 2000.

98. R. Riedl. *Order in living organisms: A systems analysis of evolution.* Wiley, New York, 1978.

99. M. Ronshaugen, N. McGinnis, and W. McGinnis. Hox protein mutation and macroevolution of the insect body plan. *Nature*, 414:914–917, 2002.

100. C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, J. C. de Jong, A. Kelso, A. I. Klimov, T. Kageyama, N. Komadina, A. S. Lapedes, Y. P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A. D. Osterhaus, G. F. Rimmelzwaan, M. W. Shaw, E. Skepner,

K. Stohr, M. Tashiro, R. A. Fouchier, and D. J. Smith. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26(Supplement 4):D31 – D34, 2008.

101. C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, J. C. de Jong, A. Kelso, A. I. Klimov, T. Kageyama, N. Komadina, A. S. Lapedes, Y. P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A. D. M. E. Osterhaus, G. F. Rimmelzwaan, M. W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R. A. M. Fouchier, and D. J. Smith. The global circulation of seasonal influenza a (H3N2) viruses. *Science*, 320(5874):340–346, 2008.

102. C. Santamaria, A. Urue, C. Videla, and *et al.* Epidemiological study of influenza virus infections in young adult outpatie: Results. *Influenza Resp Viruses*, 2:131–134, 2008.

103. Y. Satou and N. Satoh. Gene regulatory networks for the development and evolution

of the chordate heart. *Genes Dev.*, 20:2634–2638, 2006.

104. J. A. Shapiro. A 21st century view of evolution: genome system architecture, repetitive DNA, and genetic engineering. *Gene*, 345:91–100, 2004.

105. J. A. Shapiro. Retrotransposons and regulatory suites. *BioEssays*, 27:122–125, 2005.

106. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation networks of *Escherichia coli. Nature Genetics*, 31:64–68, 2002.

107. H. A. Simon. The architecture of complexity. *Proc. Amer. Phil. Soc.*, 106:467–482, 1962.

108. D. Skowronski. Influenza A (H1N1) - worldwide (11): Coincident H3N2 variation. ProMed. 2009 May 5. Available from http://www.promedmail.org, archive no. 20090505.1679.

109. A. Smith. Fossil evidence for the relationships of extant echinoderm classes and their times of divergence. In *Echinoderm Phylogeny and Evolutionary Biology*, pages 85–97. Oxford University Press, Oxford, 1988.

110. D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

111. G. Smith, D. Vijaykrishna, J. Bahl, S. Lycett, M. Worobey, O. Pybus, S. Ma, C. Cheung, J. Raghwani, S. Bhatt, J. Peirisl, Y. Guan, and A. Rambaut. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. *Nature*, 459:1122, 2009.

112. Y. Sobolevsky and E. N. Trifonov. Conserved sequences of prokaryotic proteomes and their compositional age. *J. Mol. Evol.*, 61:591–596, 2005.

113. J. Sun and M. W. Deem. Spontaneous emergence of modularity in a model of

evolving individuals. *Phys. Rev. Lett.*, 99:228107, 2007.

114. M. Suyama, D. Torrents, and P. Bork. Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*, 34:W609–W612, 2006.

115. S. Tanase-Nicola, P. B. Warren, and P. R. ten Wolde. Signal detection, modularity, and the correlation between extrinsic and intrinsic noise in biochemical networks. *Phys. Rev. Lett.*, 97:068102, 2006.

116. R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.

117. J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–4680, 1994.

118. E. N. Trifonov and I. N. Berezovsky. Molecular evolution from abiotic scratch. *FEBS Lettters*, 527:1–4, 2002.

119. E. N. Trifonov and I. N. Berezovsky. Evolutionary aspects of protein structure and folding. *Curr. Opinion. Struct. Biol.*, 13:110–114, 2003.

120. E. N. Trifonov, A. Kirzhner, V. M. Kirzher, and I. N. Berezovshy. Distinct stages of protein evolution as suggested by protein sequence analysis. *J. Mol. Evol.*, 53:394–401, 2001.

121. J. L. Turner, J. E. Fielding, H. J. Clothier, and H. A. Kelly. Influenza surveillance in victoria, 2005. *Communicable Diseases Intelligence*, 30(1):137, 2006.

122. G. W. Tyson and J. F. Banfield. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environmental Microbiology*, 10:200–207, 2008.

