ON THE THEORIES OF BIOMOLECULAR INTERACTIONS AND THEIR APPLICATIONS IN INTRINSIC PROTEIN DISORDER AND BACTERIAL SPORE GERMINATION

by

Jintao Liu

B.S. in Physics, University of Science and Technology of China, 2004

M.S. in Physics, University of Pittsburgh, 2006

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Jintao Liu

It was defended on

October 18, 2011

and approved by

Carlos J. Camacho, PhD, Associate Professor

Rob Coalson, PhD, Professor

James R. Faeder, PhD, Associate Professor

Hanna Salman, PhD, Assistant Professor

Paul Shepard, PhD, Professor

David Snoke, PhD, Professor

Dissertation Advisors: Carlos J. Camacho, PhD, Associate Professor

and Hanna Salman, PhD, Assistant Professor

ON THE THEORIES OF BIOMOLECULAR INTERACTIONS AND THEIR APPLICATIONS IN INTRINSIC PROTEIN DISORDER AND BACTERIAL SPORE GERMINATION

Jintao Liu, PhD

University of Pittsburgh, 2011

Living organisms are complex systems, where complexity arises in part from the large number of interacting components. Here I address interactions in two topics: intrinsically disordered proteins (IDP) and bacterial spore germination. In the first part, I study the role of intrinsic protein disorder in protein function with a standard thermodynamic model. IDPs are proteins without stable structure in their native states. Their ubiquitous presence undercuts the traditional view that a protein's structure determines its function. Here I propose a quantitative theory that makes predictions regarding the role of intrinsic disorder in protein structure and function. By relating disorder with the free energy of folding, I show that both catalytic and low-affinity binding proteins prefer ordered structures, whereas high-affinity binding proteins can tolerate disorder. Relevant to both transcription and signal transduction, the theory also explains how increasing disorder can tune the binding affinity to maximize the specificity of promiscuous interactions. These claims are supported by a genome-wide survey of disorder. Collectively, the study provides insights into how natural selection acts on folding stability to optimize protein function. In the second part, I study the mechanism of the initiation of bacterial spore germination and propose a quantitative model. Spores are formed by some species of gram positive bacteria (e.g., *Bacillus* and *Clostridium*) during starvation. They are metabolically dormant and can later germinate into vegetative cells when nutrients (called germinants)

reappear. The lag time of germination after encountering germinants is highly heterogeneous for spores in the same population, and the mechanism is still unclear. Here I propose a quantitative model based on the assumption that the heterogeneity is due to the variability in levels of activated germinant receptors (GR) per spore. The model produces predictions that are consistent with experiments on germination with mixtures of nutrients that trigger different types of GRs, which also suggests that signals from different GRs are summed by a common integrator.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

1.0 INTRODUCTION

Physicists have long been interested in studying biological systems. A famous example is Erwin Schrödinger's 1944 book *What is Life?* [1]. The book barely included any mathematics – an essential tool for physicists. This was partly because it was aimed for the general audience and partly because biology was "much too involved to be fully accessible to mathematics" [1]. After all, large parts of live organisms were still mysterious, such as the physical nature of heredity – the main topic of Schrödinger's book. Since then much progress has been made. We now have detailed knowledge on the inner workings of many organisms [2, 3], and physicists have played an important role in this process by inventing experimental techniques and proposing quantitative theories [4, 5]. Yet, there are still plenty of puzzles to solve and I can say this is an exciting time for physicists to jump into biology.

Living organisms are complex systems, where complexity arises, in part, from the large number of interacting components. Here I address interactions on the molecular level (Fig. 1.1) in two topics: intrinsically disordered proteins (IDP) and bacterial spore germination. The common goal is to understand the effect of these interactions on higher levels. In both cases, the networks of interactions are unknown, thus excluding the possibility of studying them based on the details of the network. For IDPs, I take the deductive approach: starting from the thermodynamics of protein folding and biomolecular interactions, I deduce the relation between protein disorder and its function and then go on to explain the finding that proteins of different

functions have different propensities for disorder. While it is a simple theory, it is not a trivial application of physics, mostly because biology also has its own principles and is not reducible to physics. I combined thermodynamics with concepts such as efficiency and specificity of biomolecular interactions, which are alien to physics but native to biology. For bacterial spore germination, I take the abductive approach, start from experimental observations, propose hypotheses, implement them in quantitative models, and validate them by comparing model predictions with new experiments. While no law of physics was used, I hope that readers will see the influence of physics, i.e., the striving for simple and elegant quantitative theories that explain and unify real world phenomena. I have benefited greatly from the pioneering works of others. Yet I also faced many challenges, since theoretical studies on both topics are still in the exploration stage.

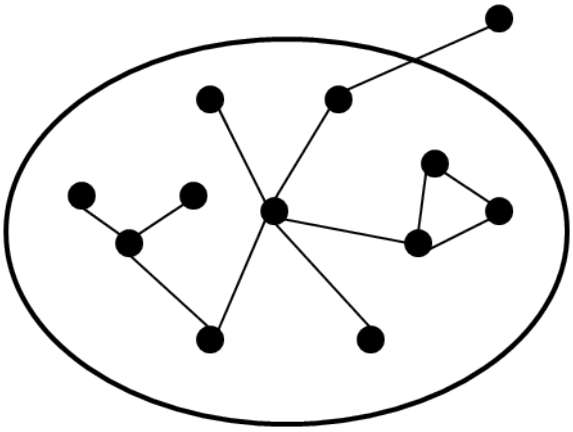

Figure 1.1: Schematic diagram of the network of interactions between the molecules inside and outside a cell. The filled circles represent molecules, the straight lines represent interactions between the molecules, and the ellipse represents the boundary of a cell.

This dissertation is organized as follows. In Chapter 2, I will introduce IDP, including historical trends of protein research, characteristics and thermodynamics of IDP, and computer predictions of disorder. In Chapter 3, I will present a genome wide survey on the relation between disorder and protein function and a thermodynamic theory explaining this relation. This work has been published and appears in condensed form in Ref. [6]. In Chapter 4, I will introduce bacterial spores and their germination, including the structure and properties of spores, measurement and stages of germination. In Chapter 5, I will present a model of the mechanism of germination and validation of this model with experimental data from the laboratory of our collaborators. Part of this work has been published and appears in Refs. [7, 8]. In Chapter 6, I will summarize my key findings and discuss the open questions.

2.0 INTRINSICALLY DISORDERED PROTEINS

2.1 INTRODUCTION

Proteins are one of the fundamental elements of life. They perform functions such as catalyzing chemical reactions, regulating gene expressions, binding ligands, transducing signals, and so on [3, 9]. They are linear polymers made of the twenty standard amino acids. The traditional view is that proteins, in their active form, have stable three dimensional structures which are determined by their amino acid sequences (Fig. 2.1), and the structures in turn encode their functions [9]. This is called the *Sequence-Structure-Function* paradigm and it is supported by the numerous protein structures that have been solved [10]. The emphasis of structure in protein research is clearly illustrated by what was written in the preface of the book *Introduction to Protein Structure* [11]: "The fundamental tenet of molecular biology, namely that one cannot really understand biological reactions without understanding the structure of the participating molecules, is at last being vindicated". However, in recent years, many proteins have been found to be without stable structure in their native states, and they are called intrinsically disordered proteins (IDP) [12-14]. Their ubiquitous presence undercuts the principle that a protein's structure determines its function [13]. Yet the sequence-structure-function paradigm is still the

standard in textbooks and IDPs receive minimal to no coverage at all [9, 11]. To understand this heavy bias, here we take a brief tour of the historical ideas.

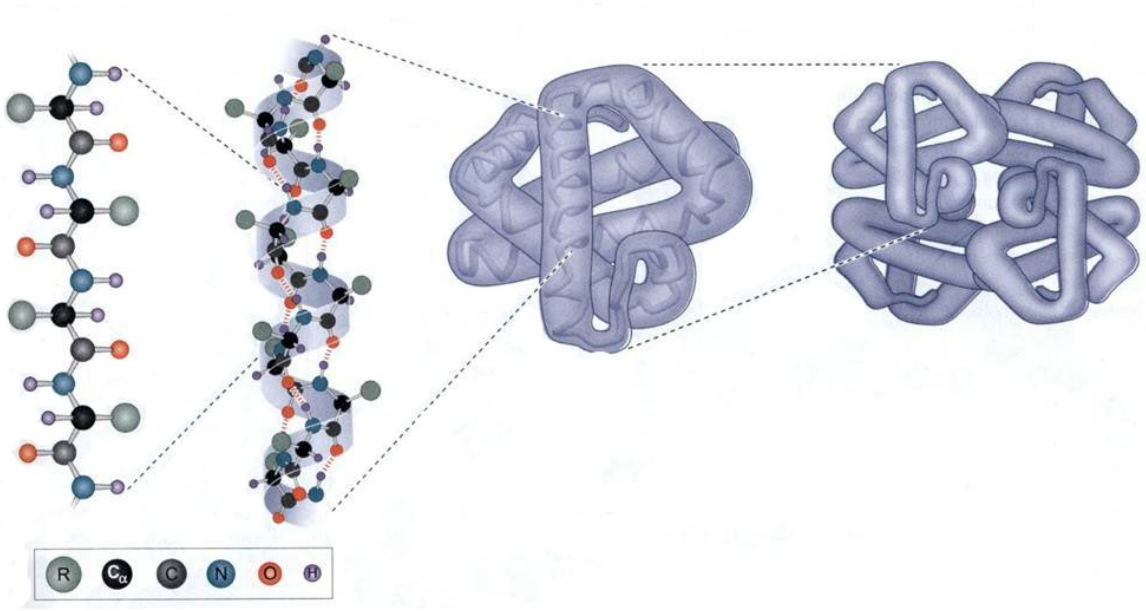

Figure 2.1: Four levels of protein structure.

A protein's amino acid sequence consists the primary structure. Amino acids within the same protein interact with each other and with water molecules and fold in sequence specific ways, where the local fold is called secondary structure and the global fold is called tertiary structure. Some proteins may form complexes with others, and the binding configuration is called quaternary structure. The labels are: R – side-chain group of an amino acid, C_α – α-carbon atom, C – carbon, N – nitrogen, O – oxygen, and H – hydrogen. (Source: Adapted from Ref. [2], p. 75, Fig. 5-7.)

Long before the determination of the first protein structure in 1958 [15], it has been speculated that proteins have well defined structures [16, 17]. It was known that enzymes, proteins that catalyze chemical reactions, are highly specific in selecting their substrates, i.e., an enzyme that works with one molecule is typically irresponsive to others even when they have similar structures and chemical properties. To explain this specificity, Emil Fischer proposed in

1894 the key-lock theory [16, 17], and claimed that the catalytic site of an enzyme must have a surface highly complementary to that of its cognate substrate, so that it does not fit to other molecules (Fig. 2.2A). To emphasize the complementarity of protein-substrate interface, proteins were perceived as rigid molecules.

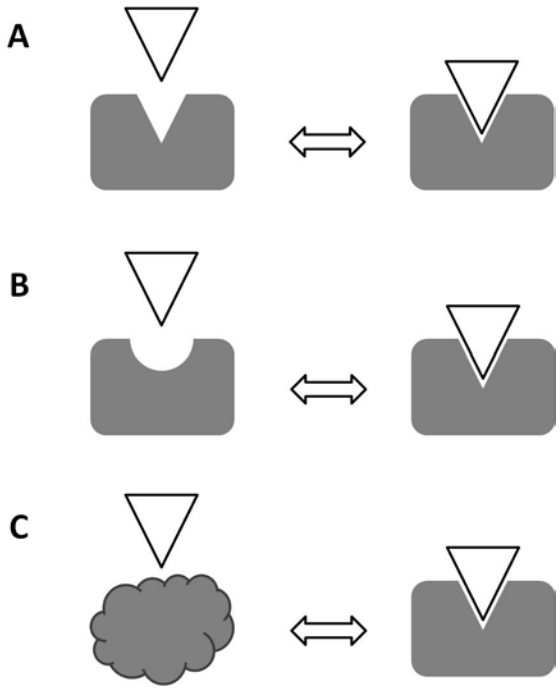

Figure 2.2 Schematic diagram of protein-ligand interaction.

Proteins are drawn as filled blocks and ligands are drawn as empty triangles. Rectangular blocks denote folded proteins and cloud shaped block denotes IDP in unfolded state. Shown are (A) the key-lock theory, (B) induced fit, and (C) induced folding.

The key-lock theory was widely accepted and generalized to all proteins [18] until D. E. Koshland pointed out in 1958 that it was insufficient to explain all the properties of enzymes [19]. For example, some enzymes could be turned on or off by the binding of regulatory molecules to locations distant from their catalytic sites. To resolve the crisis, Koshland suggested a revised explanation and called it “induced fit” theory (Fig. 2.2B). The theory assumes that

intermolecular interactions may cause appreciable changes in the three-dimensional relations of the amino acids at the active site and bring the catalytic groups into the proper orientation for reaction. The theory did not gain much acceptance initially, and a referee even wrote “the Fischer Key-Lock theory has lasted 100 years and will not be overturned by speculation from an embryonic scientist” [20]. But it eventually became the new norm, especially given that conformational changes are often seen in protein structures [11].

From history, we see a trend from viewing proteins as static objects to dynamic ones, which were no doubt influenced by the advancement of experiments. As more and more protein structures are solved, many proteins have been found to contain segments that could not be resolved in their structures, which sometimes can even be a significant fraction of a protein [21]. But those observations were usually ignored without understanding their functional significance, except in a minority of cases [21-25]. After all, great successes have been achieved by studying structured parts of proteins. Nowadays we are convinced that the induced fit theory only describes a subset of the existing proteins, and there are IDPs which do not fold on their own but can form stable structures when binding with their cognate partners (Fig. 2.2C) [26].

At the time of writing more than 68,000 experimentally determined protein structures have been deposited to the database Protein Data Bank (PDB) [10]. In contrast, there are only 643 entries in the most comprehensive database on experimentally characterized IDPs – DisProt [27]. This imbalance is not due to the rare occurrence of disorder among proteins; on the contrary, bioinformatics studies using disorder prediction techniques showed that disorder is ubiquitous, and the proportion of proteins containing long disordered regions increases with the increasing complexity of an organism [6, 12, 14, 28].

2.2 CHARACTERISTICS OF DISORDERED PROTEINS

Proteins fall onto a structural continuum, from tightly folded, to compact but flexible, and to highly extended and unstructured (Fig. 2.3) [26, 29]. Tightly folded proteins can usually be studied using X-ray crystallography, and it is found that they often also contain disordered regions, which either appear as segments missing from the electron-density map or are intentionally removed to enable crystallization [30]. A survey of the PDB database found that only ~7% of the deposited structures contain the complete amino acid sequences of the corresponding proteins, and only ~25% contain >95% of their full sequences [31]. Highly disordered proteins fail attempts of crystallization, and thus cannot be studied with X-ray crystallography. Instead, they can be studied with Nuclear Magnetic Resonance (NMR). NMR studies provide direct evidence on the existence of disorder in tightly folded proteins and the existence of highly disordered proteins [32]. They also show that disordered regions fold into stable structures upon binding to their cognate substrates (Fig. 2.3), which suggest that disorder can play important functional roles [26, 33].

In addition to X-ray crystallography and NMR, IDPs can also be studied using other techniques, including but not limited to Circular Dichroism, hydrodynamic measurements, fluorescence spectroscopy, as well as Raman spectroscopy [34]. It has been found that IDPs generally are: resistant to boiling temperature that cause ordered proteins to precipitate; insensitive to chemicals that cause ordered proteins to denature; and susceptible to proteolytic cleavage. A comprehensive review on the properties of IDPs can be found in Ref. [35].

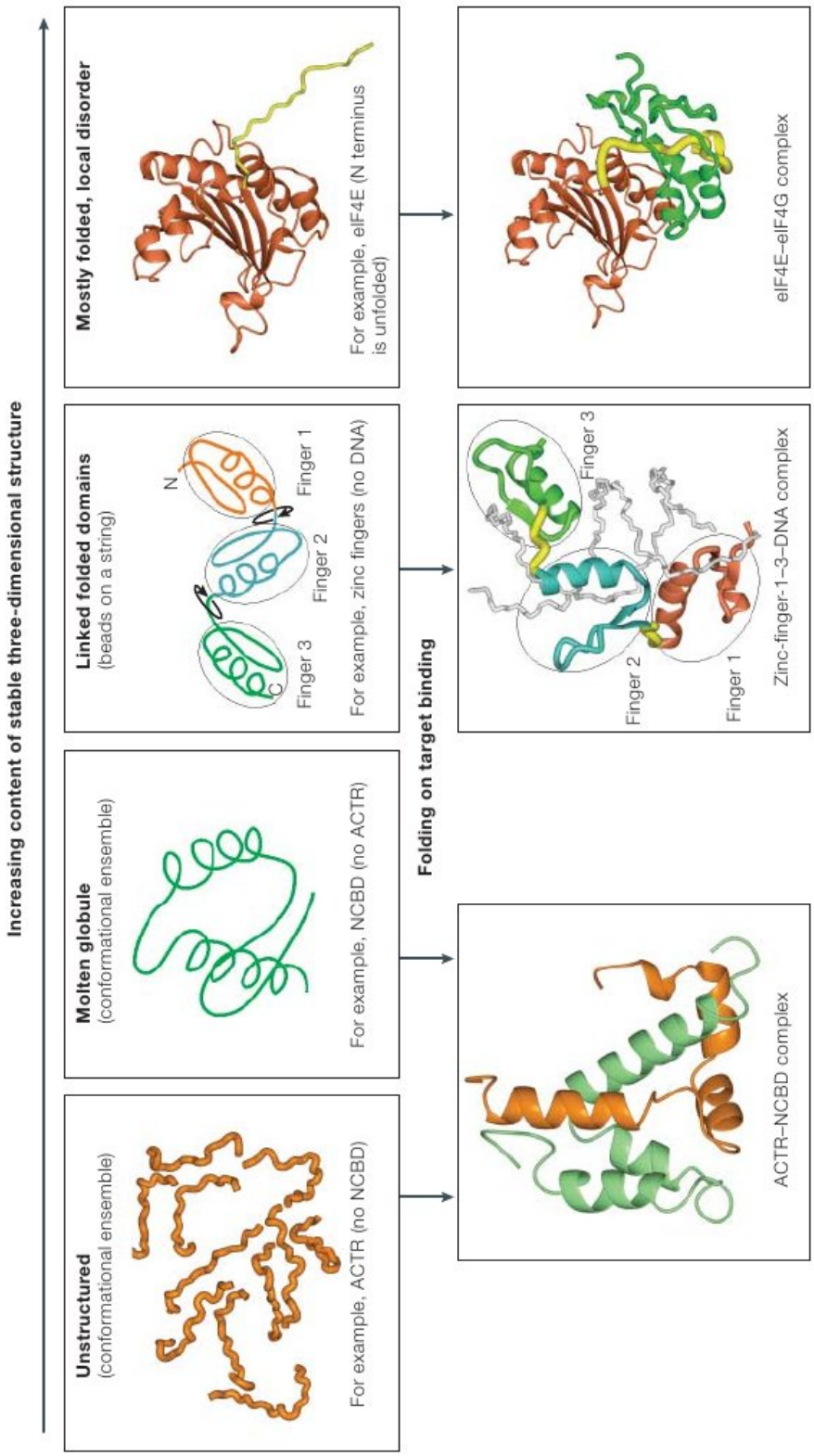

Figure 2.3: The continuum of protein structure.

(Source: Adapted from Ref. [26])

2.3 THERMODYNAMICS OF IDP

Proteins are usually studied under constant temperature and pressure, therefore the Gibbs free energy is the relevant thermodynamic potential. Figure 2.4 illustrates the free energy landscapes of proteins. At equilibrium, the most likely configuration of a protein is the one with the lowest free energy. Ordered proteins fold stably because their folded structure have significantly lower free energies than other possible configurations. In contrast, disordered proteins do not have configurations with significantly lower free energies than others, thus the lack of stable structure. However, their free energy landscapes change when interacting with their binding partners, which enable them to form stable structures, and different partners may induce different structures on the same IDP.

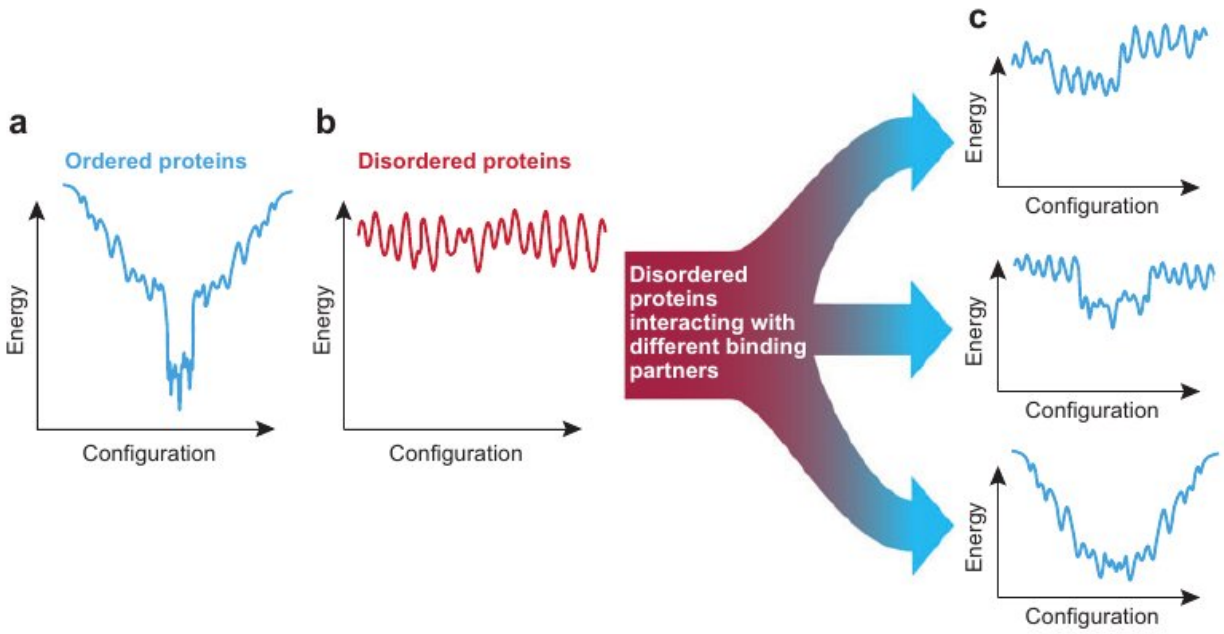

Figure 2.4: Diagram for the free energy landscapes of proteins.

Free energy landscapes of (a) a typical ordered protein and of a typical IDP in the (b) absence or (c) presence of different binding partners. The landscapes are depicted schematically in 1-D cross-sections. (Source: Adapted from Ref. [36].)

It is useful to group the configurations of an IDP into two states: If the IDP assumes the configuration when it is bound with its partner, we say it is in folded state; otherwise, it is in unfolded state. At equilibrium and in the absence of binding partners, the ratio between the probabilities for the folded and unfolded states is given by Boltzmann distribution

$$\frac{p_{\text{folded}}}{p_{\text{unfolded}}} = e^{-\Delta G_f/RT} \tag{2.1}$$

where ΔG_f is the free energy difference between folded and unfolded states, R is the ideal gas constant, and T is absolute temperature. For IDPs, $\Delta G_f > 0$ so that the unfolded state is more stable than the folded state; for ordered proteins, $\Delta G_f < 0$. In the next chapter, I will deal with the case when binding partners are present.

2.4 COMPUTER PREDICTION OF DISORDER

The difference between disordered and ordered proteins originates from their amino acid sequences. Just as sequence determines the structure of a protein, it also determines which part of a protein is disordered. This principle is the basis of all the existing algorithms predicting disorder. The first algorithm was developed by Dunker and colleagues in 1997 [37, 38]. Since then more than 50 disorder predictors have been developed. Most of them are similar in the prediction of long disordered regions, but differ in the local details of the outputs [39]. These tools enable us to study disorder in proteins that have not been experimentally characterized and to perform genome-wide studies on the functional significance of disorder. Here I introduce three of the predictors used in this work, and assessments of other predictors can be found in a number of reviews [35, 39, 40].

(1) PONDR VSL2 [41]. PONDR is the abbreviation for Predictors of Natural Disordered Regions [42], which is a family of predictors specialized for different purposes respectively, such as on disordered regions of different "flavors" [43] or lengths [37]. VSL2 is a general purpose predictor that applies to Variously characterized, Short and Long disordered regions [41]. It is a linear support vector machine [44] that was trained with 1,327 non-redundant protein sequences.

(2) FoldIndex [45]. Uversky et al found that the known list of IDPs and ordered proteins can be distinguished from the empirical formula $\langle R\rangle = 2.785\langle H\rangle - 1.151$ [46], where $\langle R\rangle$ is the average net charge per residue of a protein at pH 7.0, and $\langle H\rangle$ is the average hydrophobicity per residue [47]. In $\langle R\rangle$-$\langle H\rangle$ plots, IDPs scatter in the region above the line described by the formula, and ordered proteins scatter in the region below the line [46]. This agrees with the fact that high net charge leads to strong repulsion between the residues within the same protein and high hydrophobicity leads to greater tendency to form compact structures in water, and the folding of a protein is determined by these two competing factors. The disorder predictor FoldIndex was designed based on this idea [45].

(3) DisEMBL. This predictor is based on artificial neural networks trained for predicting three definitions of disorder [48]: a) Loops/coils as defined by the DSSP (Define Secondary Structure of Proteins) algorithm [49], a standard method for assigning secondary structure to the amino acids of proteins. Note that disorder is only found within loops, but loops/coils are not necessarily disordered. b) Hot loops, loops with a high degree of mobility as determined from B factor (Debye–Waller factor) in X-ray crystallography [50], which describes the attenuation of x-ray scattering caused by thermal motion. c) Missing coordinates in X-Ray structure. Since none of these definitions alone can give reliable predictions of order/disorder, it is recommended that

they should be combined and use the Loop predictor only as a filter to remove false disorder predictions of the other two [40].

In this dissertation, I will be mainly using VSL2, as both itself and its predecessor VSL1 [51] were evaluated as the highest ranked in the CASP7 and CASP6 (Critical Assessment of Techniques for Protein Structure Prediction) assessments respectively [52, 53]. In addition, I will also use FoldIndex and DisEMBL so that conclusions drawn from the predictions by VSL2 are guaranteed to be general rather than predictor specific.

2.5 OUTSTANDING QUESTIONS

While a large number of proteins are intrinsically disordered, the origins of this disorder are not well understood, and its ubiquitous presence undercuts the principle that a protein's structure determines its function. It has been suggested that disorder itself plays a functional role by, e.g., allowing for multiple interaction partners [54] and functional diversity [28, 55, 56], which are particularly important in cell signaling and cancer [57]. The correlation between intrinsic disorder and protein function, however, is still nebulous. These motivated me to look for the general principles that might link protein function and disorder.

In Chapter 3, I will present a quantitative theory that makes predictions regarding the role of intrinsic disorder in protein structure and function. In particular, I will discuss the implications of analytical solutions of a series of fundamental thermodynamic models of protein interactions in which disordered proteins are characterized by positive folding free energies. Without assuming any a priori structure-function relationship, the theory predicts that both catalytic and low-affinity binding proteins prefer ordered structures, whereas only high-affinity binding

proteins (found mostly in eukaryotes) can tolerate disorder. Relevant to both transcription and signal transduction, the theory also explains how increasing disorder can tune the binding affinity to maximize the specificity of promiscuous interactions. The predictions are validated by performing genome-wide surveys of disorder in both prokaryotic and eukaryotic genomes. Collectively, the study provides insight into how natural selection acts on folding stability to optimize protein function.

3.0 TOWARD A QUANTITATIVE THEORY OF INTRINSICALLY DISORDERED PROTEINS AND THEIR FUNCTION

3.1 GENOME-WIDE SURVEY OF DISORDER

Genome-wide surveys of protein disorder have shown that disorder is more prevalent in some functional categories than others [28, 55]. I revisit this question by analyzing the fraction of amino acid residues in disordered regions of both eukaryotic and prokaryotic genomes for different functional categories.

The genomes studied are: human (*Homo sapiens*), mouse (*Mus musculus*), zebrafish (*Danio rerio*), chicken (*Gallus gallus*) and *Arabidopsis thaliana* from the Swiss-Prot database [58]; yeast (*Saccharomyces cerevisiae*) from the Saccharomyces Genome Database [59]; *Escherichia coli* (K-12) from EcoCyc and EcoliHub [60]; rice (*Oryza sativa*) from the Gramene database [61]; fruit fly (*Drosophila melanogaster*) from FlyBase [62]; *Caenorhabditis elegans* from WormBase [63]; *Dictyostelium discoideum* from dictyBase [64]; *Schizosaccharomyces pombe* from the *Schizosaccharomyces pombe* GeneDB database [65]; *Bacillus anthracis* and *Pseudomonas fluorescens* from the TIGR database [66].

For each protein, the percentage of disordered amino acids was estimated by using the VSL2B predictor [41], which was trained with experimental data by using machine learning techniques and validated in comprehensive blind experiments. The predictor uses the amino acid