123. J. Valentine. *On the Origin of Phyla.* Univ. of Chicago Press, Chicago., 2004.

124. E. A. Variano, J. H. McCoy, and H. Lipson. Networks, dynamics and modularity. *Phys. Rev. Lett.*, 92:188701, 2004.

125. F. H. Veronica, T. N. Albert, R. A. Cameron, and E. H. Davidson. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc. Natl. Acad. Sci. USA*, 23:13356–13361, 2003.

126. G. P. Wagner and L. Altenberg. Complex adaptations and the evolution of evolvability. *Evolution*, 50:967–976, 1996.

127. C. Walcott. Cambrian geology and paleontology. In *Smithsonian Miscellaneous Collections*, page 14. 1914.

128. I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–64, 2007.

129. R. G. Webster. Influenza: an emerging disease. *Emerging Infectious Diseases*, 4(3):436–441, 1998.

130. J. A. Weinstein, N. Jiang, I. White, Richard A., D. S. Fisher, and S. R. Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324:807–810, 2009.

131. W. Wimsatt. Developmental constraints, generative entrenchment, and the innate–acquired distinction. In *Intergrating Scientific Disciplines*, pages 185–208. Dordrecht: Martinus-Nijhoff, 1986.

132. W. Wimsatt and J. Schank. Generative entrenchment, modularity and evolvability: When genic selection meets the whole organism. In *Modularity in development and evolution*, pages 359–394. U. Chicago Press, Chicago, 2004.

133. World Health Organization. *Wkly. Epidemiol. Rec.*, 70:53, 1995.

134. World Health Organization. *Wkly. Epidemiol. Rec.*, 71:57, 1996.

135. World Health Organization. *Wkly. Epidemiol. Rec.*, 72:57, 1997.

136. World Health Organization. *Wkly. Epidemiol. Rec.*, 73:56, 1998.

137. World Health Organization. *Wkly. Epidemiol. Rec.*, 74:57, 1999.

138. World Health Organization. *Wkly. Epidemiol. Rec.*, 75:61, 2000.

139. World Health Organization. *Wkly. Epidemiol. Rec.*, 76:57, 2001.

140. World Health Organization. *Wkly. Epidemiol. Rec.*, 77:57, 2002.

141. World Health Organization. *Wkly. Epidemiol. Rec.*, 78:57, 2003.

142. World Health Organization. *Wkly. Epidemiol. Rec.*, 79:85, 2004.

143. World Health Organization. *Wkly. Epidemiol. Rec.*, 80:65, 2005.

144. World Health Organization. *Wkly. Epidemiol. Rec.*, 80:341, 2005.

145. World Health Organization. *Wkly. Epidemiol. Rec.*, 81:81, 2006.

146. World Health Organization. *Wkly. Epidemiol. Rec.*, 82:69, 2007.

147. World Health Organization. *Wkly. Epidemiol. Rec.*, 83:77, 2008.

148. World Health Organization. *Wkly. Epidemiol. Rec.*, 84:65, 2009.

149. World Health Organization. Pandemic (H1N1) 2009 - update 64, 2009 [cited 2010 January 27]. Available from http://www.who.int/csr/disease/swineflu/updates/en/index.html.

150. World Health Organization. Recommendations for influenza viruses, 2009 [cited 2010 January 27]. Available from http://www.who.int/csr/disease/influenza/vaccinerecommendations/en/index.html.

151. World Health Organization. *Wkly. Epidemiol. Rec.*, 85:81, 2010.

152. World Health Organization. World now at the start of 2009 influenza pandemic, June 4, 2009. Available from http://www.who.int/mediacentre/news/statements/2009/.

153. G. Wray. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genetics*, 8:206–216, 2007.

154. G. Wray, J. Levinton, and L. Shapiro. Molecular evidence for deep precambrian divergences among metazoan phyla. *Science*, 274:568–573, 1996.

155. G. Yancopoulos, S. Desiderio, M. Paskind, J. Kearney, D. Baltimore, and F. Alt. Preferential utilization of the most JH-proximal VH gene segments in pre-B-cell lines. *Nature*, 311:727–733, 1984.

156. K. B. Zeldovich, P. Chen, B. E. Shakhnovich, and E. I. Shakhnovich. A first-principles model of early evolution: emergece of gene families, species, and preferred protein folds. *PLoS Comput. Biol.*, 3:e139, 2007.

157. H. Zhou, R. S. Pophale, and M. W. Deem. Computer-assisted vaccine design. In Q. Wang and Y. J. Tao, editors, *Influenza: Molecular Virology*, chapter 10. Caister Academic Press, 2010.