sequences of proteins (Fig. 3.1) as input and gives the probability that each amino acid (also called residue) is in a disordered region (Fig. 3.2), from which the percentage of disordered residues in a protein is calculated.

```
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

Figure 3.1: Amino acid sequence of the protein p53 in human.

Each letter represents one amino acid and this protein has 393 residues. (Source: UniProt [58].)

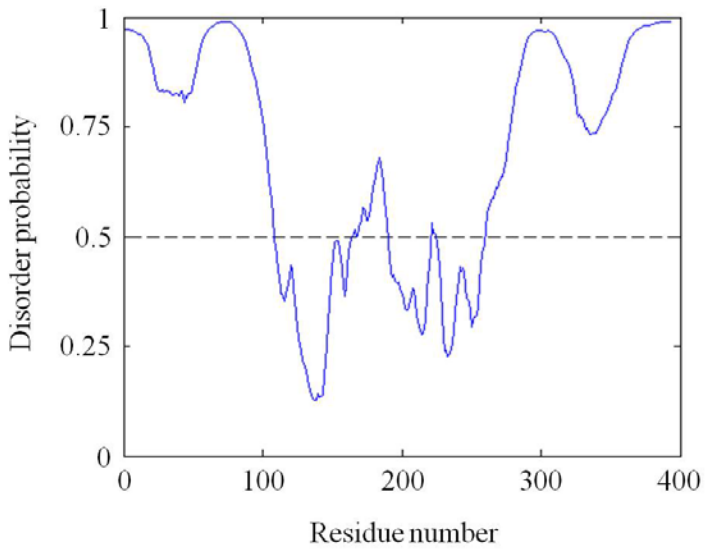

Figure 3.2: Disorder probability for the residues of human p53 protein given by VSL2B.

A probability greater than 0.5 predicts a residue to be disordered.

Figure 3.3 shows the distributions of the amount of disorder in human, yeast, and *E. coli* proteins as predicted by the disorder predictor VSL2B (also shown are the distributions after

removing proteins with more than one function; see also Fig. 3.4 for predictions of two other predictors FoldIndex [45] and DisEMBL [48]). To assign protein function, I use the *gene ontology* classification [67], in which *protein binding*, *catalytic activity*, and *transcription regulator activity* are the three largest functional categories. Contrary to the striking bias of *catalytic* and *transcription* human proteins to be significantly more ordered and disordered, respectively, disorder is neither strongly favored nor disfavored in *binding* proteins. These distinctions are still visible in yeast but are less obvious in bacterial genomes such as *E. coli*, whose proteins are found to be significantly more ordered than those found in eukaryotes across all functional categories.

Based on a more comprehensive analysis of the preference of disorder among the different functional categories, I classify the genomes into three types (Fig. 3.5): (type I) no strong preference for ordered structures in *binding* proteins but preference for disorder in *transcription* proteins, among which are human, mouse, zebrafish, chicken, rice, fruit fly, *A. thaliana*, and *D. discoideum*; (type II) no strong preference for ordered structures for either *binding* or *transcription* proteins, among which one finds yeast, *S. pombe*, and *C. elegans*; and (type III) strong preference for ordered structures in both *binding* and *transcription* proteins, among which there are *E. coli*, *B. anthracis*, and *P. fluorescens*. For *catalysis*, all genomes show a strong preference for ordered proteins. Note that prokaryotic genomes are all type III, whereas eukaryotes are either type I or II, with type I genomes being generally larger than type II.

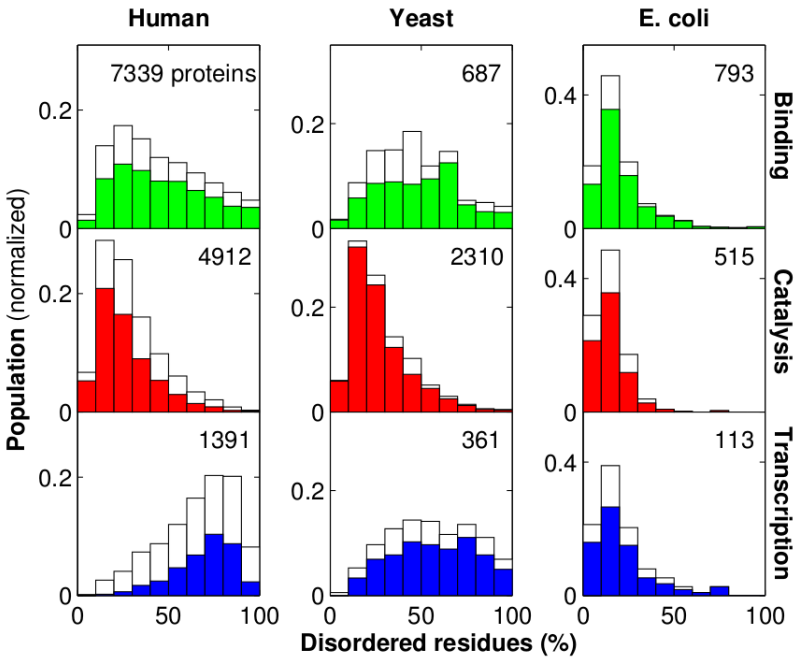

Figure 3.3: Disorder distributions predicted by VSL2B.

Normalized histograms of the percentage of disordered residues predicted by VSL2B in the sequence of human (*H. sapiens*), yeast (*S. cerevisiae*) and *E. coli* (K-12) proteins within the gene ontology categories of *protein binding*, *catalytic activity*, and *transcription regulator activity*. The distributions after removing the overlap between the three categories are shown by the lower bars (shaded). All distributions are normalized to the total number of proteins in each category noted in the upper right corner of each frame. In humans, contrary to the bias of transcription and catalytic proteins to be significantly more disordered and ordered, respectively, binding proteins indicate that disorder is neither strongly favored nor disfavored. The statistical significance of these results, based on a Kolmogorov–Smirnov test [68], is $P < 10^{-150}$. In yeast, although binding and catalytic proteins show the same trend as occurs in higher eukaryotes, transcription proteins overall show no significant preference for order or disorder. In *E. coli*, all three functions show strikingly similar distributions favoring ordered structures. Similar distributions were found in other eukaryotic and prokaryotic genomes.

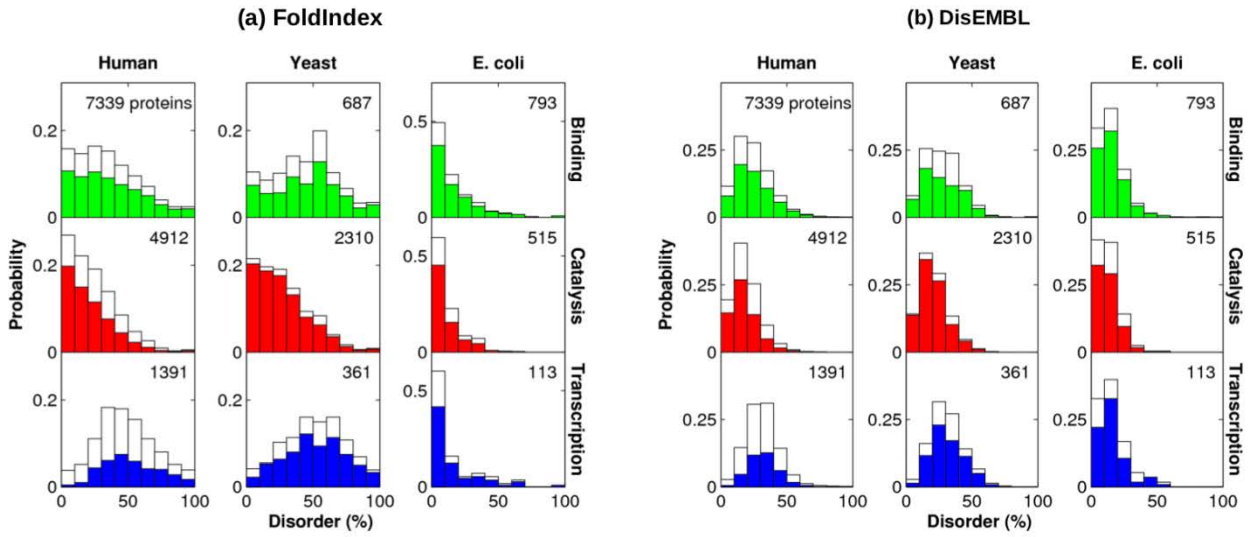

Figure 3.4: Disorder distribution predicted by FoldIndex and DisEMBL.

Normalized histograms of the percentage of disordered residues in the sequence of human, yeast (*S. cerevisiae*) and *E. coli* (K-12) proteins within the Gene Ontology categories of *Protein Binding, Catalytic Activity* and *Transcription Regulator Activity* using (a) FoldIndex and (b) DisEMBL respectively. The distributions after removing the overlap between the three categories are shown by the lower bars (shaded). These two predictors show similar biases as in Fig. 3.3, but with the caveat of consistently under-predicting disorder with respect to the VSL2B predictor. Note that FoldIndex is specialized in predicting regions that have low hydrophobicity and high net charge, and DisEMBL is specialized in predicting highly mobile loops and regions lacking electron density in crystal structures [40].

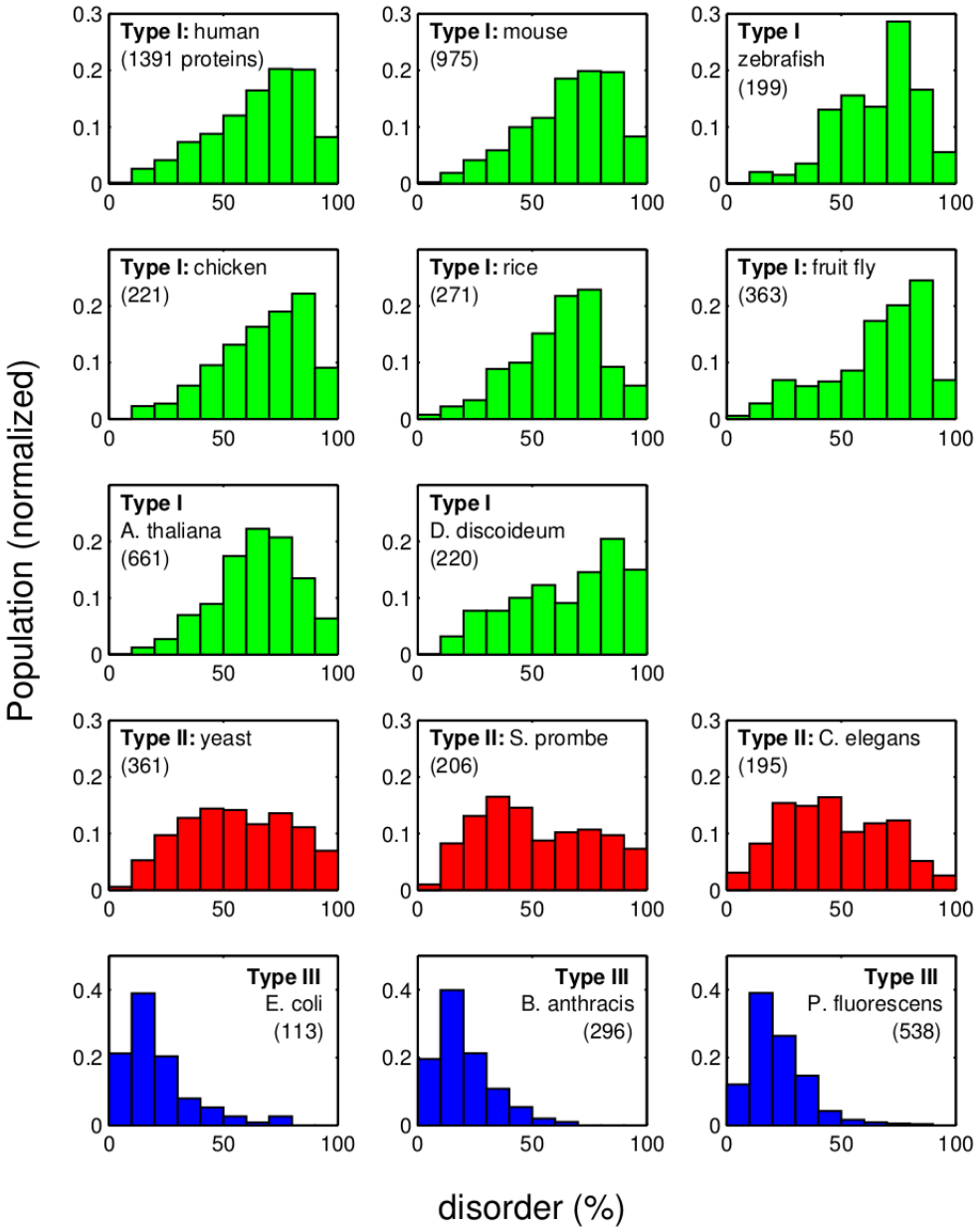

Figure 3.5: Disorder distribution of Transcriptional proteins in different genomes.

The genomes are classified into three groups: Type I, no strong preference for ordered structures in *Binding* proteins, but preference for disorder in *Transcription*; Type II, no strong preference for ordered structures for either *Binding* or *Transcription*; Type III, strong preference for ordered structures in both *Binding* and *Transcription*; For *Catalysis*, all genomes show a strong preference for ordered proteins.

3.2 THERMODYNAMIC MODEL

The analysis in Section 3.1 suggests that selection pressures act on protein disorder to optimize particular aspects of protein function, raising the question of what universal properties may have driven proteins involved in *Binding*, *Catalysis*, and *Transcription* to evolve along different pathways. Here I show that a simple thermodynamic model of molecular interactions can elucidate the role of disorder in binding and catalysis.

As described by Dyson and Wright [26], proteins in the cellular environments may have disorder in long loops, end terminals, hinge regions, domains, and even covering their full sequences (Fig. 2.3). However, in a complex, these motifs acquire well-defined 3D structures. Common descriptors to all these forms of disorder are the folding free energy (ΔG_f) of the motifs participating in the molecular interaction and the dissociation constant (K_d) of the interaction, where a positive folding free energy corresponds to a disordered protein [22]. In this model, folding is defined as a two-state equilibrium between the unfolded state (U) and the folded state (F) [69]. Thus, the ratio of the concentrations of folded to unfolded proteins that are unbound is given by

$$[\mathrm{F}]_{\mathrm{eq}}/[\mathrm{U}]_{\mathrm{eq}} = e^{-\Delta G_f/RT} \tag{3.1}$$

where "eq" denotes equilibrium, R is the ideal gas constant and T is absolute temperature. Molecular interactions are described by a simple binding model that assumes that only folded proteins bind the substrate (Fig. 3.6, conformational selection [70]), i.e.,

$$\begin{array}{ccc} \mathrm{F} & + \mathrm{S} \overset{K_d^c}{\rightleftharpoons} & \mathrm{FS} \\ {\scriptstyle \Delta G_f}\uparrow\downarrow & & \\ \mathrm{U} & & \end{array} \tag{3.2}$$

where K_d^c is the binding affinity between F and S

$$K_d^c \equiv [\mathrm{F}]_{\mathrm{eq}}[\mathrm{S}]_{\mathrm{eq}}/[\mathrm{FS}]_{\mathrm{eq}} \tag{3.3}$$

which implicitly accounts for the effects of interface area, shape, hydrogen bonds, and other interactions, hence I will call it complementary affinity. Note that the size of the interface provides a natural upper bound on the number of contacts contributing to the interaction. In this sense, higher complementarity is often associated with a large interface, although in some cases it can be caused by other factors (e.g., small-molecule drugs often have binding affinities between 10^{-9} to 10^{-12} molar, 1 molar (M) = 1 mol/L). K_d^c is equivalent to the experimental binding affinity between the protein (irrespective of whether it is in the U or F state) and its substrate $K_d^{\mathrm{exp}} \equiv \left([\mathrm{U}]_{\mathrm{eq}} + [\mathrm{F}]_{\mathrm{eq}}\right)[\mathrm{S}]_{\mathrm{eq}}/[\mathrm{FS}]_{\mathrm{eq}}$ if the protein is ordered ($[\mathrm{U}]_{\mathrm{eq}} \ll [\mathrm{F}]_{\mathrm{eq}}$). On the other hand, if the protein is disordered ($[\mathrm{U}]_{\mathrm{eq}} \gtrsim [\mathrm{F}]_{\mathrm{eq}}$), then using Eq. 3.1 we have

$$K_d^{\mathrm{exp}} = K_d^c\left(1 + e^{\Delta G_f/RT}\right) \tag{3.4}$$

Note that in this formulation, K_d^c characterizes the strength of the binding interaction for the folded protein and is independent of the folding free energy ΔG_f, allowing for a clear distinction between binding and folding.

Aside from conformational selection [70], disordered proteins could also function through induced folding (Fig. 3.6) [13, 71] or a combination of the two [72]. However, as will be demonstrated later, the conclusions do not lose generality because I only rely on equilibrium or steady state properties. For each functional category, I relate a measure of optimal performance to ΔG_f over the range of parameters found in nature. With the exception of *transcription*, where further discussion is needed, I will show that this general model accounts for the observed distributions in Figs. 3.3 and 3.4 if one assumes that natural selection acts on ΔG_f to optimize protein function. In the following, I discuss the key relations between folding stability and function.

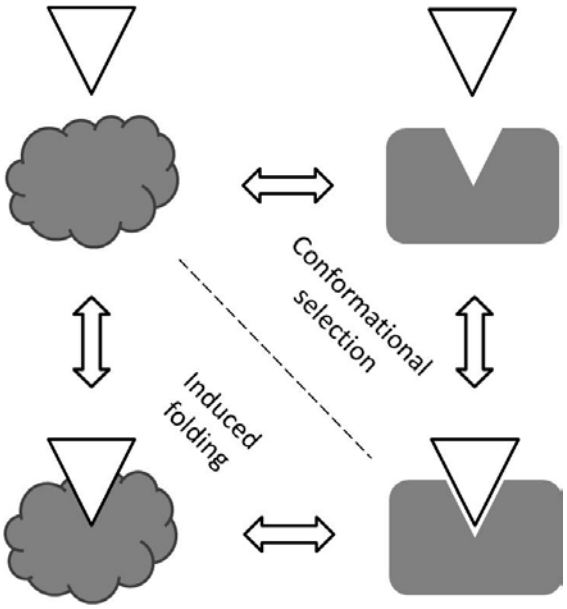

Figure 3.6: Schematic diagram of folding and binding of IDP.

Proteins are drawn as filled blocks and ligands are drawn as empty triangles. Cloud shaped and rectangular blocks denote IDP in unfolded and folded states respectively. The upper and lower pathways are called conformational selection and induced folding respectively.

For *binding* proteins, combine Eqs. 3.1, 3.3, and 3.4, the equilibrium complex concentration is given by

$$[\mathrm{FS}]_{\mathrm{bind}} = \frac{1}{2}\left[\left(c_p + c_s + K_d^{\mathrm{exp}}\right) - \sqrt{\left(c_p + c_s + K_d^{\mathrm{exp}}\right)^2 - 4c_p c_s}\right] \tag{3.5}$$

where $c_p \equiv [\mathrm{U}] + [\mathrm{F}] + [\mathrm{FS}]$ and $c_s \equiv [\mathrm{S}] + [\mathrm{FS}]$ are the total protein and substrate concentration, respectively. At given K_d^c, $[\mathrm{FS}]_{\mathrm{bind}}$ reaches a maximum $[\mathrm{FS}]_{\mathrm{bind}}^{\mathrm{max}}$ if $\Delta G_f \ll 0$. The curves in Fig. 3.7A show the ratio $[\mathrm{FS}]_{\mathrm{bind}}/[\mathrm{FS}]_{\mathrm{bind}}^{\mathrm{max}}$ as a function of folding free energy (ΔG_f), in the absence of excess protein or substrate ($c_p = c_s = 1$ μM). This ratio defines a measure of the efficiency of protein binding to produce maximum amount of complex. For the physiologically relevant range of K_d^c between 10^{-5} and 10^{-10} M, a binding efficiency of, say,

90% or higher, is obtained for folding-stability thresholds of $\Delta G_f < -1.2$ kcal/mol and $\Delta G_f <$ 2.9 kcal/mol, respectively (see ref. [73], where a similar analysis was used to relate peptide immunogenicity and folding stability). Specifically, only strongly interacting proteins with $K_d^{\exp} < 1.2 \times 10^{-7}$ M can efficiently bind disordered proteins ($\Delta G_f > 0$). As shown in Fig. 3.7A, a more stringent criterion of 97% binding efficiency also leads to a wide range of stability thresholds, where now $K_d^{\exp} < 1 \times 10^{-8}$ M can tolerate disorder. An excess of protein ($c_p > c_s$) or substrate ($c_s > c_p$) can accommodate a slightly larger amount of disorder (Fig. 3.8A), but this does not affect our main conclusion that highly complementary interactions are more tolerant of disorder, whereas the binding efficiency of low-complementarity interactions is rapidly diminished by disorder.

For *catalysis*, there is an additional step of substrate conversion to product P via the FS complex,

$$\begin{array}{ccccccc} & & \mathrm{F} & + & \mathrm{S} & \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} & \mathrm{FS} \overset{k_{\mathrm{cat}}}{\rightarrow} \mathrm{F} + \mathrm{P} \\ \Delta G_f & \uparrow\downarrow & & & & & \\ & & \mathrm{U} & & & & \end{array} \tag{3.6}$$

where k_1, k_{-1}, and k_{cat} are reaction rate constants. Within the Michaelis-Menten limit [9], i.e., assuming the concentration of FS reaches steady state

$$d[\mathrm{FS}]/dt = k_1[\mathrm{F}][\mathrm{S}] - (k_{-1} + k_{\mathrm{cat}})[\mathrm{FS}] = 0 \tag{3.7}$$

one can derive the catalytic rate

$$V_{\mathrm{cat}} \equiv \frac{d[\mathrm{P}]}{dt} = k_{\mathrm{cat}}[\mathrm{FS}] = \frac{k_{\mathrm{cat}} c_p [\mathrm{S}]}{K_m^c \left(1 + e^{\Delta G_f / RT}\right) + [\mathrm{S}]} \tag{3.8}$$

where $K_m^c \equiv (k_{-1} + k_{\mathrm{cat}})/k_1$ is the Michaelis constant. At given K_m^c, V_{cat} reaches a maximum $V_{\mathrm{cat}}^{\max}$ if $\Delta G_f \ll 0$. Figure 3.7B show the ratio $V_{\mathrm{cat}}/V_{\mathrm{cat}}^{\max}$, which defines a measure of the

efficiency of catalysis, as functions of ΔG_f with different values of K_m^c. For typical K_m^c values between 10^{-1} M and 10^{-6} M there is a relatively invariant threshold of the folding free energy, $\Delta G_f = -1.0$ kcal/mol, above which catalysis becomes suboptimal (i.e., $V_{\rm cat}/V_{\rm cat}^{\rm max} \leq 90\%$). This threshold is maintained even for substrate concentrations as high as 10^{-5} M (Fig. 3.8B). Thus, catalytic function is optimized when thermodynamics strongly favor the ordered state. Interestingly, because to have a fast conversion rate the strength of the enzyme–substrate interaction characterized by the Michaelis constant K_m must be much weaker than standard protein–protein K_d, enzymes can also be thought of as a special case of extremely weak binding proteins, i.e., ordered.

In the discussions above, I assumed folding is a two-state equilibrium between the unfolded state and the folded state. However, the conclusions also apply to folding through multiple states. Here I assume the protein first goes from the unstructured state U to an intermediate state I and then to the foled state F

$$\mathrm{U} \overset{\Delta G_f^{(1)}}{\longleftrightarrow} \mathrm{I} \overset{\Delta G_f^{(2)}}{\longleftrightarrow} \mathrm{F} \tag{3.9}$$

We have $[\mathrm{U}]_{\rm eq}/[\mathrm{I}]_{\rm eq} = e^{\Delta G_f^{(1)}/RT}$ and $[\mathrm{I}]_{\rm eq}/[\mathrm{F}]_{\rm eq} = e^{\Delta G_f^{(2)}/RT}$. Thus

$$K_d^{\rm exp} \equiv \frac{\left([\mathrm{U}]_{\rm eq} + [\mathrm{I}]_{\rm eq} + [\mathrm{F}]_{\rm eq}\right)[\mathrm{S}]_{\rm eq}}{[\mathrm{FS}]_{\rm eq}} = K_d^c \left(1 + e^{\Delta G_f^{(2)}/RT} + e^{\Delta G_f^{(1)}/RT} e^{\Delta G_f^{(2)}/RT}\right) \tag{3.10}$$

where $K_d^c \equiv [\mathrm{F}]_{\rm eq}[\mathrm{S}]_{\rm eq}/[\mathrm{FS}]_{\rm eq}$. Compare Eqs. 3.4 and 3.10, disorder is now characterized by two folding free energies instead of one. But Eq. 3.5, which is used in Fig. 3.7A, is unchanged so the conclusions from the two-state folding analysis still hold, and they can be easily generalized to *N*-state folding.

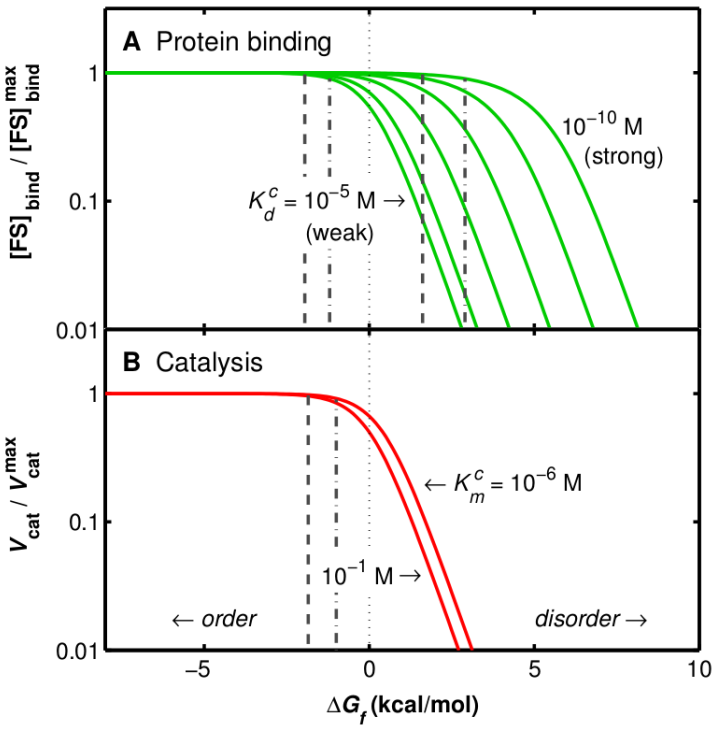

Figure 3.7: Binding and catalytic efficiency.

(A) Ratio of complex concentration $[FS]_{bind}$ as given by Eq. 3.5 to maximum concentration $[FS]^{max}_{bind}$ ($\Delta G_f \ll 0$). $c_p = c_s = 1$ µM. Vertical dash-dotted lines indicate the folding free energy for 90% (dashed lines for 97%) binding efficiency ($[FS]_{bind}/[FS]^{max}_{bind}$ with $K^c_d = 10^{-5}$ and 10^{-10} M, respectively. To maintain high binding efficiency, weak binding requires negative ΔG_f (prefers order), whereas strong binding allows positive ΔG_f (tolerates disorder). (B) Fractional production rate for catalytic activity relative to maximum catalytic rate V^{max}_{cat} ($\Delta G_f \ll 0$) as given by Eq. 3.8 ([S] = 1 µM). The vertical dash-dotted line indicates the folding free energy for 90% (dashed line for 97%) catalytic efficiency (V_{cat}/V^{max}_{cat}) with all relevant K^c_m. To maintain high catalytic efficiency, negative ΔG_f (ordered structure) is required for the whole range of physiological parameters shown here. Note that to allow for fast conversion, enzyme-substrate interactions (characterized by the Michaelis constant K_m) are limited to much weaker interactions than those of binding proteins (K_d).

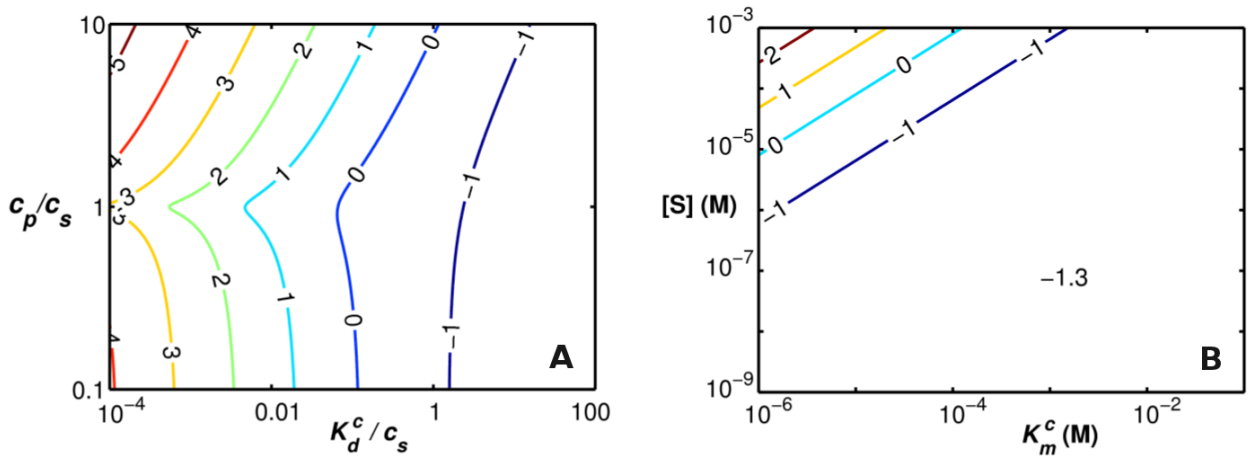

Figure 3.8: Expanded view of stability thresholds.

(A) Contour plot of stability thresholds ΔG_f^{bind} for 90% binding efficiency ($[\text{FS}]_{\text{bind}}/[\text{FS}]_{\text{bind}}^{\text{max}} = 0.9$, where $[\text{FS}]_{\text{bind}}$ is the protein-substrate complex concentration) as a function of the dimensionless quantities c_p/c_s and K_d^c/c_s, where c_p and c_s are the total protein and substrate concentration respectively, and K_d^c is the the "complementary" binding affinity. The region between −1 kcal/mol < ΔG_f^{bind} < 3 kcal/mol covers almost the full phase space of physiological parameters. (B) Contour plot of stability thresholds ΔG_f^{cat} for 90% catalytic efficiency ($V_{\text{cat}}/V_{\text{cat}}^{\text{max}} = 0.9$, where V_{cat} is the catalytic rate) under the full range of Michaelis constant K_m^c and substrate concentration [S]. The stability threshold is around −1 kcal/mol for almost the full range of parameters.

I also assumed the folding and binding of IDPs follow the conformational selection pathway (Fig. 3.6). However, the conclusions also apply to the induced folding pathway. Here I assume the unfolded protein U first associates with the substrate S, then transforms into the folded state F with the assistance from the substrate and becomes bound to it (induced folding)

$$\mathrm{U} + \mathrm{S} \overset{K_d^{(1)}}{\longleftrightarrow} \mathrm{U} \cdot\cdot \mathrm{S} \overset{K_d^{(2)}}{\longleftrightarrow} \mathrm{FS} \tag{3.11}$$

We have

$$\begin{aligned} K_d^{(1)} &= [\mathrm{U}]_{\mathrm{eq}}[\mathrm{S}]_{\mathrm{eq}}/[\mathrm{U} \cdot\cdot \mathrm{S}]_{\mathrm{eq}} \\ K_d^{(2)} &= [\mathrm{U} \cdot\cdot \mathrm{S}]_{\mathrm{eq}}/[\mathrm{FS}]_{\mathrm{eq}} \\ c_p &= [\mathrm{U}] + [\mathrm{U} \cdot\cdot \mathrm{S}] + [\mathrm{FS}] \\ c_s &= [\mathrm{S}] + [\mathrm{U} \cdot\cdot \mathrm{S}] + [\mathrm{FS}] \end{aligned} \tag{3.12}$$

One can derive

$$\left(1 + K_d^{(2)}\right)^2 [\mathrm{FS}]_{\mathrm{eq}}^2 - \left[\left(c_p + c_s\right)\left(1 + K_d^{(2)}\right) + K_d^{(1)} K_d^{(2)}\right][\mathrm{FS}]_{\mathrm{eq}} + c_p c_s = 0 \tag{3.13}$$

Note that the situation where induced folding applies is when FS is much more stable than U ·· S ($K_d^{(2)} \ll 1$). Thus

$$[\mathrm{FS}]_{\mathrm{eq}}^2 - \left(c_p + c_s + K_d^{(1)} K_d^{(2)}\right)[\mathrm{FS}]_{\mathrm{eq}} + c_p c_s = 0 \tag{3.14}$$

Since the total free energy change of the whole folding/binding process is independent of the pathway it takes, we have (compare Eqs. 3.2 and 3.11)

$$RT \ln\left(K_d^{(1)}/c_0\right) + RT \ln\left(K_d^{(2)}/c_0\right) = \Delta G_f + RT \ln(K_d^c/c_0) \tag{3.15}$$

where $c_0 = 1$ M, the left hand side Eq. 3.15 corresponds to induced folding and the right hand side corresponds to conformational selection. The equation can be simplified as

$$K_d^{(1)} K_d^{(2)} = e^{\Delta G_f/RT} K_d^c \tag{3.16}$$

Solve for $[\mathrm{FS}]_{\mathrm{eq}}$ using Eqs. 3.14 and 3.16, one obtains

$$[\mathrm{FS}]_{\mathrm{eq}} = \frac{1}{2}\left[\left(c_p + c_s + e^{\Delta G_f/RT}K_d^c\right) - \sqrt{\left(c_p + c_s + e^{\Delta G_f/RT}K_d^c\right)^2 - 4c_p c_s}\right] \qquad (3.17)$$

Induced folding applies when $[\mathrm{U}]_{\mathrm{eq}} \gg [\mathrm{F}]_{\mathrm{eq}}$ ($\Delta G_f \gg RT$), where the Eq. 3.17 is equivalent to Eq. 3.5, so the conclusions on *Binding* proteins based on conformational selection also applies when induced folding is assumed instead. This is also true for *Catalysis* since it is a special case of *Binding*.

3.3 SPECIFICITY OF PROMISCUOUS INTERACTIONS

Here I show that disorder also provides a mechanism to distinguish between two substrates that differ in binding affinity by a relatively small amount, say 1.5 kcal/mol (Fig. 3.9). For strong binding (K_d^{exp} small), the amount of complex formation with each substrate is almost indistinguishable. A positive ΔG_f , however, can tune K_d^{exp} (Eq. 3.4) to maximize the discrimination between binding of the two substrates while at the same time maintaining a high level of binding to the higher-affinity substrate. Note that the experimental affinity required to bring about this optimal specificity is lower the higher the concentration of protein or substrate. Our finding is reminiscent of Schulz's high-complementarity (or small K_d^c), low-affinity (or large K_d^{exp}) rationalization of the flexibility of nucleotide binding proteins [22], which has also been applied in the context of signal transduction [26] as well as the suggestion of Dunker et al. [74] that disorder uncouples complementarity (K_d^c) and affinity (K_d^{exp}). Here I define "specificity" as simply providing better discrimination among similar physical interactions, a more common usage of the concept [71] that is likely to play a critical role in complex cellular networks.

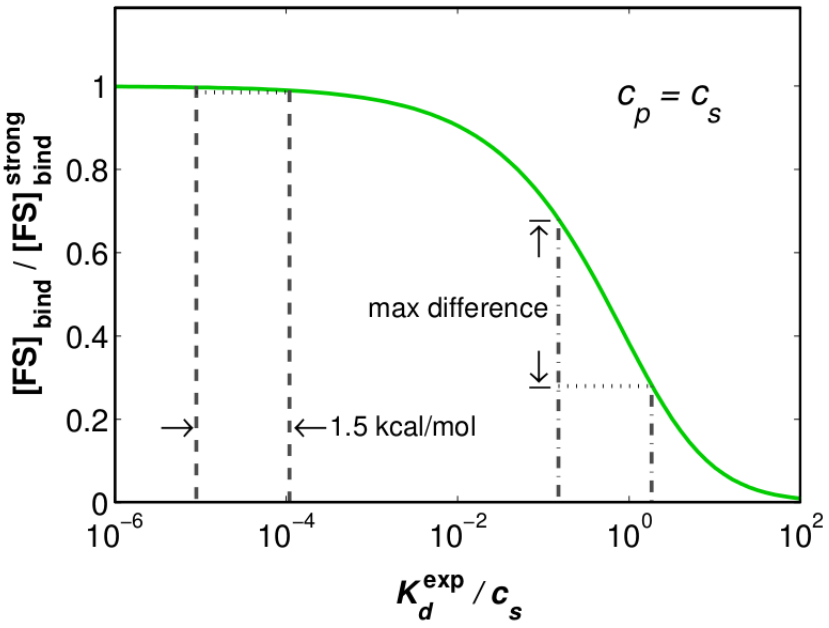

Figure 3.9: Maximum discrimination in binding to similar substrates.

The solid curve shows the equilibrium complex concentration $[FS]_{bind}$ (Eq. 3.5) normalized by the strong binding limit $[FS]^{strong}_{bind}$ ($K_d^{exp} \to 0$). $c_p = c_s$ is used without losing generality. Each pair of vertical lines shows the relative amount of bound complexes formed by two different substrates with a binding free energy difference of 1.5 kcal/mol. For strong binding, the complex concentration saturates, and there is almost no difference in the amount of complex formed by either substrate (dashed lines). On the other hand, decreasing the experimental binding affinity by destabilizing the folded state (F) enhances complex formation by the stronger binding substrate relative to the weaker one (dash-dotted lines).

3.4 DISCUSSION

The survey indicates that the distribution of the amount of disorder depends strongly on protein function, and a first-principles thermodynamic analysis explains the nature of this relationship. For proteins whose main function is to bind other proteins, the amount of disorder that can be tolerated without degrading function is quite broad, depending on the complementarity of the interaction. Catalytic proteins have a strong preference for a stable folded state with $\Delta G_f \lesssim 1$ kcal/mol, consistent with the notion that catalysis has strong conformational requirements, as conjectured by Pauling [18] in the pre-structure age and more recently discussed by other researchers (see, e.g., ref. [75]). Note, however, that although protein stability below the aforementioned threshold (Figs. 3.7B and 3.8B) does not improve catalysis any further [76], this pre-organized state leaves ample room for conformational changes that might be required to bring about efficient catalysis. Finally, I show that disorder can be used to maximize the specificity of promiscuous interactions relevant to transcription and signal transduction.

Instead of rationalizing our findings in terms of adaptability or other processes that are not easily quantifiable, I restrict the discussion to the experimentally derived parameters defined in our models, making our predictions both experimentally and quantitatively more relevant. For instance, Fig. 3.9 shows that for μM concentrations, highly complementary complexes, say, $K_d^c \sim$ nM, will yield maximum discrimination if folding instability lowers $K_d^{\exp}$ to μM. This extra discrimination is likely to play a role in the differential regulation of promiscuous binding domains such as SH2/3s, whose typical affinities agree with the predictions of the model [77]. More interestingly, the theory also elucidates the dependence on concentration of the experimental affinity that optimizes specificity (Fig. 3.9).

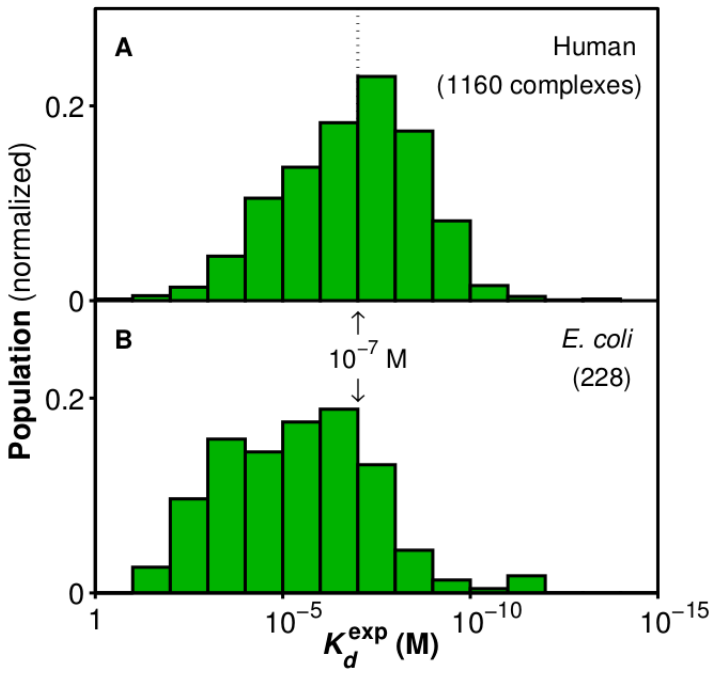

Figure 3.10: Distributions of experimentally measured protein-ligand binding affinities.

Data are taken from the PDBbind database (version 2007). The overall distributions are consistent with our hypothesis that the lack of disorder in prokaryotes could be due to their relatively weaker binding affinities ($\gtrsim 10^{-7}$ M).

The theory predicts that lower-affinity interactions are expected to involve proteins with less disorder, which may help explain why disorder is less prevalent in prokaryotes (type III) than eukaryotes (types I and II). Indeed, the strikingly similar distributions for *E. coli* shown in Fig. 3.3 suggest that disorder does not play a role in function (similar data are observed for other prokaryotes). Without disorder, protein binding efficiency would imply $K_d^{\exp} \gtrsim 10^{-7}$ M. A survey of the protein-ligand interactions in the PDBbind database [78] (Fig. 3.10) confirms not only that bacterial proteins may indeed bind small ligand molecules more weakly than humans proteins but also that there is a sharp drop in the number of *E. coli* ligands (20% compared with 50% for human) with $K_d^{\exp}$ smaller than the predicted threshold of 10^{-7} M. From the point of

view of evolution, the drop of $K_d^{\exp}$ is also consistent with the intuition that short-lived microorganisms have less need to form long-lived complexes.

It is important to stress that protein-functional assignments are still incomplete [67]. Indeed, for the genomes I analyzed, only a subset of all proteins has at least one assigned function, e.g., ~75%, 88% and 32% of human, yeast, and *E. coli*, respectively. As already mentioned, our analysis encompasses motifs participating in the molecular interactions. Hence, for multi-site/domain proteins a specific function should not necessarily require folding of the entire protein. Figure 3.11 further expands on the amount of intrinsic disorder in multifunctional proteins as well as on the correlation of disorder and protein length. For the most part, I find that proteins with both binding and transcription functions have a disorder distribution similar to transcription, whereas the distribution for proteins with binding and catalytic functions is more similar to catalytic. For these subsets, I failed to observe significant correlations between disorder and protein length. For *E. coli*, most proteins are ordered. However, the few highly disordered proteins involved in transcription are all relatively small, resulting in a weak negative correlation. The small sets of proteins with both catalytic and transcription functions as well as all three functions (including binding) show a positive correlation with length while seemingly encompassing a combination of the disorder distributions of each individual functional category. Further analysis of disorder as a local property of the functioning site is likely to reveal insights into how evolution has coupled structure and functions to cope with the increasing complexity of higher organisms.

Ultimately, the theory might provide more subtle quantitative predictions for the interplay between disorder and function for specific proteins. Although current experimental technologies cannot readily analyze weakly stable proteins, let alone positive folding free energies,

computational techniques might help to fill this gap. Although there are other aspects not considered here, such as the role of disorder in aggregation and degradation, our findings show how disorder has opened a new dimension in the regulation of molecular interactions for eukaryotes and, most certainly, humans. Collectively, our findings suggest that protein folding should be viewed as a continuum in which folding stability is just one more parameter that evolution uses to optimize function.

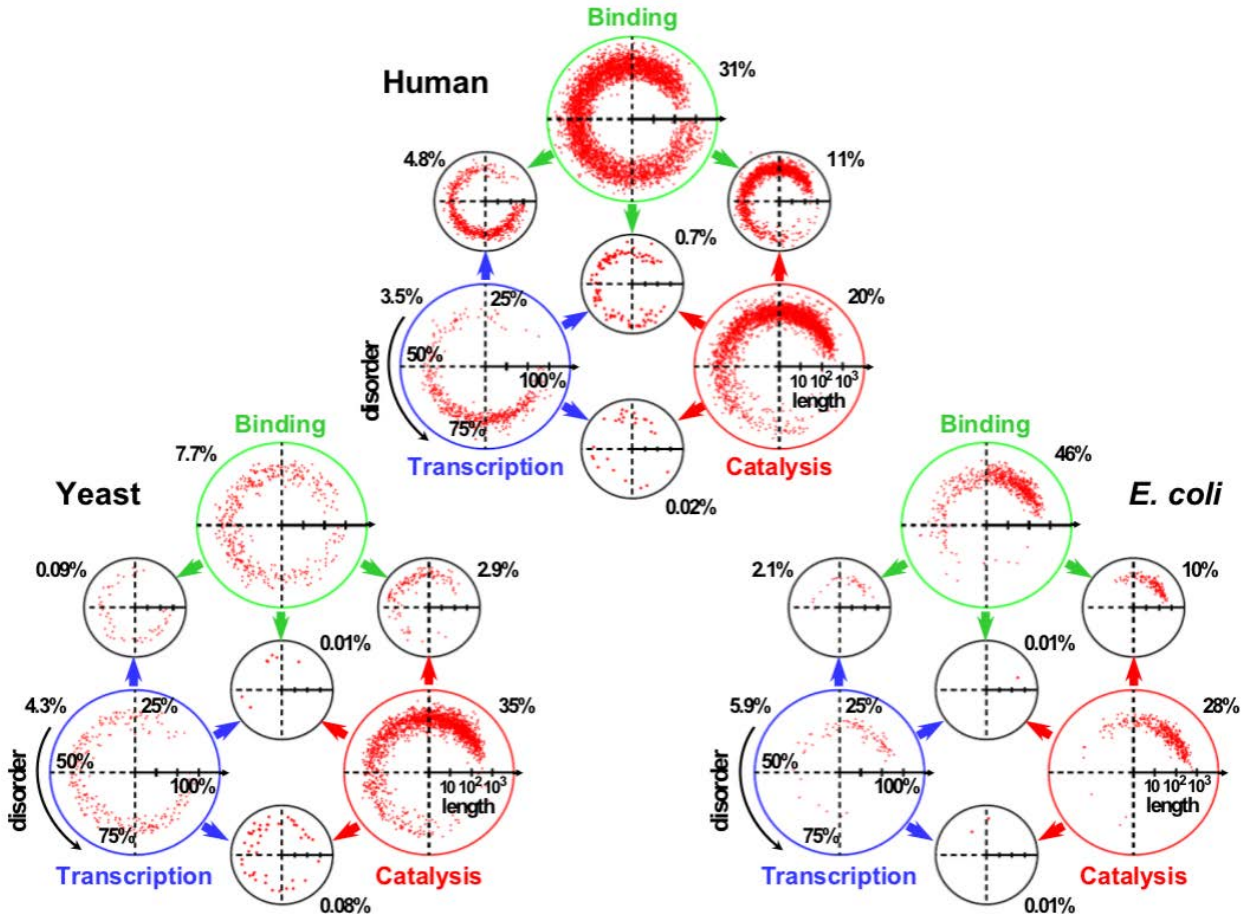

Figure 3.11: Intrinsic disorder as a function of protein length.

Plotted are proteins with (non-overlapping) binding, transcription, and catalytic function (large circles), and for proteins with more than one function, as indicated by the colored arrows from each individual functional category (smaller circles). For each polar coordinate plot, the radial and angular (counterclockwise) coordinates correspond to protein length in a log-scale and the percentage of residues that are classified as disordered for the protein (as in Fig. 3.3), respectively. For clarity, percent disorder and protein length are labeled only in transcription and catalysis plots, respectively. Indicated outside each circle is the percentage of proteins in each functional category relative to the total number of proteins for which the function has been annotated for each organism (i.e., 15,260, 5,900, and 1,362 for human, yeast and *E. coli*, respectively). The figure shows that disorder does not correlate with protein length for well-sampled functional categories. The analysis of disorder in multifunctional proteins also reveals interesting patterns. Specifically, binding does not seem to impact the level of disorder of either transcription or catalytic proteins, whereas disorder in proteins with both catalytic and transcription functionalities appear to follow either one of the patterns found for the individual functions.

4.0 BACTERIAL SPORE GERMINATION

4.1 INTRODUCTION

Some gram positive bacteria (e.g., *Bacillus* and *Clostridium*) stay in the division/growth cycle when nutrients are abundant and conditions are favorable; otherwise, they protect themselves by transforming into spores (Fig. 4.1) [79]. Spores are radically different from normal bacterial cells [80, 81]. They have low water content, low protein mobility, and near-undetectable metabolic activity. They are also highly resistant to harsh environmental conditions such as heat, radiation, and toxic chemicals. These properties make spores long-lived and hard to kill. However, they constantly monitor their surroundings and can initiate germination into normal cells within a few minutes when nutrients, called germinants, reappear (Fig. 4.1).

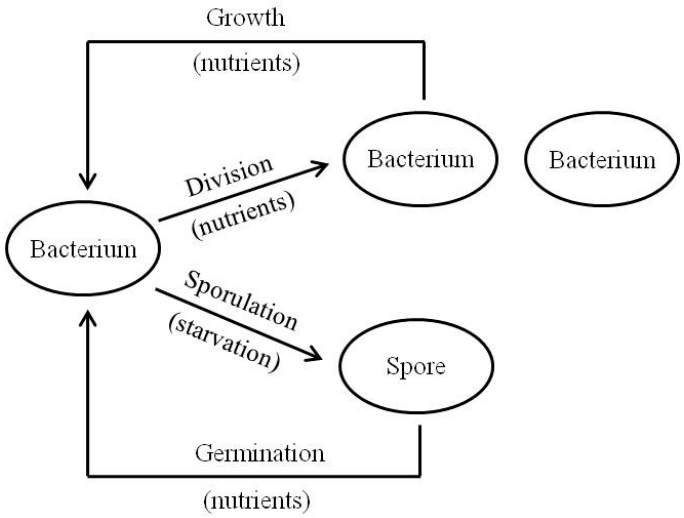

Figure 4.1: Life cycle of spore forming bacteria.

4.2 MEASUREMENTS OF GERMINATION

4.2.1 Single spore measurements

Germination of single spores has been characterized with electron microscopy, phase contrast microscopy, and Raman scattering spectroscopy. Electron microscopy yields the greatest details on the structure of spores and their morphological changes during germination [82, 83]. As shown in Fig. 4.2A, a spore is typically structured as (from interior to exterior): a core containing DNA and other essential molecules surrounded by thin layers of an inner membrane and a cell wall; a thick shell called the cortex mainly consisting of peptidoglycan; a thin layer of outer membrane; a thin layer called the coat mainly consisting of proteins, which protects the spore from reactive chemicals and predators; and finally, an additional layer called the sporangium, which is present only in some species and is comprised of the remains of the wall from the original bacterium. During germination, the cortex is degraded, the part of the spore surrounded by inner membrane and cell wall transforms into a vegetative cell, and the coat is torn apart and abandoned (Fig. 4.2B). Electron microscopy has the limitation that spores have to be killed and fixed with chemicals before the observation. Therefore, it cannot be used to measure the kinetics of germination. In contrast, phase contrast microscopy and Raman scattering spectroscopy can be used to monitor germination in real time without disrupting the process.

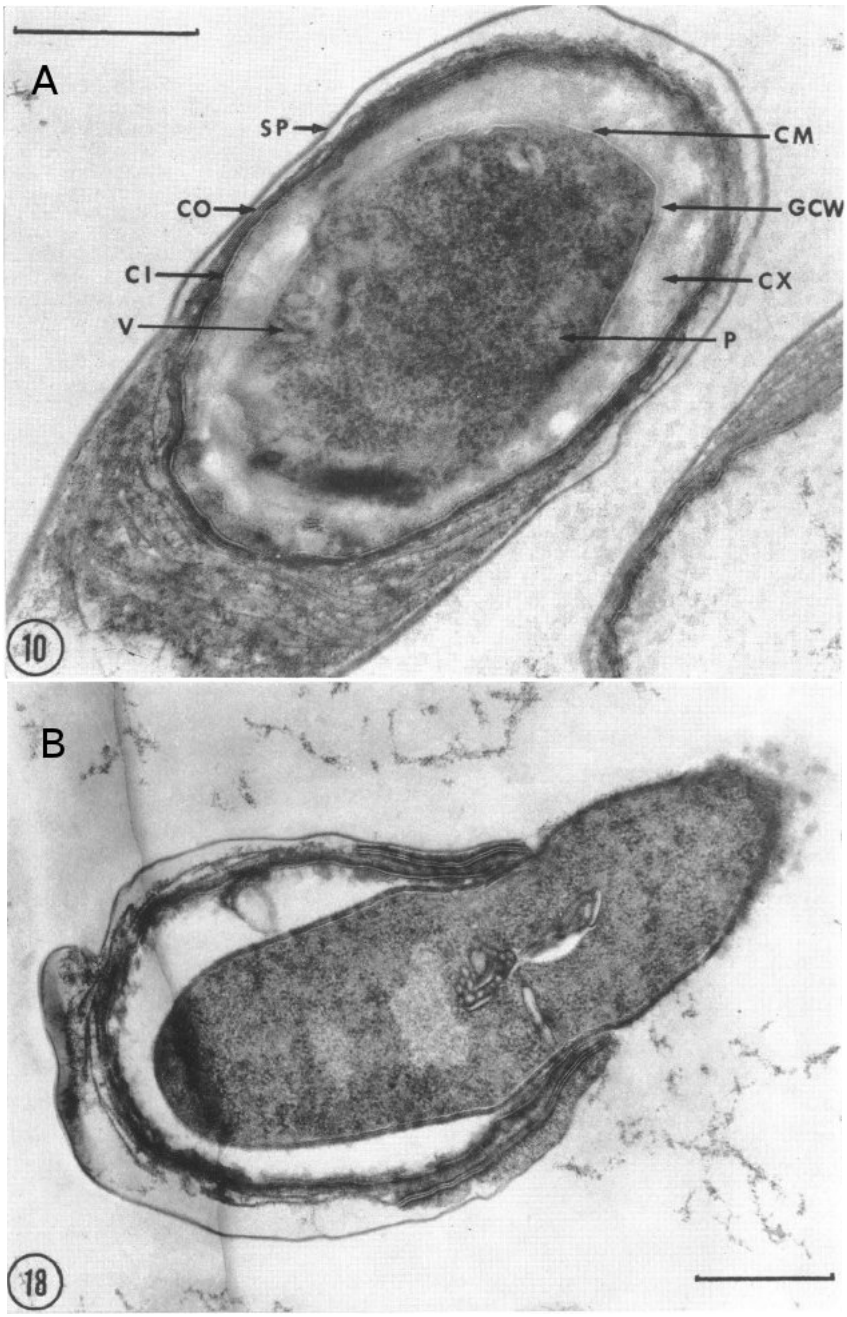

Figure 4.2: Electron micrographs of longitudinal sections of germinating *Clostridium pectinovorum* spores. (A) Early stage of germination. The labels are: P - protoplast, CM - core membrane, GCW - germ cell wall, CX - cortex, CI - inner coat, CO - outer coat, V - vesicular structures, and SP - sporangium. (B) Last stage of germination, where vegetative cell emerges from fractured spore coat. Markers represent 0.5 µm. Other spore species have similar structures. (Source: Adapted from Ref. [83].)

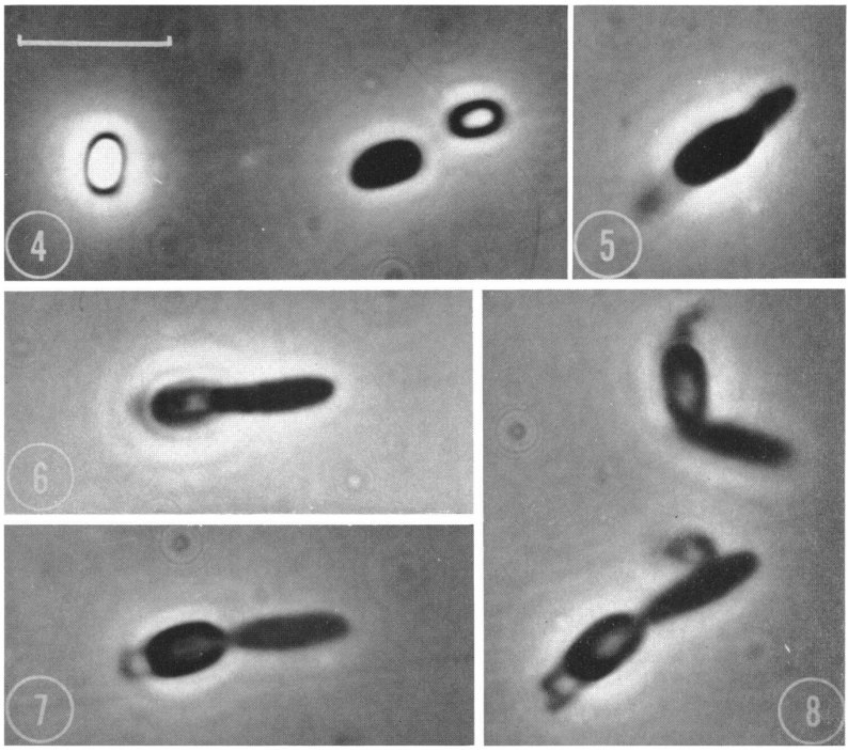

Figure 4.3: Phase contrast micrographs of germinating *Clostridium pectinovorum* spores. Shown are the loss of refractility and progressive emergence of vegetative cell. Marker represents 5 µm and applies to all the images. The sequence of events is labeled with numbers. Other spore species undergo similar changes during germination. (Source: Adapted from Ref. [83].)

Figure 4.3 shows the process of germination observed with phase contrast microscopy [83]. Spores are initially phase bright. After the introduction of germinants, they first become phase dark, indicating that they have initiated germination. Then vegetative cells gradually emerge from spores' coats. Early experiments on hundreds of spores using this technique [84] showed that, while the transition from phase bright to phase dark is rapid (~10 seconds), the time it takes for a spore to initiate the transition varies widely from spore to spore (from a few seconds to more than 10 minutes). This heterogeneity in germination is not due to genetic differences or variations in germination condition, but appears to be due to stochasticity in gene expression and

variations in culture conditions during sporulation [7]. It is advantageous for the survival of spore populations, as the fast germinating spores make sure opportunities to establish new colonies are not missed when conditions improve and the slow germinating spores guarantee the whole population is not jeopardized when the improvement is only temporary. However, it also greatly complicates spore eradication in the food and health care industries, as small percentages of spores germinate extremely slowly (called superdormant spores) and are not easily killed before germination [85-87].

Recently, Raman scattering spectroscopy has also been used to characterize spore germination [88-91]. A major event during the early stage of germination is the release of spores' large depot (~10% of spore dry weight) of Ca^{2+} and dipicolinic acid (DPA) [80]. The two form chelated complexes inside the spore core, thus I will call them CaDPA. CaDPA can be detected by Raman scattering spectroscopy, as it gives rise to strong Raman scattering of light at certain wavelengths, and the amount of CaDPA is proportional to the intensity of the scattering [88, 92]. In a series of recent experiments, individual spores were confined with optical traps, and the amount of CaDPA inside each spore was measured during germination [89, 90]. It was found that there is usually a delay between the addition of germinant and the beginning of CaDPA release, and the length of this delay varied from spore to spore, but the durations of CaDPA release were roughly the same for all the spores and were short compared to the delay before the release (Fig. 4.4). These experiments show that the heterogeneity in germination originates from processes prior to the release of CaDPA.

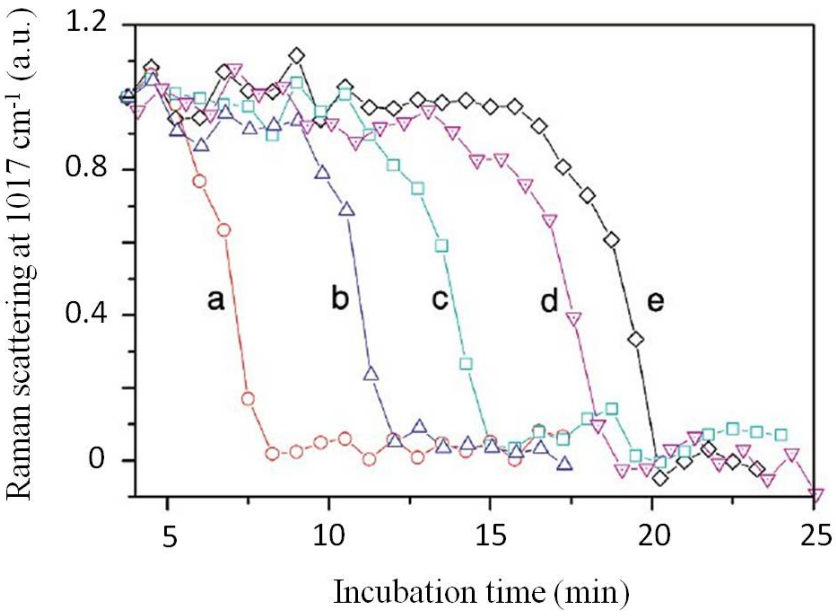

Figure 4.4: Raman scattering from germinating spores.

***Bacillus subtilis* spores were germinated with L-alanine. Relative intensities of the 1017 cm^{-1} CaDPA Raman band from five individual spores are plotted, which show that the incubation time – time since the addition of germinants – required for the initiation of CaDPA release ranges between 5-15 minutes, and the duration of the release is around 3 minutes for all the spores. (Source: Adapted from Ref. [89].)**

4.2.2 Population measurements

While single spore measurements yielded many details on what happens during germination, they are time consuming to perform and the extracted properties are noisy. Population measurements have also been developed, which are advantageous in exploring germination under many different experimental conditions.

An early method that has been commonly used is to measure the optical density (OD) of solutions containing spores, which is a measure of how much light of certain wavelength (typically 600 nm) is absorbed by the sample

$$\mathrm{OD} = \log_{10}(I_0/I) \tag{4.1}$$

where I_0 and I are intensities of the incident and transmitted light respectively. OD decreases as spores germinate, and this property has been used to study the effects of various factors on germination [93-96]. However, this method does not give quantitative measures on germination as it is not clear how to convert changes in optical density to percentage of spore germination, which is essential for this dissertation.

Recently, a new technique has been developed to yield quantitative measures of germination [97, 98]. It takes advantage of the fact that spores release CaDPA upon germination. DPA^{2-} can form fluorescent complexes with Tb^{3+}, so Tb^{3+} is added to the solution containing spores at the beginning of each experiment. Therefore, by measuring the fluorescence intensity, one obtains the amount of CaDPA released by the spores. Since each spore from the same population contains about the same amount of CaDPA [92], the method provides a quantitative measure on the amount of spores that have released CaDPA. In addition, it can also be used to study another major event in the early stage of germination – commitment. Commitment happens before CaDPA release. While germinants are required for spores to initiate germination, they do not have to be present for the whole period until CaDPA release. Experiments showed that spores continue to germinate after the interactions between germinants and their corresponding GRs were blocked [93, 96, 99], demonstrating that commitment is the moment where no germinant is needed anymore and spores irreversibly proceed to later stages of germination. Similar to CaDPA release, the time it takes for a spore to commit to germination is also heterogeneous [99].

4.3 STAGES OF GERMINATION

To help the readers understand the problems I am addressing in Chapter 5, here I divide germination of a single spore into four stages based on two well characterized events – commitment to germination and CaDPA release – and name them as: commitment, pre-CaDPA release, CaDPA release, and post-CaDPA release (Fig. 4.5).

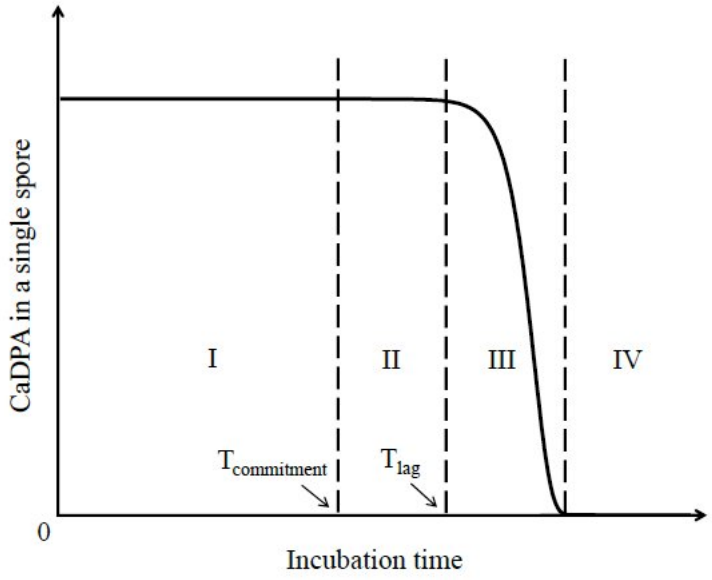

Figure 4.5: Four stages of spore germination with nutrient germinants.

(I) commitment to germination, (II) pre-CaDPA release, (III) CaDPA release, and (IV) post-CaDPA release. Note that durations of the various stages are not drawn to scale, and that there can be a small amount (~10%) of slow CaDPA release prior to T_{lag} starting at the beginning of stage I with spores of some species [88, 89, 91, 100].

Stage I, which is a major source of germination heterogeneity [99], starts from the addition of germinant and ends the moment that the spore commits to germination. Germinants are typically amino acids, sugars or purine nucleosides. They trigger germination by binding to a group of

germinant receptors (GR) on the inner membrane of spores [80]. Germinant-GR interaction plays a major role in Stage I and a large set of experiments can be and have been performed, in particular with *B. subtilis* spores since: i) the GRs are already known; ii) many germinants and inhibitors for each GR are known; iii) levels of GRs can be controlled and likely even measured directly in the near future; and iv) there are spores with GR variants that exhibit modified germinant-GR interaction. The next question is what other molecules and processes are involved in Stage I, and this will be key to understanding how nutrient-GR interaction triggers germination, why $T_{commitment}$ varies between spores and is decreased by heat activation and increasing the GR level per spore.

Stage II starts from the moment of commitment and ends at the initiation of CaDPA release. Little is known about Stage II, in large part because there is as yet no method for direct measurement of the length and heterogeneity of its duration in individual spores. The molecular changes involved in Stage II are unknown, except that the release of monovalent ions (H^+, Na^+, K^+) should begin during one of the first two stages, since this event is prior to the release of CaDPA [101]. While the initial germinant is no longer required at this stage, it is unclear whether germinant-GR interaction still plays any role.

Stage III starts from the initiation of CaDPA release and ends at the completion of CaDPA release. This stage has been well characterized by a series of recent experiments [88-91]. The duration of Stage III is relatively constant and unaffected by heat activation, germinant concentration and levels of GRs per spore. In addition, experiments measuring the uptake of a nucleic acid dye by single germinating *Bacillus* spores showed that the uptake starts at the beginning of Stage III [102], implying the initiation of drastic changes inside the spore core at this time, although what these changes are is not clear. One remaining question is what

determines the duration of CaDPA release, which lasts three minutes in wild type *B. subtilis* spores [89] but only takes 30 seconds in wild type *B. cereus* spores [90]. The protein CwlJ is certainly one of the players in this stage, since its deletion results in significantly slower CaDPA release [89].

The final stage in germination, Stage IV, starts from the completion of CaDPA release and covers the remainder of the germination process, including the progression into spore outgrowth. Significant heterogeneity has also been observed in this stage but little is known about the contributing factors or mechanisms [7]. Since I am only concerned with the mechanism that initiates spore germination, I will not discuss Stage IV further.

4.4 OUTSTANDING QUESTIONS

The duration of the early stage of germination varies a lot between spores, and, depending on germinant concentration, certain fraction of spores may not germinate even at very long times [80]. This heterogeneity of germination has long been a concern in food industry and health care, and has been studied mainly via either direct or indirect measurement of CaDPA release. However, it was realized only recently that the major source of the heterogeneity is actually in the commitment step [99]. It is now becoming clear that the commitment step is the key to understand the mechanism of germination initiation and the reason for germination heterogeneity. Yet there is still no definite answer on what mechanism is responsible for the triggering of germination and what determines how long a spore takes to commit to germination.

In Chapter 5, I will propose a quantitative model based on the assumption that the heterogeneity in germination is due to the variability in levels of activated GRs per spore. The

model has three key components: the distribution of GR numbers in a spore population, the concentration dependence of nutrient binding and activation of the particular GR, and the threshold number of bound GRs for a spore to germinate. The GR distribution can be directly determined from experiments, and the other two components were determined by fitting data for percentage of germination as a function of nutrient concentration. The model has been used to predict germination of spores with mixtures of nutrients that trigger different types of GRs and produced results that were consistent with experiments, which suggests that signals from different types of GRs are summed by a common signal integrator.

5.0 A QUANTITATIVE MODEL OF GERMINATION

5.1 INTRODUCTION

In this chapter, I propose a quantitative model on Stage I of germination – the commitment step. I will use the model to address the following two questions: (1) what determines how long it takes for a spore to commit to germination, i.e., what cause the heterogeneity of germination; (2) how are signals alerting the presence of germinants processed inside spores.

The model is inspired by an earlier model by Woese et al. [103]. To understand the mechanism of germination and the reason for its heterogeneity, Woese *et al* postulated that germination occurs (spores become phase dark) when the level of some unknown substance P in a spore reaches a threshold P_c, and production of P is catalyzed by some unknown "germination enzyme" E in the spore (Fig. 5.1). The rate of accumulation of P is assumed to be

$$\frac{dP}{dt} = Kn - k_2 P \tag{5.1}$$

where K and k_2 are constants determining the rate of production and degradation of P respectively, n is the number of E in a spore that are activated by germinants. Before the addition of germinants $P = 0$; after the addition of germinants

$$P(t) = \frac{Kn}{k_2}(1 - e^{-k_2 t}). \tag{5.2}$$

The incubation time it takes for a spore to germinate (when $P(t) > P_c$) is

$$t = \begin{cases} -\frac{1}{k_2}\ln\left(1 - \frac{P_c k_2}{Kn}\right), & n > P_c k_2/K \\ \infty, & n \leq P_c k_2/K \end{cases} \tag{5.3}$$

The model thus accounts for the heterogeneity of germination as due to variations in the number of germination enzymes n from spore to spore, and can also account for superdormant spores [85, 87] – spores that stay dormant even after long incubation times – as due to an insufficient number of germination enzymes.

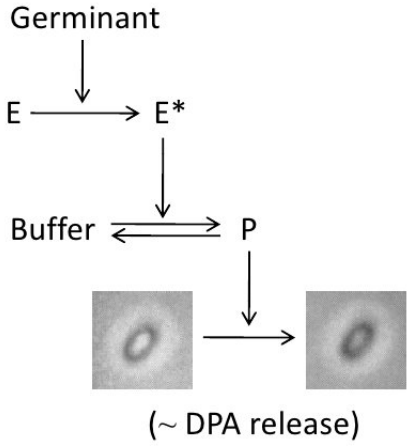

Figure 5.1: Schematic diagram of the model proposed by Woese et al.

The molecules corresponding to the "germination enzyme" postulated by Woese et al. do exist; they are the germinant receptors (GRs) on the inner member of spores. Evidence that GRs are responsible for nutrient-induced germination[1] is (1) spores with GRs knocked out do not respond to nutrients [80, 104], (2) overexpressing a GR leads to faster germination and a higher percentage of germinating spores in the presence of the cognate germinant [105]. However, the biochemical process corresponding to the accumulation of P is still unknown. In addition, the

[1] Germination can also be induced by high pressure and non-nutrients such as exogenous CaDPA and dodecylamine, which are out of the scope of this dissertation.

model has never been used to make quantitative predictions that can be verified against experiments. In the following, I will develop a more general approach to modeling spore germination that encompasses key elements of the Woese model but is more closely linked to exeperimental data both in its parameterization and its predictions.

5.2 GERMINANT RECEPTORS

Nutrients trigger spore germination through interactions with GRs. GRs are proteins located in the inner membranes of spores [80, 106]. In *B. subtilis*, the spore species studied in this dissertation, there are three types of GRs – GerA, GerB, and GerK – which respond to different nutrients (Fig. 5.2). GerA can be triggered by L-alanine or L-valine. Neither GerB nor GerK alone can trigger germination. However, GerB together with GerK can be triggered by AGFK (a mixture of L-asparagine, glucose, fructose, and K^+). In addition, compared with triggering GerA alone, faster germination is observed when GerB or GerK is also triggered with their corresponding germinants, indicating that GerB and GerK can also facilitate germination through GerA (Fig. 5.2). At last, three point mutations were isolated on the GerB GR (the mutated GR is called GerB*) that enabled it to trigger germination without GerK by binding to L-asparagine (or a few other nutrients) alone [107, 108].

It is unclear how multiple GRs act together to trigger germination [108]. In this chapter, I focus on the simplest system – spores of *B. subtilis* FB10 strain, in which GerB is substituted with GerB*, either GerA or GerB* alone can trigger germination with a single nutirient and without co–receptors, and the role of GerK can be ignored. This will lay the groundwork for understanding wild type spores where more complex interactions between the GRs are involved.

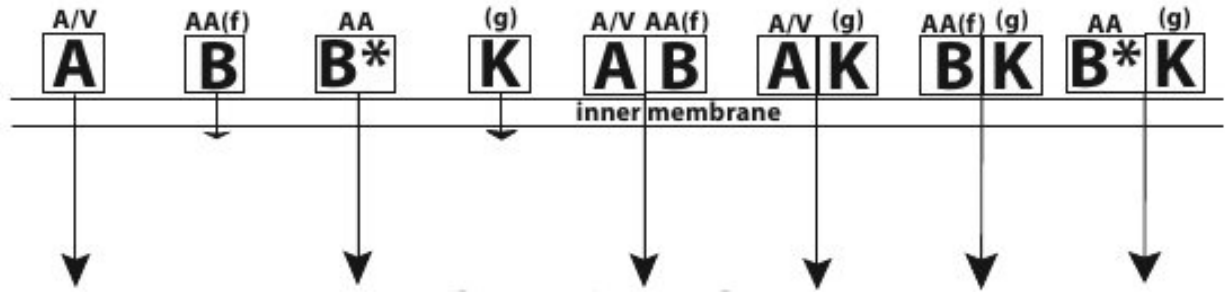

Figure 5.2: The GRs and their cognate germinants in *B. subtilis* spores.

Abbreviations used for receptors are in boxes and are as follows: A, GerA; B, GerB; B*, GerB*; K, GerK. Abbreviations used for germinants or cogerminants are as follows: A, L-alanine, AA, L-alanine, L-asparagine, L-serine, or L-threonine; f, fructose; g,glucose; V, L-valine. "/" indicates "or", cogerminants are shown in parentheses. Long arrows mean germination is induced through the GR(s) with the corresponding germinants, and short arrows means no significant germination is induced. (Source: Adapted from Ref. [108].)

5.3 THEORY

Here I introduce the theoretical basis of two critical building blocks of the spore germination model: Hill equation for germinat-GR interaction, and gamma distribution for GR level distribution.

5.3.1 The Hill equation

I will use the Hill equation to model germinant-GR binding. The equation that bears his name was first used by A. V. Hill [109] to explain experimental data on the absorption of oxygen by the protein hemoglobin in red blood cells. Hill postulated that n hemoglobin molecules aggregate into a complex that then binds oxygen according to the equation

$$\mathrm{Hb}_n + n\mathrm{O}_2 \overset{K_n}{\leftrightarrow} \mathrm{Hb}_n(\mathrm{O}_2)_n \tag{5.4}$$

where, according to the Law of Mass Action, the parameter $K_n \equiv [\mathrm{Hb}_n(\mathrm{O}_2)_n]/[\mathrm{Hb}_n][\mathrm{O}_2]^n$ is a constant (called the association constant) and square brackets denote concentrations. The percentage saturation of hemoglobin with O_2 is then given by

$$y = \frac{K_n[\mathrm{O}_2]^n}{1 + K_n[\mathrm{O}_2]^n} \tag{5.5}$$

which is nowadays called the Hill equation, and n is called the Hill coefficient. Hill noted that this equation fit all known dissociation curves of oxyhemoglobin with a very high degree of accuracy except that n did not turn out to be integer. Hill should have stopped here, but to resolve this problem, he further postulated that hemoglobin formed aggregates of different sizes: HbO_2, $\mathrm{Hb}_2(\mathrm{O}_2)_2$, $\mathrm{Hb}_3(\mathrm{O}_2)_3$, etc., so the exact expression for y should be

$$y = \sum_{n=1,2,3,\ldots} \frac{a_n K_n[\mathrm{O}_2]^n}{1 + K_n[\mathrm{O}_2]^n} \tag{5.6}$$

where a_n is the relative abundance of $\mathrm{Hb}_n(\mathrm{O}_2)_n$. He then claimed Eq. 5.5 to be an approximation of Eq. 5.6. Hill's theory was several decades later found to be wrong, as crystal structures of hemoglobin showed that they are tetramers with each subunit binding one oxygen molecule, and more realistic models were proposed [110, 111]. However, Hill's equation as a simple empirical formula survived, as it captures the sigmoid shape binding curves typically have.

Nowadays, Hill equation is still widely used to analyze protein-ligand interactions. In this thesis, I use a different form of the equation

$$y = x^n/(K_d^n + x^n) \tag{5.7}$$

where K_d has the dimension of concentration and is called apparent dissociation constant. Typical uses of the Hill equation are to provide rough measures of the binding affinity between

ligand and protein through K_d (the smaller it is, the stronger is the interaction) and the cooperativity between different subunits of the same protein (or protein complex) through n (the larger n is, the stronger the cooperative interaction). However, the physical interpretation of the Hill coefficient depends on details of the binding model and no meaning should be inferred from it unless parameters of the correct mechanism can be identified [112]. In the following, I briefly demonstrate that the Hill coefficient has different meanings in different situations.

First, consider the following binding scheme [112]

$$\mathrm{R} + 2\mathrm{L} \overset{K_{d1}}{\longleftrightarrow} \mathrm{RL} + \mathrm{L} \overset{K_{d2}}{\longleftrightarrow} \mathrm{RL_2} \tag{5.8}$$

At equilibrium, we have $K_{d1} = [\mathrm{R}][\mathrm{L}]/[\mathrm{RL}]$, $K_{d2} = [\mathrm{RL}][\mathrm{L}]/[\mathrm{RL_2}]$, and $[\mathrm{R}]_{\mathrm{tot}} = [\mathrm{R}] + [\mathrm{RL}] + [\mathrm{RL_2}]$, and we can derive

$$\frac{[\mathrm{RL_2}]}{[\mathrm{R}]_{\mathrm{tot}}} = \frac{[\mathrm{L}]^2}{K_{d1}K_{d2} + K_{d2}[\mathrm{L}] + [\mathrm{L}]^2} \tag{5.9}$$

When $K_{d2} \ll K_{d1}$, i.e., when there is marked positive cooperativity between the binding of two ligands on the same receptor (the binding of the first ligand significantly increases R's affinity for the second ligand), fitting of the binding curve with the Hill equation leads to a Hill coefficient close to 2.

Then consider the following modified binding scheme [112]

$$\mathrm{R} + 2\mathrm{L} \overset{K_{d1}}{\longleftrightarrow} \mathrm{RL} + \mathrm{L} \overset{K_{d2}}{\longleftrightarrow} \mathrm{RL_2} \overset{K_{d3}}{\longleftrightarrow} \mathrm{R^*L_2} \tag{5.10}$$

where there is an additional step that converts R into an active state R^* once it is fully bound. At equilibrium, we have $K_{d3} = [\mathrm{RL_2}]/[\mathrm{R^*L_2}]$, and $[\mathrm{R}]_{\mathrm{tot}} = [\mathrm{R}] + [\mathrm{RL}] + [\mathrm{RL_2}] + [\mathrm{R^*L_2}]$, and we can derive

$$\frac{[\mathrm{R^*L_2}]}{[\mathrm{R}]_{\mathrm{tot}}} = \frac{[\mathrm{L}]^2}{K_{d1}K_{d2}K_{d3} + K_{d2}K_{d3}[\mathrm{L}] + (1 + K_{d3})[\mathrm{L}]^2} \tag{5.11}$$

When $K_{d1} = K_{d2} \gg K_{d3}$, i.e., when there is no cooperativity between the binding of two ligands on the same receptor (the binding of the first ligand has no effect on R's affinity for the second ligand) but there is strong preference for the active state once R is fully bound, fitting of the binding curve with the Hill equation also leads to a Hill coefficient close to 2.

It is clear that although we could obtain Hill coefficients close to 2 for both schemes, they arise for different reasons.

5.3.2 Gamma distribution

I will use gamma distribution to model the levels of GRs on individual spores. The expression levels of many proteins in cells follow the gamma distribution [113], which can be derived with the following model (derivations adapted from Ref. [114]):

$$\begin{array}{ccccc} \text{DNA} & \xrightarrow{k_1} & \text{mRNA} & \xrightarrow{k_2} & \text{protein} \\ & & \gamma_1 \downarrow & & \gamma_2 \downarrow \\ & & \emptyset & & \emptyset \end{array} \tag{5.12}$$

which assumes that the gene encoding the protein is transcribed into mRNA with a first order rate constant of k_1, the mRNA is then translated into protein with a first order rate constant of k_2, and finally mRNA and protein are degraded with first order rate constants of γ_1 and γ_2 respectively. Let us denote protein concentration with $x(t)$, which varies with time t, and denote the probability density for a cell to have the protein concentration at x with $p(x)$. Note that x is treated as a continuous variable. An analysis where the protein level is treated as a discrete variable can be found in Ref. [115]. The two approaches leads to analytically different but numerically similar distributions. The evolution of $p(x)$ with t is described by the following master equation:

$$\partial p(x)/\partial t = \partial[\gamma_2 x p(x)]/\partial x + k_1 \int_0^x dx' w(x,x') p(x') \tag{5.13}$$

The first term on the right hand side of Eq. 5.13 corresponds to the decrease in x caused by protein degradation and cell growth/proliferation, where $\gamma_2 = \ln 2/T + \ln 2/T_1$, with T being the half-life of the protein and T_1 being the length of a cell cycle. It can be visualized as a continuous current of probability density toward $x = 0$ with a magnitude of $\gamma_2 x p(x)$, so the rate of change in $p(x)$ from this part is given by the divergence of the current $\partial[\gamma_2 x p(x)]/\partial x$.

The second term on the right hand side of Eq. 5.13 corresponds to the change in x caused by the production of protein. First, according to the model, mRNA is been produced at rate k_1, and its lifetime follows exponential distribution (since its degradation is described by a first order rate constant γ_1) with a mean value $1/\gamma_1$. During the life of a mRNA, proteins are been produced at the constant rate k_2, thus the number of proteins translated from each mRNA also follows exponential distribution, with a mean value of $b = k_2/\gamma_1$. Second, for many proteins one can assume the lifetime of mRNA is short compared to the lifetime of the protein, as it was observed in experiments that protein production happens in bursts, with each event resulting in an exponentially distributed number of proteins [116]. Finally, one can assume the burst size $x - x'$ is independent of the original protein concentration x, thus the transport of the probability density p from x' to x is described by

$$w(x,x') = b^{-1} e^{-(x-x')/b} - \delta(x - x') \tag{5.14}$$

where the δ function accounts for the decrease of $p(x)$ due to bursts.

At steady state $\partial p/\partial t = 0$, the master equation can be solved by using the Laplace transformations, and we obtain the gamma distribution:

$$p(x) = x^{a-1} e^{-x/b}/b^a \Gamma(a) \tag{5.15}$$

where $a = k_1/\gamma_2$ is the mean number of mRNAs produced (also the mean number of protein bursts) during the average lifetime of a protein. The parameters a and b determine the shape and scale of the distribution respectively, and the mean is $\langle x \rangle = ab$.

5.4 SINGLE GERMINANT MODEL

First consider germination with single germinant so only one type of GR is involved. Figure 5.3 shows the key elements of the model for germination kinetics (Fig. 5.3D) based on three key elements: the distribution of GR numbers in a spore population (Fig. 5.3A), the concentration dependence of nutrient germinant binding and activation of the particular GR (Fig. 5.3B), and the dependence of commitment time for a spore on the number of bound GRs (Fig. 5.3C).

Of the three basic components of the model only the GR distribution (Fig. 5.3A) can be directly determined from experiments. The other components, the GR activation and the commitment time curves will be determined by fitting data for germination kinetics as a function of germinant concentration. The GRs are present at low levels, with 24-40 molecules per spore for the GerB receptor [117] and other GRs may also be expressed at similar levels. In addition, recent experiments with spores expressing fluorescently-labeled GRs have provided detailed quantitative information about receptor distributions in spores [118]. Here, instead of the Poisson distribution assumed by Woese et al., I assume the GR number on a single spore follows a gamma distribution (Section 5.3.2), as was recently shown to be the case for most of the proteins in the proteome of bacteria *E. coli* [113]:

$$p(N; a, b) = N^{a-1} e^{-N/b} / \Gamma(a) b^a \tag{5.16}$$

where N is the total number of GRs on a spore, a is the shape parameter and b is the scale parameter. This can be verified by tagging GRs with fluorescent proteins and measuring the fluorescent intensities of each spore [118]. Indeed, preliminary data showed distributions of the fluorescent intensity could be fit with Eq. 5.16 and yielded shape parameters a ranging between 2-5 (courtesy of P. Setlow). More experiments and analyses are required to obtain a better measure of the parameters, as the effects of the fluorescent tags on the expressions of the GRs are still unclear. Here I will leave a as a tunable parameter (as will be shown below, the value of b is irrelevant in this model).

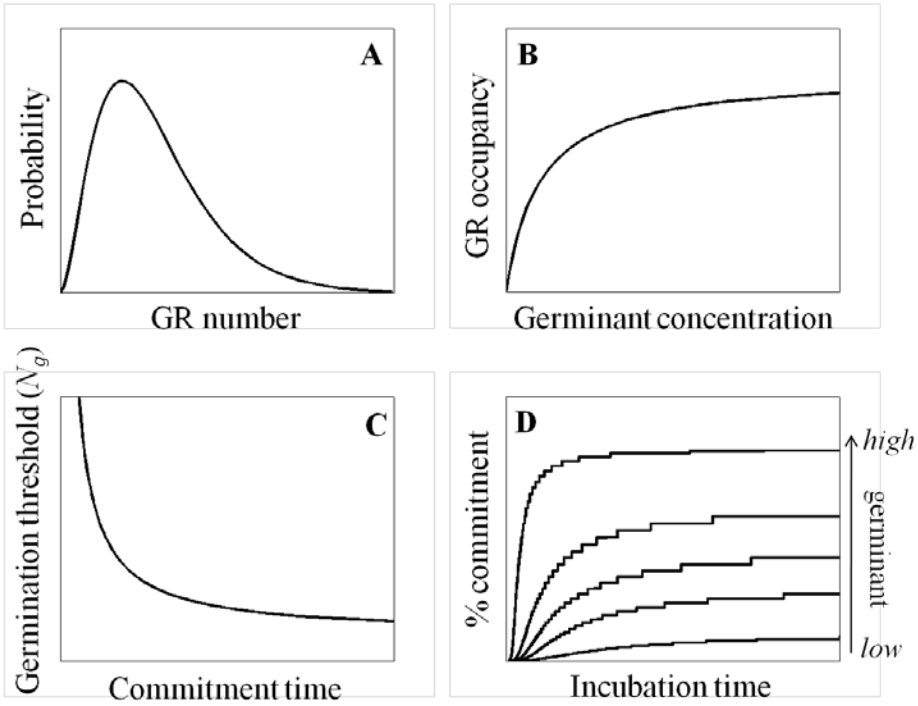

Figure 5.3: Model for spore germination with nutrient germinants.

(A-C) Components of the model. (A) Distribution of number of GRs of a given type per spore in a spore population. (B) Fraction of GRs bound to germinant as a function of germinant concentration. (C) Minimum number of bound GRs required for a spore to commit to germination by certain time. (D) Schematic diagram of germination kinetics generated by the model for different germinant concentrations. The speed and extent of germination increase as germinant concentration is increased.

I also assume GRs are activated by germinant binding (Fig. 5.3B). In addition, binding and activation are assumed to be fast in comparison to the time required for germination. This step was only minimally studied by Woese et al., probably due to the unclear nature of the "germination enzyme" at that time. The number of bound GRs in a spore N^* is the product of the total number of GRs N in the spore and the average GR occupancy $h(c)$ at the given germinant concentration c,

$$N^*(c) = N \times h(c). \tag{5.17}$$

Since we do not know the mechanism of germinant-GR interaction, I will use the Hill equation, derived in Section 5.3.1, to describe the binding of germinants to GRs:

$$h(c) = c^\alpha/(K^\alpha + c^\alpha) \tag{5.18}$$

where α is the Hill coefficient and K is the apparent binding affinity between germinant and GR. One can see that increasing germinant concentration increases levels of bound GRs per spore, as does increasing N.

Our third assumption is that commitment time depends only on the number of bound GRs in a spore. This assumption is a simplification because other spore properties, such as the expression levels of other components of the germination pathway or spore water content are also likely to affect the germination rate [7]. To capture the phenomenon that the percentage of spores that germinate decreases as the nutrient concentration is decreased, an activation threshold should exist, so that below a threshold number of bound GRs germination does not occur, but above the threshold commitment time decreases with increasing numbers of bound GRs, N^*. Another way of expressing this assumption is that the *germination criterion* for a spore to commit to germination before time t is

$$N^* > N_g(t) \tag{5.19}$$

where $N_g(t)$ decreases monotonically over time (Fig. 5.3), so that the more bound GRs in a spore, the faster it commits to germination.

Now I combine the three assumptions together. Using Eqs. 5.17 and 5.19, the germination criterion can also be written as

$$N > N_g(t)/h(c) \tag{5.20}$$

Then using Eq. 5.16, the fraction of spores committed to germination at time t with germinant concentration c (Fig. 5.3D) is given by

$$F(t,c) = \sum_{N>N_g(t)/h(c)} p(N;a,b) \tag{5.21}$$

As will be seen below, it is more convenient to treat N as a continuous variable, then

$$F(t,c) = \int_{N_g(t)/h(c)}^{\infty} p(N;a,b)dN \tag{5.22}$$

Since the gamma distribution has the property that $p(N;a,b)dN = p(N/\mu\,;a,b/\mu)d(N/\mu)$, where $\mu = ab$ is the average number of GRs per spore, we have

$$F(t,c) = \int_{n_g(t)/h(c)}^{\infty} p(n;a,1/a)dn, \tag{5.23}$$

where $n \equiv N/\mu$ is the normalized number of GRs on a spore, $n_g(t) \equiv N_g(t)/\mu$ is the normalized germination threshold. Equation 5.23 enables us to apply the model without knowing the absolute levels of GRs in the spores, which simplifies the fitting procedure.

5.5 DOUBLE GERMINANT MODELS

To extend the model to the case of two germinants activating two different types of GRs, here I generalize the three key elements of the single germinant model.

For GR levels, each GR should still follow the gamma distribution, but there is the new question that whether the levels of the two GRs on the same spore are correlated. Correlations are likely to exist since the expressions of different GRs are regulated by the same transcription factors during sporulation, including both the same RNA polymerase sigma factor, σ^G, as well as the DNA binding protein SpoVT [119]. This can be verified by tagging different GRs with fluorescent proteins of different color and measuring the fluorescent intensities from different GRs on the same spores [118]. Indeed, preliminary data showed correlation coefficients between 0.3-0.4 (courtesy of P. Setlow). More experiments and analyses are required to obtain a better measure of the levels of the GRs and their correlation, as the effects of the fluorescent tags on the expressions of the GRs are still unclear. Here I will assume the levels of different GRs are correlated and leave the correlation coefficient as a tunable parameter.

For GR-germinant interaction, there is the question of whether the activation of one GR affects the activation of a different GR. Such effects are seen in the chemosensing system in bacteria of a number of species, and many chemosensory proteins are associated in large arrays in the plasma membrane of bacteria, with this association allowing direct cooperative interactions between different proteins [120-123]. Indeed, recent work has shown that the great majority of a spore's GRs are co-localized in one small region of the spore's inner membrane [118], and this seems likely to promote cooperative interactions. Cooperativity between different GRs is also strongly suggested by the requirement for both GerB and GerK GRs for germination of wild type *B. subtilis* spores with the AGFK mixture [80]. For the spore strain studied in this

dissertation, FB10, we need to consider possible interactions between GerA and GerB*. However, the point mutations giving rise to the GerB* GRs may disrupt interactions between them, just as they eliminate the GerK GR requirement in order for the GerB* GR to respond to L-asparagine alone [107, 108]. Here I start from the assumption that different GRs do not interact, so Eqs. 5.17 and 5.18 are still applicable, but different GRs have different values for the parameters K and α. This simplification will be validated when the model predictions are compared with experiments, and I leave more complex models for future study.

Finally, I generalize the germination criterion with the help of phase diagrams of germination (Fig. 5.4). With a single germinant (Fig. 5.4A), the criterion for a spore to commit to germination before time t is assumed to be Eq. 5.19, which can also be written as

$$n^*/n_g(t) > 1 \tag{5.24}$$

where $n^* \equiv N^*/\mu$ is the normalized number of GRs on the spore. With double germinants, the new criterion should fall back to Eq. 5.24 when the concentrations of one germinant is set to zero, so that it is $n_1^*/n_g^{(1)}(t) > 1$ if $n_2^* = 0$ and is $n_2^*/n_g^{(2)}(t) > 1$ if $n_1^* = 0$, where the subscripts/superscripts 1 and 2 denote to the two GRs involved respectively. In addition, a recent experiment [8] found that percentages of *B. subtilis* spore germination with mixtures of low concentrations of germinants acting on different GRs were much higher than the sums of the percentages of germination with individual germinants alone (Fig. 5.5). This phenomenon was not seen with spores lacking GRs responsible for recognizing one or several components of the germinant mixtures. Therefore, different GRs function synergistically in triggering germination. Given the above constrains, there are still many ways different GRs could function together, such as the random model shown in Fig. 5.4B. Instead of enumerating all the possibilities, here I only

consider two representative ones, following the rule that no complex mechanism should be introduced if simple ones are able to explain the data.

(I) SUM Model (Fig. 5.4C). Spores "count" both GRs and base their decisions on the collective number of bound GRs, and germinate if

$$n_1^*/n_g^{(1)}(t) + n_2^*/n_g^{(2)}(t) > 1 \tag{5.25}$$

The underlying assumption of this model is that bound GRs are activated and initiate certain downstream signals. When mixtures of germinants are used, the activation of one type of GR is independent of the activation of other GRs, but their downstream signals are summed together. Finally, once the total signal strength reaches certain threshold, the spore commits to germination.

(II) OR Model (Fig. 5.4D). Spores "count" both GRs but only base their decisions on the dominant GR, and germinate if

$$n_1^*/n_g^{(1)}(t) > 1 \quad \text{or} \quad n_2^*/n_g^{(2)}(t) > 1 \tag{5.26}$$

The underlying assumption is that the downstream signals from different GRs are summed separately instead of together, and a spore commits to germination if the signal from either GR is strong enough.

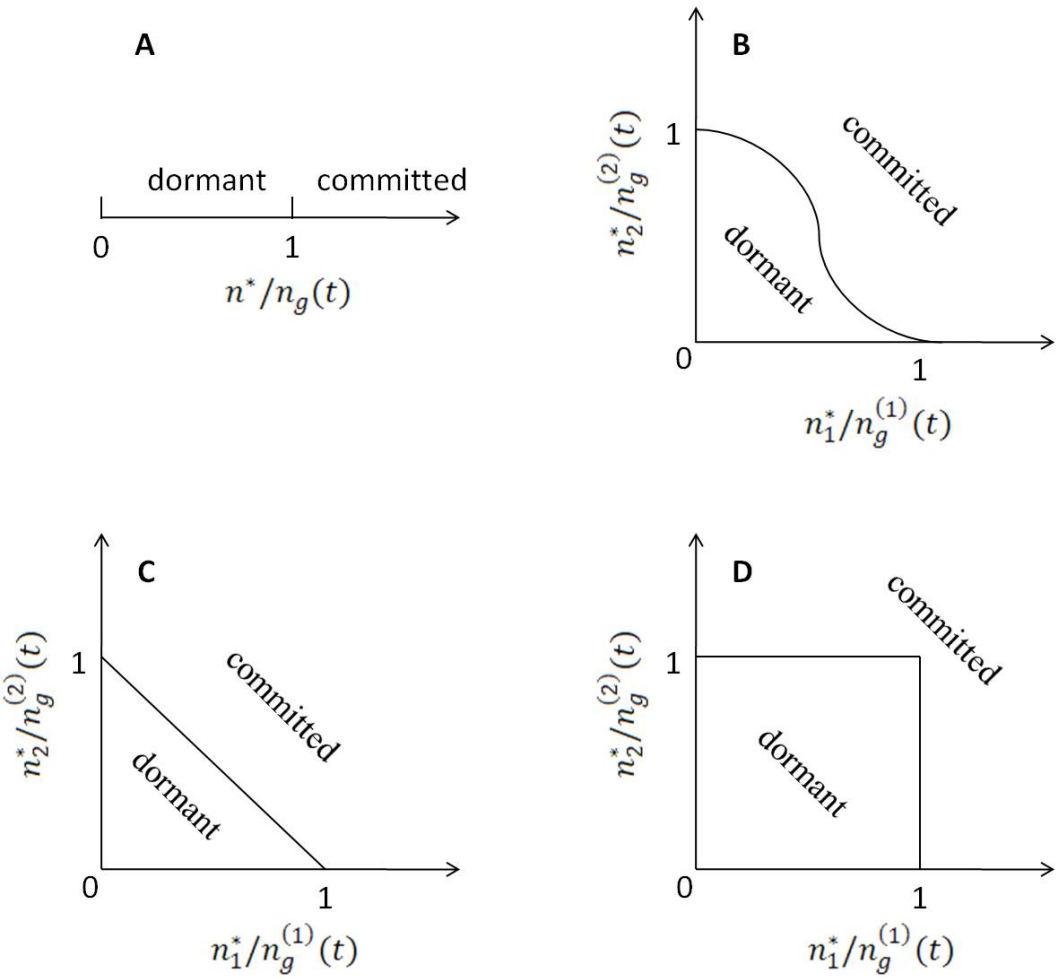

Figure 5.4: Phase diagrams of germination.

According to the model, at any time *t* a spore is in one of two states: dormant or committed to germination. (A) With a single germinant, a spore has committed to germination if $n^*/n_g(t) > 1$, otherwise it is still dormant. With double germinants, the state of a spore depends on both GRs via ($n_1^*/n_g^{(1)}(t)$, $n_2^*/n_g^{(2)}(t)$) and the sate space is divided by a curve connecting (1,0) and (0,1). Shown are (B) a random model, (C) the SUM model, and (D) the OR model.

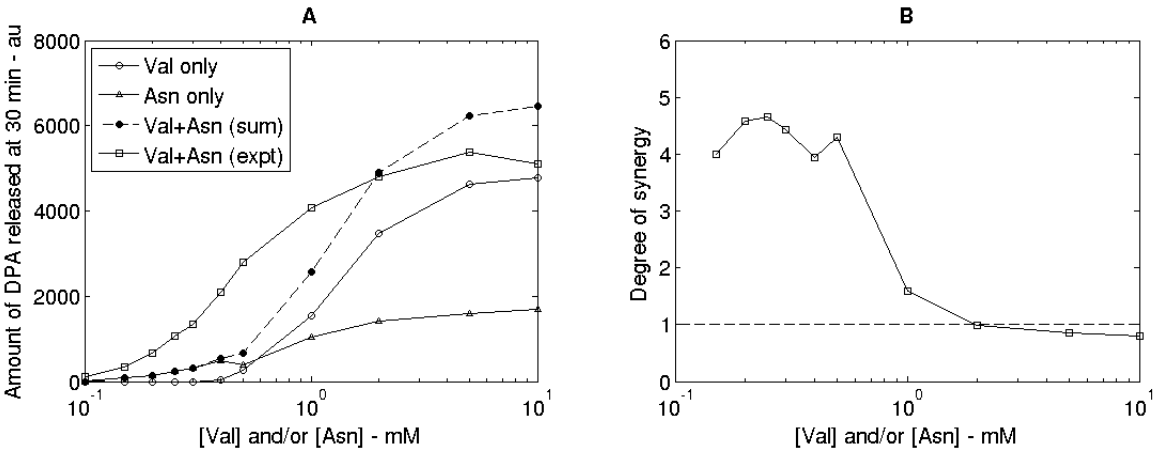

Figure 5.5: Germination of FB10 spores with L-valine or/and L-asparagine.

A) Amount of CaDPA release from *gerB* B. subtilis* spores germinating with L-valine and/or L-asparagine and B) degree of synergy when both germinants were used. Spores of *B. subtilis* FB10 (has the GerB* GR) were germinated with various concentrations of L-valine and/or L-asparagine, and amount of CaDPA (in arbitrary units) released at 30 min after addition of germinants are shown (similar results were obtained for other time points). The symbols used in A) are: ○ - L-valine germination; △ - L-asparagine germination; ● - sum of the rates with L-valine and L-asparagine germination alone; and □ - experimental curve with L-valine plus L-asparagine germination. Degree of synergy was defined as the ratio between experimentally measured amount of CaDPA release (proportional to percentage of germination) with L-valine plus L-asparagine germination and the sum of the corresponding amounts of CaDPA release with L-valine and L-asparagine germination alone. (Collaboration with X. Yi and P. Setlow.)

5.6 RESULTS

In this section, I apply the models to FB10 spores. The procedure is illustrated in Fig. 5.6. Experiments are performed by our collaborators, where spores are germinated with L-valine or/and L-asparagine. Note L-valine activates the GerA GR and L-asparagine activates the GerB* GR. First, I will extract the model parameters by fitting the single germinant model separately to germination data with L-valine alone and data with L-asparagine alone. Then I will apply the parameters to the double germinant models to predict germination with mixtures of L-valine and L-asparagine, and validate the predictions against experiments, thus allowing me to determine which mechanism might be used to process germination signals inside spores.

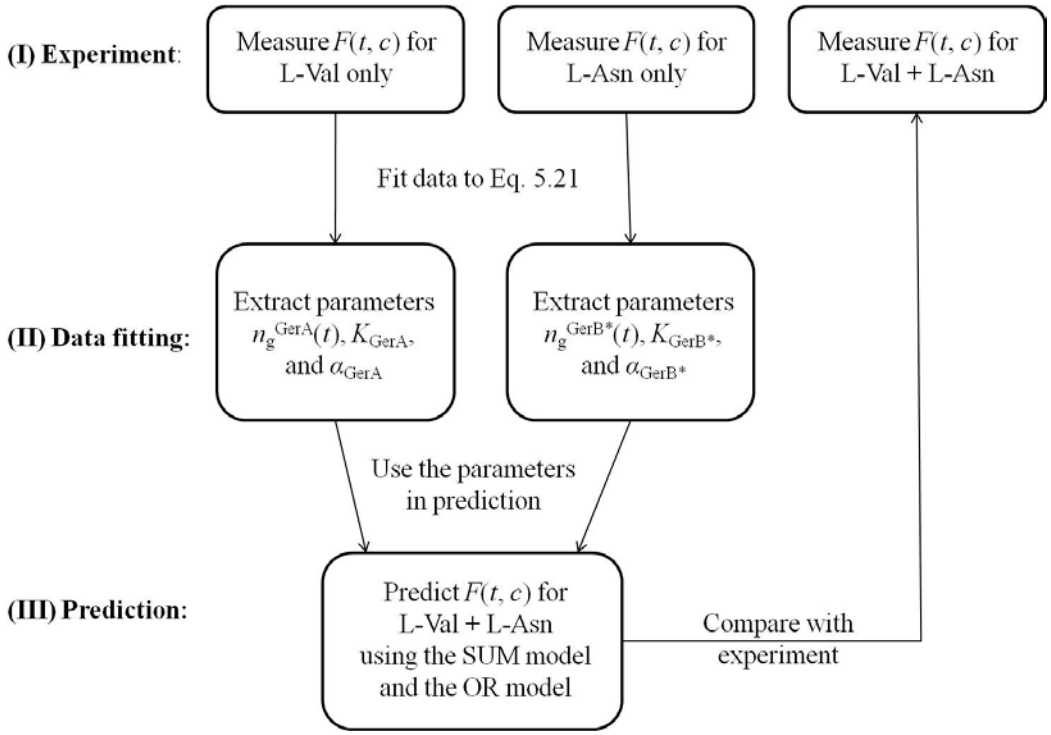

Figure 5.6: Flowchart of model implementation.

5.6.1 Experiments

Germination kinetics are often sensitive to spore preparation procedures, e.g. percentages of germination at given incubation time may differ significantly between different trials, making it unreliable to mix data from different trials. Thus high throughput experiments are critical for obtaining data for the quantitative models presented here, as I need data for $F(t, c)$ – fraction of commitment – over a broad range of germinant concentrations. However, high throughput assays measuring $F(t, c)$ are not yet available. Here I use CaDPA release data, which are relatively easy to obtain, to approximate the commitment data. Note that the approximation is only reasonable at incubation times longer than ~17 minutes (Fig. 5.7), since at earlier times the delay between commitment and CaDPA release becomes significant enough that the percentage of spores having released CaDPA is far lower than the percentage of spores having committed to germination.

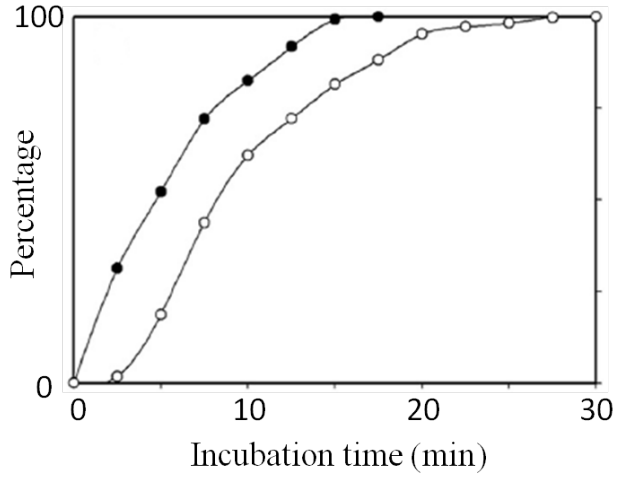

Figure 5.7: Percentages of commitment and CaDPA release during germination of FB10 spores. Commitment and CaDPA release are labeled with ● and ○ respectively. Spores were germinated in 150 µM L-asparagine. (Source: Adapted from [99].)

Experiments were performed (by our collaborators) in a 96-well plate, so that germination under 96 different conditions could be measured simultaneously. Each well contained 100 μl of solution with around 10^7 spores and a fixed amount of Tb^{3+}, and germination was measured by recording the fluorescence emission from Tb-DPA [97, 98], the intensity of which is proportional to the amount of CaDPA released by the spores into the solution. Figure 5.8 shows the CaDPA release of FB10 spores with various concentrations of L-valine or/and L-asparagine, which starts soon after the addition of germinants and lasts for more than 30 minutes. Since the duration of the CaDPA release for individual spores is relatively short (~3 minutes, Fig. 4.4), the fluorescence intensity is roughly proportional to the amount of spores that have released their CaDPA. One can see that germination is heterogeneous, with some spores germinating in the first few minutes but others taking more than half an hour. In addition, the percentage of germination increases and the average germination time decreases as the nutrient levels is increased.

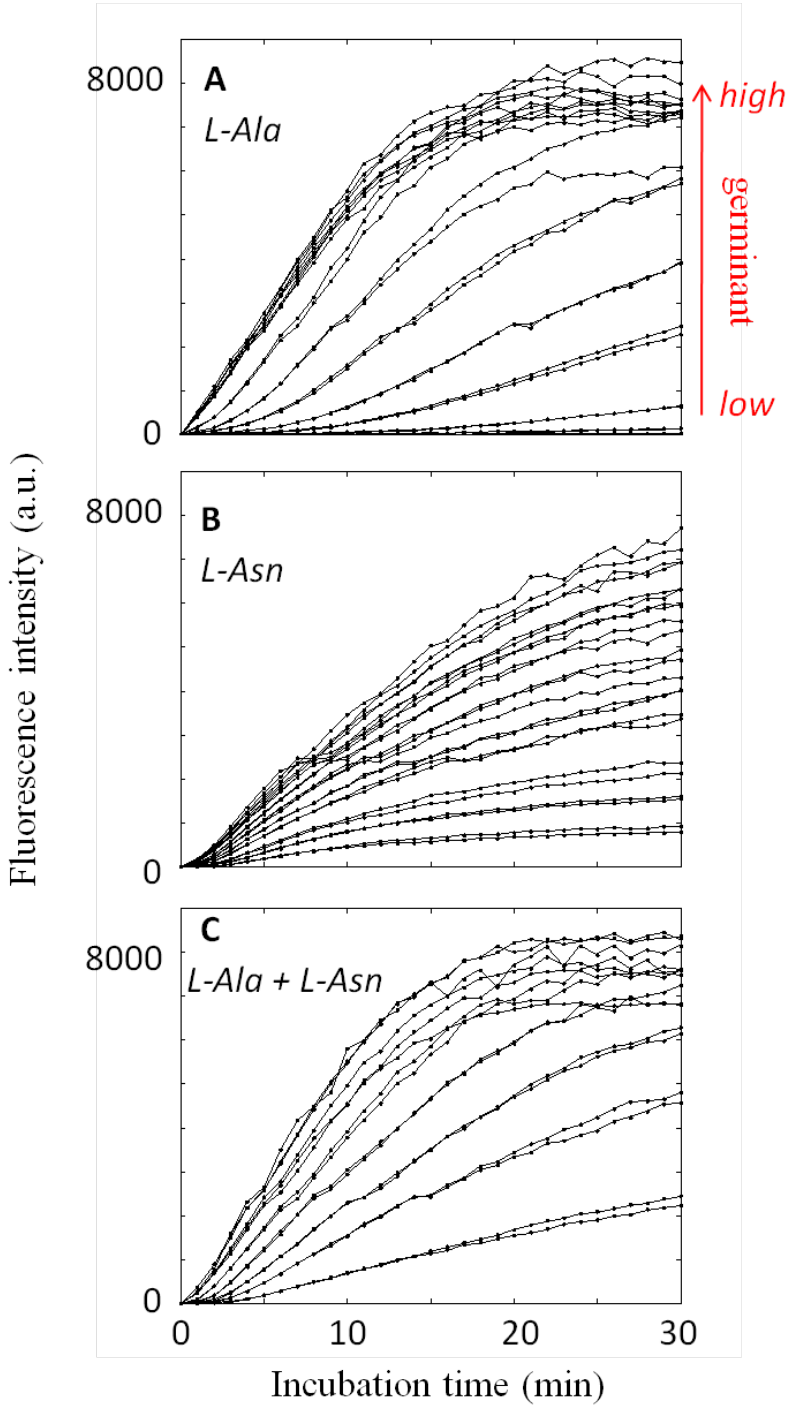

Figure 5.8: Germination of FB10 spores.

The germinants are (A) L-alanine, (B) L-asparagine, and (C) equal concentrations of L-alanine plus L-asparagine. Germinant concentrations range from 0.1 to 100 mM. Two trials were performed for each concentration. Data provided by X. Yi and P. Setlow.

5.6.2 Data fitting

Here I extract the apparent binding affinity between germinant and GR K, the Hill coefficient of the binding α, and the function $n_g(t)$ from the single germinant data. Equation 5.23 is fit to the data shown in Fig. 5.8A and B respectively (a MATLAB script carrying out the fitting is included in Appendix A). To convert the fluorescence intensity to fraction of germination, I assume 100% of the spores are germinated at saturating concentrations of L-valine plus L-asparagine (10 mM each), since few spores are superdormant (take extremely long time to germinate) to high concentrations of nutrient mixtures [85, 87]. Figure 5.9 shows the extracted parameters for L-valine (binding to GerA) and L-asparagine (binding to GerB*) germination respectively. Note that only parameters for the time period between 17 and 30 min are shown, due to limitations of the experiment (the experimental readings far underestimated the actual amount of committed spores at earlier stage, and were likely complicated by the outgrowth of germinated spores at later stage).

The estimated parameters include the apparent binding affinity K and the Hill coefficient α between germinants and their cognate GRs, and the corresponding threshold $n_g(t)$ for a spore to commit to germination by time t. First, n_g, as expected, decreases with time, i.e., the more bound GRs on a spore, the less time it takes to commit to germination. GerA showed lower n_g values than GerB* for all the time points, which raises the question of whether the two GRs have the same efficacy in triggering germination, e.g., given that a spore with a certain number bound GerA needs a certain amount of incubation time to commit to germination, will it take the same amount of time to commit to germination if it had the same number of bound GerB* instead? To answer this question, the average level of GerA on a spore relative to that of GerB* will be

needed. Second, the values of K fall in the typical range for protein-ligand binding (Fig. 3.10), and the binding of L-asapragine to GerB* is almost an order of magnitude stronger than that of L-valine to GerA. Finally, the Hill coefficients α were close to 1, suggesting that each GR may have only one binding site and function as monomer. Note, I did not assume K and α were constants over time, but extracted their values for different time points independently. The results suggest that it is also reasonable to treat them as constants. Finally, K, α, and $n_g(t)$ are the minimum set of parameters to describe the data, as they determine the inflection point, the steepness, and the final height of the curve in Fig. 5.9A respectively.

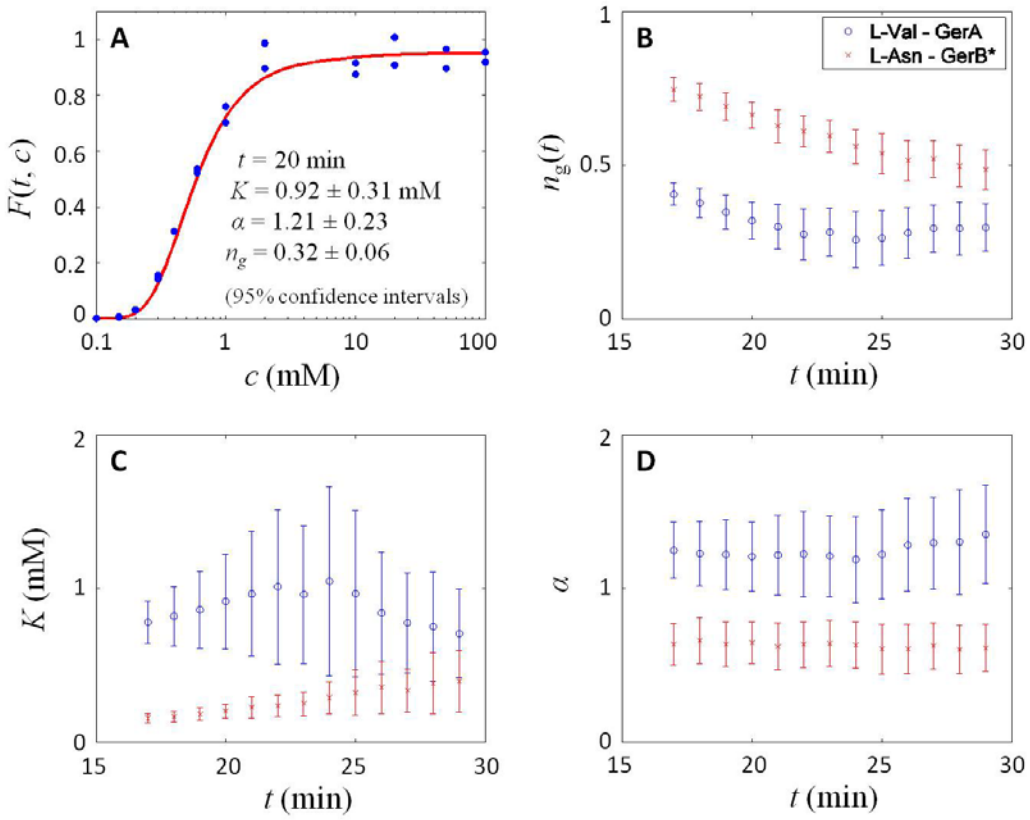

Figure 5.9: Data fitting.

(A) Fitting of Eq. 5.23 to the percentages of CaDPA release at 20 min yielded the values of the model parameters at 20 min. Repeating this procedure for all the time points ranging between 17 and 30 min yielded the values of (B) n_g, (C) K, and (D) α as functions of time. The tunable parameter a (for gamma distribution) is set to 3.5.

5.6.3 Model prediction

In this section, I use the extracted parameters (Fig. 5.9) to predict experiments with double nutrients (Fig. 5.8C). In other words, I use the single germinant data at a given time point (Fig. 5.10, plotted with symbols + and ○) as input to predict germination with double nutrients at the same time point and validate the prediction against the corresponding data (Fig. 5.10, plotted with symbol ●).

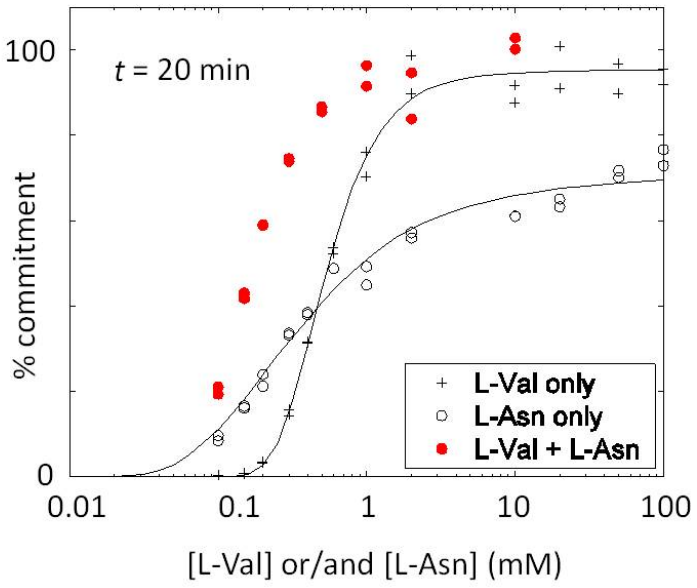

Figure 5.10: Percentages of FB10 spores committed to germination at 20 min with various concentrations of L-valine or/and L-asparagine.

The solid lines are plotted with Eq. 5.23 using the extracted parameters (at 20 min) shown in Fig. 5.9.

The prediction procedure is as follows. (I) Generate 10,000 "spores" by generating 10,000 pairs of random numbers (n_{GerA}, $n_{\mathrm{GerB*}}$) with each pair representing the normalized number of GerA and GerB* GRs on a spore (Fig. 5.11). Both n_{GerA} and $n_{\mathrm{GerB*}}$ follow the gamma

distribution $p(n; a, 1/a)$, and as a simplification I assume the same shape parameters a for GerA and GerB*. In addition, n_{GerA} and $n_{\text{GerB*}}$ are correlated with correlation coefficient r. The algorithm generating random numbers following this joint gamma distribution is described in Appendix B. (II) At given germinant concentrations ($c_{\text{L-Val}}$, $c_{\text{L-Asn}}$) and incubation time t, calculate the normalized number of bound GRs (n^*_{GerA}, $n^*_{\text{GerB*}}$) using Eqs. 5.17 and 5.18 and the corresponding parameters values in Fig. 5.9. (III) The model predicts that the fraction of spores committed to germination at time t with double germinant (n^*_{GerA}, $n^*_{\text{GerB*}}$) is the fraction of "spores" satisfying Eq. 5.25 (the SUM model) or Eq. 5.26 (the OR model).

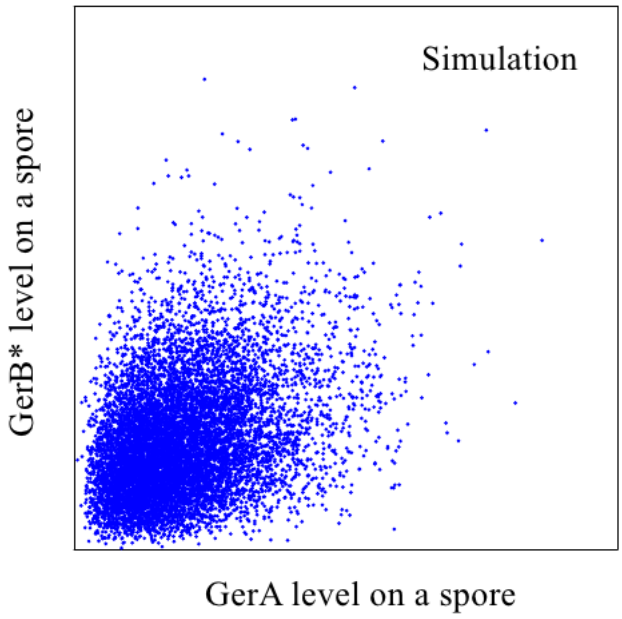

Figure 5.11: Scatter plot of the normalized numbers of GerA and GerB* on individual spores.

($a = 3.5, r = 0.35$.)

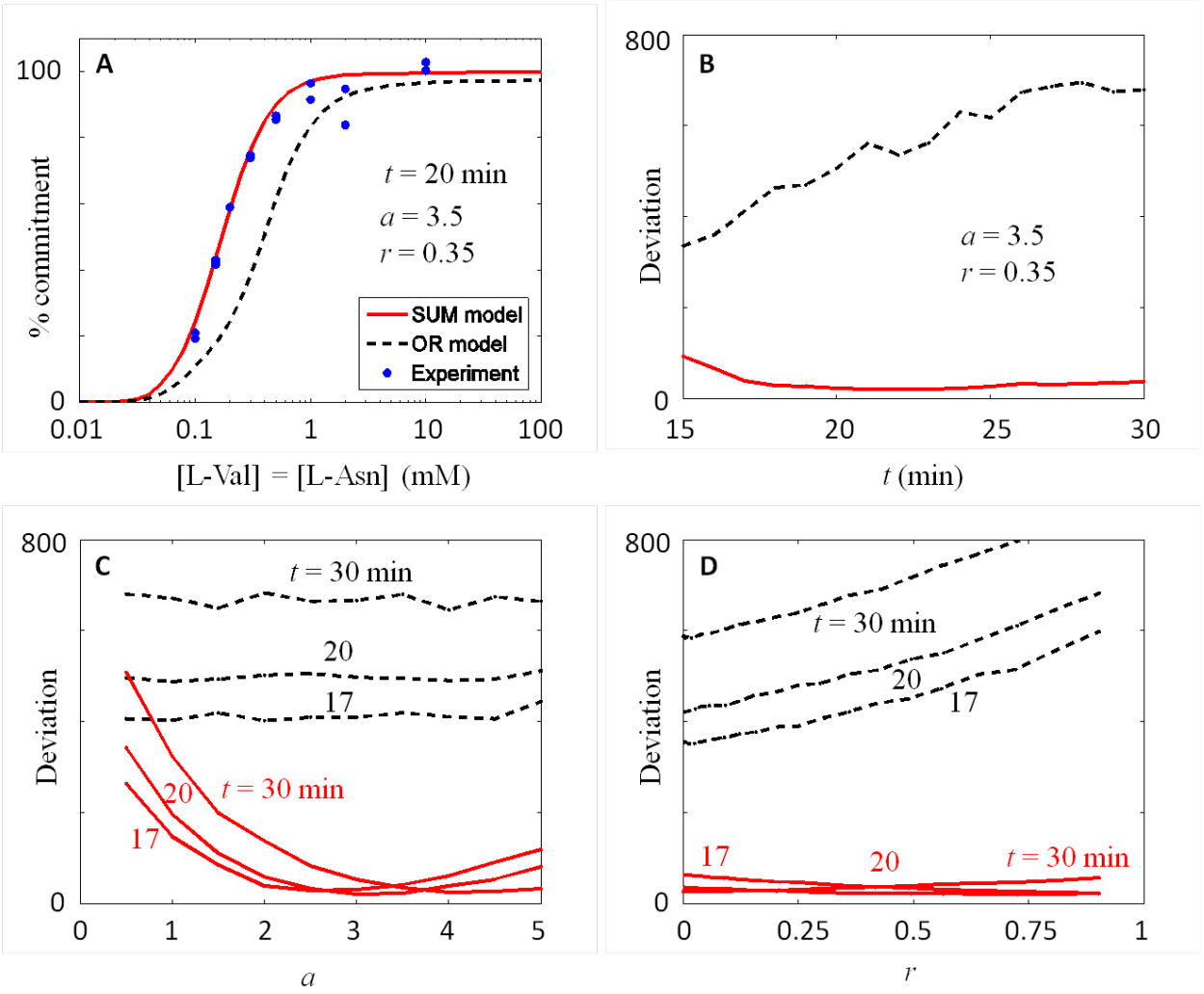

Figure 5.12: Model predictions for the germination of FB10 spores with equal concentrations of L-valine and L-asparagine.

(A) Percentage of commitment at 20 min after the addition of germinants. The models assumed the shape parameter $a = 3.5$ and correlation coefficient $r = 0.35$. (B-D) Deviations (mean square difference) between model prediction and experiment. (B) Deviations at different incubation time (assuming $a = 3.5$ and $r = 0.35$). (C) Deviations at 17, 20, and 30 min as functions of a (assuming $r = 0.35$). (D) Deviations at 17, 20, and 30 min as functions of r (assuming $a = 3.5$).

Figure 5.12A shows the model predictions compared with experiment at 20 min after the addition of germinants. It is clear that the data is better fit by the SUM model than the OR model, and similar results were obtained for all the time points between 17 and 30 min (Fig. 5.12B). Note I assumed $a = 3.5$ for the shape parameter of GerA/GerB* level distribution (Eq. 5.16) and a correlation coefficient $r = 0.35$ for the correlation between the levels of GerA and GerB* on the same spore. Figures 5.12C and D show that the predictions of the SUM model are optimal at around $a = 3.5$ and are insensitive to the correlation coefficient r, while the OR model fails over the whole range of parameters.

5.6.4 Goodness of fit and prediction

In Section 5.6.2, I obtained the model parameters from fitting of the model to experimental data. The fitting procedure yielded a set of parameter values that minimize the mean square difference between the model and the data (see Appendix A). However, it is not the only set of parameter values that could describe the data. Instead, there should be a whole range of parameter values that may agree with the data with different probabilities [124]. In the following, I carry out an ensemble analysis that reflects this probabilistic view of parameter estimation.

The expected value of some property of a model is given by [125]:

$$E[f(\Theta)|M,Y] = \int_{\Theta_{\min}}^{\Theta_{\max}} f(\Theta)P(\Theta|M,Y)d\Theta \tag{5.27}$$

where $f(\Theta)$ is a generic function of the model parameters, and $P(\Theta|M,Y)$ is the probability distribution of Θ given the model M and data Y. Using Bayes theorem

$$P(\Theta|M,Y) = \frac{P(Y|\Theta,M)\cdot P(\Theta|\mathrm{M})}{P(Y)} \tag{5.28}$$

where $P(Y|\Theta, M)$ is the conditional probability of simulating data Y given Θ and M, $P(\Theta|\mathrm{M})$ is the probability distribution of Θ prior to any knowledge about data Y, and $P(Y)$ is the evidence for a model. Both $P(Y)$ and $P(\Theta|\mathrm{M})$ can be considered as a constants [124]. Finally, it is reasonable to assume that $P(Y|\Theta, M)$ is representable as a multivariate Gaussian [124]

$$P(Y|\Theta, M) = \text{const} \times \exp\left\{-\sum_i \frac{[Y_i - M_i(\Theta)]^2}{2\sigma_i^2}\right\} \tag{5.29}$$

with σ_i denoting the standard deviation of data Y_i.

The distribution $P(\Theta|M, Y)$ can be estimated using the Metropolis-Hasting algorithm [124, 126, 127], which is a Monte Carlo technique that could generate random walks in the parameter space for Θ following to the desired distribution. The procedure is as follows: (1) Select with equal probability one of the parameters Θ_m, with $\Theta_m \in \{K, \alpha, n_g(t_1), n_g(t_2), \dots\}$. Note that based on the results in Fig. 5.9, here I assume the parameters K and α are constant over time; (2) Propose an update from Θ_m to $\Theta'_m = \Theta_m + \Delta_m$, where Δ_m is drawn with constant probability from an interval $[-\Delta_m^{\max}, \Delta_m^{\max}]$; (3) Accept the proposed step with probability min[1, $P(Y|\Theta', M)/P(Y|\Theta, M)$], where $\Theta' = (\Theta_1, \Theta_2, \dots, \Theta'_m, \dots)$. If the proposed step is accepted, the next Θ in the random walk is $\Theta^+ = \Theta'$, otherwise $\Theta^+ = \Theta$. At the beginning of a random walk, Θ is randomly initialized. Each random walk is consisted of 10^4 "warm-up" steps to equilibrate followed by 2×10^5 accumulation steps, and Θ is sampled after every 200 steps so that an ensemble of 1,000 sets of parameters are collected. The values of $\Delta_m^{\max}$ are chosen so that the acceptance ratio is around 0.3. I also assume σ_i = 0.1, i.e., the experimental data on percentage of germination is accurate up to $\pm 10\%$, which should be an upper bound.

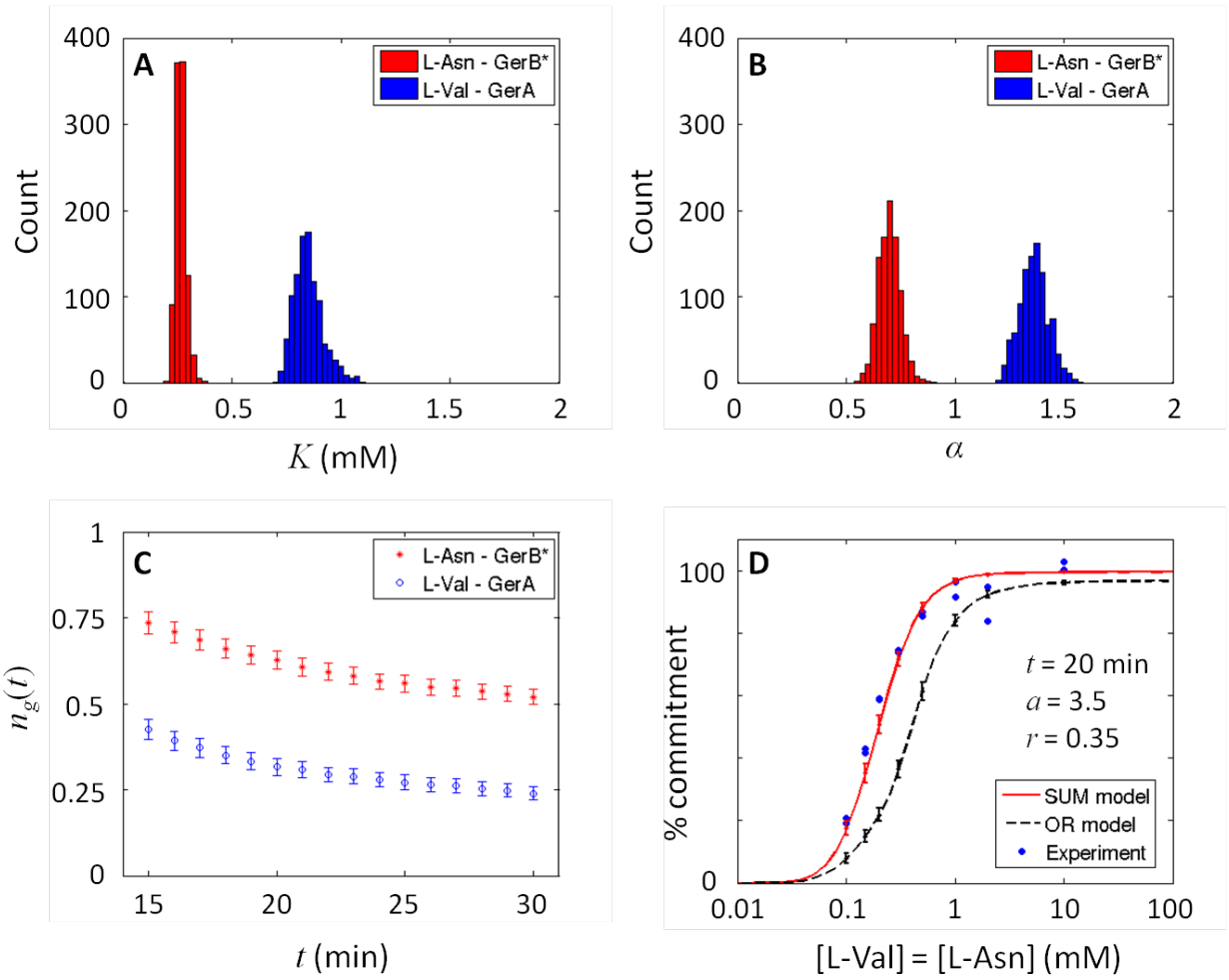

Figure 5.13: Data Fitting and model prediction using ensemble method.

(A) Histograms of *K* for the binding of L-asparagine to the GerB* GR and the binding of L-valine to the GerA GR respectively. (B) Histograms of α. (C) Mean value and standard deviation (plotted as error bar) of germination criteria as function of commitment time. (D) Predictions of the SUM model and the OR model using the ensemble of parameters shown in (A-C). Each parameter set from the ensemble gives rise to one prediction, the lines and the error bars show the ensemble average and the standard deviations of the predictions. Experimental data is also shown to compare with the model.

The ensemble of parameters estimated using the Monte Carlos method is shown in Figs. 5.13A-C. The parameters K and α are plotted in histograms, which reflect the likelihood of the parameter values representing the data given the model. The values obtained are consistent with that shown in Fig. 5.9, and the widths of the histograms provide a measure of the goodness of fit. The ensemble average and standard deviation of the parameter $n_g(t)$ for the time range 15-30 mins is also shown, and the result is also consistent with that shown in Fig. 5.9, but with noticeably smaller error bars. The parameters are then used to make predictions, and Fig. 5.13D shows the SUM model shows good agreement with experiment for the whole ensemble of parameters.

5.7 DISCUSSION

The agreement of the SUM model with experiments suggests that spores "count" their total number of bound GRs and use this information as a major factor in deciding whether or not to germinate. The mechanism by which spores count is not known, but there are at least two possibilities.

The first possible mechanism is that any single active GR has a certain probability to trigger germination, and the more active GRs there is the more likely a spore is to germinate and the faster the process is. (As an analogy, any activated grenade in an arsenal has certain probability to trigger the explosion of the entire arsenal, and the more activated grenades there are the more likely and the sooner is the explosion). However, this mechanism predicts that spores are unstable since the GRs may maintain a low level of activity either due to trace levels

of germinants in the environment or to spontaneous GR activation, and this contradicts the fact that spores can stay dormant for extremely long periods.

The second possible mechanism is that germination is determined by the total number of active GRs on a spore, i.e., signals from all the active GRs are integrated inside the spore, and germination is triggered when the total signal strength is above a certain threshold. There is currently no direct evidence for such an "integrator", although it has been suggested that the GerD protein, loss of which greatly decreases germination via GRs, might serve such a function [128]. However, since there was synergism between the GerA and GerB* GRs in germination when GerD was absent, GerD is not likely to be the "integrator" postulated above. In addition, GerD appears not to be present in spores of *Clostridium* species [129], and although synergism between multiple germinants has not been investigated in spores of *Clostridium* species, one might expect such a phenomenon to be widespread. A second way that signals from different germinants could be integrated is via some major germination event itself, such as CaDPA release. The release of the great majority of CaDPA during the germination of individual spores takes only a few min, although this is preceded by a long lag period (T_{lag}), the length of which varies considerably between individual spores [100]. While all events that take place during T_{lag} are not known, there is a slow release of CaDPA during this period [130]. This slow CaDPA release in T_{lag} would likely alter spore properties such as core water content and the strain on the spore's peptidoglycan cortex, and these changes could then trigger fast CaDPA release and completion of spore germination [80, 130]. In this scenario perhaps the whole spore core/cortex functions as the integrator, and there is no dedicated integrator molecule.

The model provides a simple explanation for superdormant spores, which is that the superdormant fraction comes from the low-expression end of the GR distribution (Fig. 5.3A).

This is supported by the experimental observation that superdormant spores germinate normally with CaDPA or dodecylamine [85, 87], which do not involve the activation of the GRs, but remain dormant with respect to the initial nutrients used to isolate them. What is less clear is why superdormant spores are also less responsive to other nutrients that stimulate different GRs yet germinate relatively normally with nutrient mixtures that stimulate multiple GRs [85]. One possible reason is that the levels of different GRs on single spores are correlated, so that a spore with low levels of one GR is also likely to have low levels of other GRs. Another possibility is that the diminished responses of superdormant spores to single nutrients combined with normal responses to nutrient combinations is a reflection of cooperative interactions between different GRs.

The model also provides the basis for including the effects of other factors, such as heat activation or pH, on the kinetics or extent of germination. Heat activation is a common laboratory procedure of applying sub-lethal heat shock on spores before germinants are added (for the experiments used in this study, spores were heat activated for 30 min at 75°C and cooled on ice for $\leq$ 1 hour before use [8]). It is known that this procedure leads to increased germination rates but the reason is still unclear. The model suggests two possible explanations: (1) Heat activation increases the affinity of GRs for their cognate germinants, thus increasing the number of bound GRs in a spore for a given germinant concentration; (2) Heat activation increases the efficacy of each bound GR, shortening commitment time for a given number of bound GRs and/or decreasing the germination threshold. The latter seems to be favored by the fact that non-heat-activated spores give much higher yields of superdormant spores than do optimally activated populations [85]. However, the two possibilities may coexist, and other factors not yet considered in the model may also play some role.

6.0 CONCLUSIONS AND OUTLOOK

6.1 INTRINSICALLY DISORDERED PROTEINS

Through a thermodynamic model, I showed that evolution may act differentially upon the level of disorder for proteins of different functions, which is supported by a genome-wide survey of disorder. For proteins whose main function is to bind other proteins, the amount of disorder that can be tolerated without degrading function is quite broad, depending on the complementarity of the interaction. Catalytic proteins have a strong preference for a stable folded state, consistent with the notion that catalysis has strong conformational requirements. More interestingly, disorder can be used to maximize the specificity of promiscuous interactions relevant to transcription and signal transduction. An interesting hypothesis arising from the theory is that lower affinity interactions are expected to involve proteins with less disorder, which may help explain why disorder is less prevalent in prokaryotes than eukaryotes. The hypothesis is consistent with a preliminary survey of protein-ligand interactions using the PDBbind database that suggests that bacterial proteins bind small ligand molecules more weakly than humans proteins.

There is still questions about the role of disorder in eukaryotic transcriptional proteins and prokaryotic proteins. Eukaryotic transcription factors have long been noticed to have ordered DNA binding domains and disordered transcription activation domains [21]. The genome wide

survey further confirmed the ubiquitous presence of disorder in eukaryotic transcriptional proteins, in contrast to the lack of disorder in prokaryotes. To understand the difference, it is useful to compare how transcription is activated in different ways in the two types of organisms [2]. Prokaryotic transcription activators generally have separate activation and DNA-binding regions, where the activation region interacts and recruits RNA polymerase onto the DNA. Mutant transcription factors that bind DNA but could not interact with RNA polymerase do not activate transcription. Eukaryotic transcription activators, in contrast, rarely activate transcription through direct interaction with RNA polymerases. Instead, they recruit complex transcription machinery which in turn recruits the RNA polymerase. It has been suggested that, by being disordered, transcription factors could rapidly recruit the components of the transcription machinery through a fly-casting mechanism [131], and facilitate the assembly of the complex and the binding to the DNA. The theory presented here further suggest that disorder may also facilitate the process by tuning the binding affinities between the transcription factors and maximizing the discrimination between cognate and non-cognate binders.

6.2 BACTERIAL SPORE GERMINATION

I analyzed germination of FB10 spores with L-valine or/and L-asparagine – a simplified case where only the GerA and GerB* GRs are involved. I proposed a model (Fig. 6.1A) assuming: (1) There is no interaction between the GerA and GerB* GRs; (2) GRs are activated by germinant binding, and the active GRs generate certain downstream signal. The strength of the signal from each type of GR is proportional to the number of the corresponding GRs that are activated. (3) Signals from GerA and GerB* are summed by a common integrator and the total signal strength

is used to determine how fast a spore commits to germination, and the stronger the signal the faster a spore germinates. The agreement between the predictions of the model and the experiments suggests that these are reasonable assumptions. Based on the first assumption, we can conclude that the interactions between GerA and GerB* should be weak if not nonexistent. Figure 6.1B shows an extended model to include this possibility, where a GR can either function alone or together with a different GR. The extended model would be plausible if GRs function as complexes, where both homo-oligomers and hetero-oligomers can be formed.

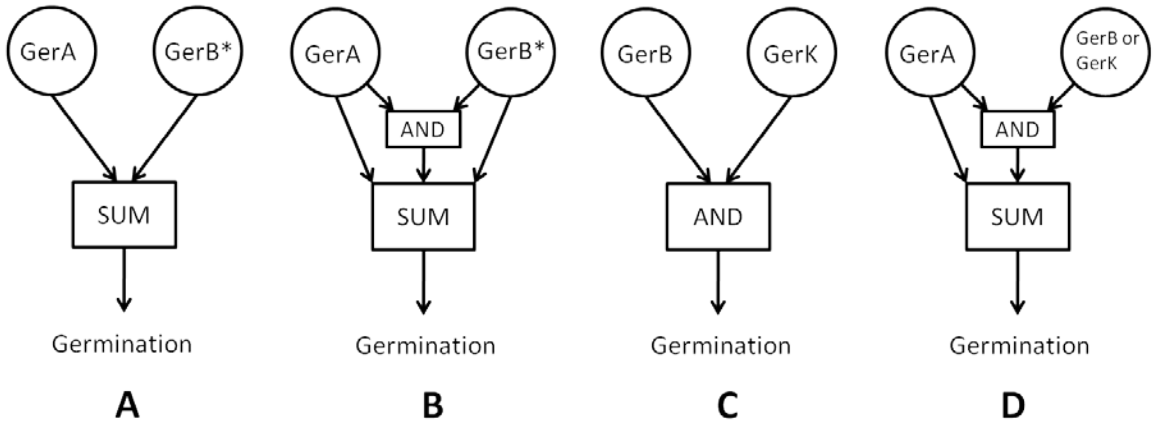

Figure 6.1: Possible mechanisms of signal processing inside spores.

To generalize the model to wild type *B. subtilis* spores, where there is GerB instead of GerB*, and where the GerK GR also play important roles, more complex mechanisms of signal processing need to be considered. Figure 6.1C illustrates a mechanism where two types of GRs need to be activated simultaneously in order to trigger germination, which seems to apply to germination of wild type *B. subtilis* spores with AGFK, where both GerB and GerK are activated, but neither GerB nor GerK alone can trigger germination [80]. Figure 6.1D shows a model where one type of GR can either function alone or together with a second GR, while the

second GR has to function together with the first one. This mechanism seems to apply to the GerA GR in wild type *B. subtilis* spores, where the activation of GerA alone is sufficient to trigger germination, and the rate of germination is increased if either GerB or GerK is also activated, although neither of them can trigger germination alone [108]. Finally, these models seem to be supported by the recent finding that different GRs colocalize in discrete clusters on the inner membranes of spores [118].

APPENDIX A

MATALAB SCRIPT FOR DATA FITTING

```
function spore_fit(filename,gam_a)

% This matlab script fits the single germinant model to experimental data on fractions of
%spores committed to germination at various germinant concentrations.
%
%--INPUT--
%
%filename: name of the file containing experimental data
%
%Data should be in a text file with two columns, with the first column being germinant
%concentrations in units of mM, and the second colum being the corresponding fractions
%of spores committed to germination at a given time point.
%
%gam_a: scale parameter for gamma distribution
%
%The model assumes the number of germinant receptors (GR) on a spore follows gamma
%distribution. Gamma distribution has two parameters: a shape parameter that
%determines the shape of the distribution, and a scale parameter that determines the
%mean value of the distribution. The model only depends on the shape parameter.
%
%--OUTPUT--
%
%The script generates a plot showing the experimental data (fraction of commitment vs.
%germinant concentration), the model parameters from fitting of the experimental data,
%and a curve generated by the model using the derived parameters.
%The model parameters are:
%
%K: apparent binding affinity between germinant and its cognate GR
%alpha: Hill coefficient of the binding
%n_g: normalized number of bound GRs (relative to the average number of GRs on a spore)
```

```
%       required for a spore to commit to germination no later than the given time point.
%
%--EXAMPLE--
%
%spore_fit('data.txt',3)

data = load(filename);
conc = data(:,1); % germinant concentration (mM)
fcom = data(:,2); % fraction of commitment at a given time point

f_type = fittype('1 - gamcdf( (1+(K/c)^alpha)*ng, gam_a, 1/gam_a )','problem',{'gam_a'},...
            'dependent',{'y'},'independent',{'c'},'coefficients',{'K', 'alpha', 'ng'});
f_opt = fitoptions('Method','NonlinearLeastSquares','Lower',[0, 0, 0],'Upper',[Inf, Inf, Inf],...
           'Startpoint',[median(conc), 1, 1]);
f_result = fit(conc,fcom,f_type,f_opt,'problem',{gam_a});

tmp = coeffvalues(f_result);
K    = tmp(1);
alpha = tmp(2);
ng   = tmp(3);

tmp = confint(f_result,0.95);
err_K    = (tmp(2,1) - tmp(1,1)) / 2;
err_alpha = (tmp(2,2) - tmp(1,2)) / 2;
err_ng   = (tmp(2,3) - tmp(1,3)) / 2;

figure(1)
clf
set(gca,'fontsize',16)
tmp_x = 10.^[-2:.01:2];
tmp_y = 1 - gamcdf( (1+(K./tmp_x).^alpha)*ng, gam_a, 1/gam_a );
semilogx(tmp_x,tmp_y,'r','linewidth',2)
hold on
semilogx(conc,fcom,'ob','markerfacecolor','b')
hold off
axis([min(conc) max(conc) 0 1.2])
xlabel('Germinant concentration (mM)')
ylabel('Fraction of commitment')
title(['spore_fit('' filename ''',' num2str(gam_a) ')'],'interpreter','none')
legend(['Model (95% confidence):' ...
    '\newlineK = ' num2str(K,3) ' \pm ' num2str(err_K,2) ' mM' ...
    '\newline\alpha = ' num2str(alpha,3) ' \pm ' num2str(err_alpha,2) ...
    '\newlineN_g/N_{average} = ' num2str(ng,3) ' \pm ' num2str(err_ng,2)], ...
    'Experimental data','location','southeast')
```

APPENDIX B

ALGORITHM GENERATING CORRELATED GAMMA DISTRIBUTIONS

Multivariate gamma distributions can be generated from multivariate normal distributions [132]. Consider $\boldsymbol{X_u} = (X_{1u}, \dots, X_{pu})$ distributed as a p-variate normal vector with mean vector zero and covariance matrix $\boldsymbol{M} = (\sigma_{ij})$. It can be shown that $Z_i = \sum_{u=1}^{n} X_{iu}^2$ follows gamma distribution

$$p(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b}$$

with $a = n/2$ and $b = 2\sigma_{ii}$. Since the X_i's are correlated, hence Z_i's are also correlated and we obtain a possible p-variate gamma distribution.

Now the problem reduces to generating the p-variate normal distribution $\boldsymbol{X} = (X_1, \dots, X_p)$. Suppose $\boldsymbol{Y} = (Y_1, \dots, Y_p)$ is a vector with each element follows the standard normal distribution and no correlation between the elements. Since linear combinations of normal random variables is again normal, we could let $\boldsymbol{X} = \boldsymbol{C}^T \boldsymbol{Y}$, where $\boldsymbol{C}$ be a $p \times p$ matrix, and the problem again reduces to finding $\boldsymbol{C}$ such that

$$\boldsymbol{C}^T \boldsymbol{C} = \boldsymbol{M}$$

Since $\boldsymbol{M}$ is symmetric positive-definite, $\boldsymbol{C}$ is the Cholesky decomposition of $\boldsymbol{M}$.

Finally, following is a MATLAB script generating the number of GerA and GerB* on a population of spores:

```
N_spore = 10000;                % number of spores
N_GerA = zeros(1,N_spore);      % number of GerA GR on each spore
N_GerB = zeros(1,N_spore);      % number of GerB* GR on each spore
a = 3;                          % gamma distribution parameter: a (2*a must be integer)
b = 1/a;                        % gamma distribution parameter: b
r = .35;                        % correlation between levels of GerA and GerB*
M = b/2 * [1 sqrt(r);  sqrt(r) 1];   % covariance matrix
C = chol(M);                    % Cholesky decomposition
for i = 1:N_spore
   Y = randn(2, 2*a);           % standard normal distribution
   X = C' * Y;
   N_GerA(i) = sum(X(1,:).^2);
   N_GerB(i) = sum(X(2,:).^2);
end
```

BIBLIOGRAPHY

1. Schrödinger, E., *What is life?: the physical aspect of the living cell ; with, Mind and matter ; & Autobiographical sketches*. 1992: Cambridge University Press.
2. Watson, J.D., *Molecular biology of the gene*. 2004: Pearson/Benjamin Cummings.
3. Alberts, B., *Molecular biology of the cell*. 2008: Garland Science.
4. Judson, H.F., *The Eighth Day of Creation: Makers of the Revolution in Biology*. 2004: Cold Spring Harbor Laboratory Press.
5. Cairns, J., G.S. Stent, and J.D. Watson, *Phage and the origins of molecular biology*. 2007: Cold Spring Harbor Laboratory Press.
6. Liu, J., J.R. Faeder, and C.J. Camacho, *Toward a quantitative theory of intrinsically disordered proteins and their function.* Proceedings of the National Academy of Sciences, 2009. **106**(47): p. 19819-19823.
7. Setlow, P., J. Liu, and J.R. Faeder, *Heterogeneity in bacterial spore populations*, in *In Bacterial Spores: Current Research and Applications*, E. Abel-Santos, Editor. 2011, Horizon Scientific Press.
8. Yi, X., et al., *Synergism between Different Germinant Receptors in the Germination of Bacillus subtilis Spores.* J Bacteriol, 2011. **193**(18): p. 4664-4671.
9. Berg, J.M., J.L. Tymoczko, and L. Stryer, *Biochemistry*. 2002, New York: W.H. Freeman.
10. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide Protein Data Bank.* Nat Struct Biol, 2003. **10**(12): p. 980.
11. Branden, C. and J. Tooze, *Introduction to protein structure*. 1999, New York,NY: Garland.
12. Romero, P., et al., *Thousands of proteins likely to have long disordered regions.* Pac Symp Biocomput, 1998: p. 437-448.
13. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* J Mol Biol, 1999. **293**(2): p. 321-331.
14. Dunker, A., et al., *Intrinsic Protein Disorder in Complete Genomes.* Genome Informatics, 2000. **11**: p. 161-171.
15. Kendrew, J.C., et al., *A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis.* Nature, 1958. **181**(4610): p. 662-666.
16. Fischer, E., *Einfluss der Configuration auf die Wirkung der Enzyme.* Berichte der deutschen chemischen Gesellschaft, 1894. **27**(3): p. 2985-2993.
17. Koshland, D.E., *The Key-Lock Theory and the Induced Fit Theory.* Angew. Chem. Int. Ed. Engl., 1995. **33**(23-24): p. 2375-2378.
18. Pauling, L., *Nature of forces between large molecules of biological interest.* Nature, 1948. **161**: p. 707-709.

19. Koshland, D.E., *Application of a Theory of Enzyme Specificity to Protein Synthesis.* Proc Natl Acad Sci U S A, 1958. **44**(2): p. 98-104.
20. Koshland, D.E., *How to get paid for having fun.* Annu Rev Biochem, 1996. **65**: p. 1-13.
21. Sigler, P.B., *Acid blobs and negative noodles.* Nature, 1988. **333**(6170): p. 210-212.
22. Schulz, G.E., *Molecular Mechanism of Biological Recognition*, M. Balaban, Editor. 1979, Elsevier/North-Holland Biomedical Press. p. 79-94.
23. Huber, R. and J. William S. Bennett, *Functional significance of flexibility in proteins.* Pure & Appl. Chem., 1982. **54**: p. 2489-2500.
24. Pontius, B.W., *Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association.* Trends Biochem Sci, 1993. **18**(5): p. 181-186.
25. Kriwacki, R.W., et al., *Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity.* Proc Natl Acad Sci U S A, 1996. **93**(21): p. 11504-11509.
26. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
27. Sickmeier, M., et al., *DisProt: the Database of Disordered Proteins.* Nucl. Acids Res., 2007. **35**(suppl_1): p. D786-793.
28. Ward, J.J., et al., *Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life.* J Mol Biol, 2004. **337**(3): p. 635-645.
29. Dunker, A.K. and Z. Obradovic, *The protein trinity -- linking function and disorder.* Nat Biotech, 2001. **19**(9): p. 805-806.
30. Dunker, A.K., et al., *Intrinsically disordered protein.* Journal of Molecular Graphics and Modelling, 2001. **19**(1): p. 26-59.
31. Le Gall, T., et al., *Intrinsic disorder in the Protein Data Bank.* J Biomol Struct Dyn, 2007. **24(4)**: p. 325-342.
32. Dyson, H.J. and P.E. Wright, *Unfolded Proteins and Protein Folding Studied by NMR.* Chem Rev, 2004. **104**(8): p. 3607-3622.
33. Dyson, H.J. and P.E. Wright, *Coupling of folding and binding for unstructured proteins.* Curr Opin Struct Biol, 2002. **12**(1): p. 54-60.
34. Rose, G.D., *Unfolded proteins*. 2002, Amsterdam; Boston: Academic Press.
35. Tompa, P., *Structure and function of intrinsically disordered proteins*. 2010, London: Chapman & Hall.
36. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept.* Annu. Rev. Biophys., 2008. **37**(1): p. 215-.
37. Romero, P., et al. *Identifying Disordered Regions in Proteins from Amino Acid Sequence*. in *Proc. IEEE International Conference on Neural Networks, Huston, TX, June 1997*. 1997.
38. Garner, E., et al., *Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization.* Genome Inform Ser Workshop Genome Inform, 1998. **9**: p. 201-213.
39. He, B., et al., *Predicting intrinsic disorder in proteins: an overview.* Cell Res, 2009. **19**(8): p. 929-949.
40. Ferron, F., et al., *A practical overview of protein disorder prediction methods.* Proteins: Structure, Function, and Bioinformatics, 2006. **65**(1): p. 1-14.

41. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder.* BMC Bioinformatics, 2006. **7**(1): p. 208.
42. Romero, P., et al., *Sequence complexity of disordered protein.* Proteins: Structure, Function, and Genetics, 2001. **42**(1): p. 38-48.
43. Vucetic, S., et al., *Flavors of protein disorder.* Proteins: Structure, Function, and Genetics, 2003. **52**(4): p. 573-584.
44. Vapnik, V.N., *Statistical learning theory*. 1998, New York: Wiley.
45. Prilusky, J., et al., *FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded.* Bioinformatics, 2005. **21**(16): p. 3435-3438.
46. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins: Structure, Function, and Genetics, 2000. **41**(3): p. 415-427.
47. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* Journal of Molecular Biology, 1982. **157**(1): p. 105-132.
48. Linding, R., et al., *Protein Disorder Prediction: Implications for Structural Proteomics.* Structure, 2003. **11**(11): p. 1453-1459.
49. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-2637.
50. Woolfson, M.M., *An introduction to X-ray crystallography*. 1997: Cambridge University Press.
51. Obradovic, Z., et al., *Exploiting heterogeneous sequence properties improves prediction of protein disorder.* Proteins: Structure, Function, and Bioinformatics, 2005. **61**(S7): p. 176-182.
52. Jin, Y. and R.L. Dunbrack Jr., *Assessment of disorder predictions in CASP6.* Proteins: Structure, Function, and Bioinformatics, 2005. **61**(S7): p. 167-175.
53. Bordoli, L., F. Kiefer, and T. Schwede, *Assessment of disorder predictions in CASP7.* Proteins: Structure, Function, and Bioinformatics, 2007. **69**(S8): p. 129-136.
54. Haynes, C., et al., *Intrinsic Disorder Is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes.* PLoS Comput Biol, 2006. **2**(8): p. e100-.
55. Xie, H., et al., *Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions.* J Proteome Res, 2007. **6**(5): p. 1882-1898.
56. Romero, P.R., et al., *Alternative Splicing in Concert with Protein Intrinsic Disorder Enables Increased Functional Diversity in Multicellular Organisms.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8390-8395.
57. Iakoucheva, L.M., et al., *Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins.* J Mol Biol, 2002. **323**(3): p. 573-584.
58. UniProtConsortium, *The Universal Protein Resource (UniProt).* Nucleic Acids Research, 2008. **36**(suppl 1): p. D190-D195.
59. Hong, E.L., et al., *Gene Ontology annotations at SGD: new data sources and annotation methods.* Nucl. Acids Res., 2008. **36**(suppl_1): p. D577-581.
60. Keseler, I.M., et al., *EcoCyc: A comprehensive view of Escherichia coli biology.* Nucl. Acids Res., 2009. **37**: p. D464-70.
61. Liang, C., et al., *Gramene: a growing plant comparative genomics resource.* Nucl. Acids Res., 2008. **36**(suppl_1): p. D947-953.

62. Drysdale, R. and FlyBaseConsortium, *FlyBase: A Database for the Drosophila Research Community*, in *Drosophila: Methods and Protocols --- Methods in Molecular Biology*, C. Dahmann, Editor. 2008, Humana Press. p. 45-59.
63. Bieri, T., et al., *WormBase: new content and better access.* Nucleic Acids Research, 2007. **35**(suppl 1): p. D506-D510.
64. Fey, P., et al., *dictyBase--a Dictyostelium bioinformatics resource update.* Nucl. Acids Res., 2009. **37**(suppl_1): p. D515-519.
65. Aslett, M. and V. Wood, *Gene Ontology annotation status of the fission yeast genome: preliminary coverage approaches 100%*. Yeast, 2006. **23**(13): p. 913-919.
66. Haft, D.H., J.D. Selengut, and O. White, *The TIGRFAMs database of protein families.* Nucleic Acids Research, 2003. **31**(1): p. 371-373.
67. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nat Genet, 2000. **25**(1): p. 25-29.
68. DeGroot, M.H. and M.J. Schervish, *Probability and statistics*. 2002, Boston: Addison-Wesley.
69. Zwanzig, R., *Two-state models of protein folding kinetics.* Proc Natl Acad Sci U S A, 1997. **94**(1): p. 148-150.
70. Tsai, C.-J., et al., *Structured disorder and conformational selection.* Proteins: Structure, Function, and Genetics, 2001. **44**(4): p. 418-427.
71. Spolar, R. and M. Record, Jr, *Coupling of local folding to site-specific binding of proteins to DNA.* Science, 1994. **263**(5148): p. 777-784.
72. Rajamani, D., et al., *Anchor residues in protein-protein interactions.* Proc Natl Acad Sci U S A, 2004. **101**(31): p. 11287-11292.
73. Camacho, C.J., Y. Katsumata, and D.P. Ascherman, *Structural and Thermodynamic Approach to Peptide Immunogenicity.* PLoS Comput Biol, 2008. **4(11)**(11): p. e1000231-.
74. Dunker, A., et al., *Protein disorder and the evolution of molecular recognition: theory, predictions and observations.* Pac Symp Biocomput, 1998: p. 473-484.
75. Yang, L.-W. and I. Bahar, *Coupling between Catalytic Site and Collective Dynamics: A Requirement for Mechanochemical Activity of Enzymes.* Structure, 2005. **13**(6): p. 893-904.
76. Shoichet, B.K., et al., *A relationship between protein stability and protein function.* Proc Natl Acad Sci U S A, 1995. **92**(2): p. 452-456.
77. Ladbury, J.E. and S. Arold, *Searching for specificity in SH domains.* Chemistry & Biology, 2000. **7**(1): p. R3-R8.
78. Wang, R., et al., *The PDBbind Database: Methodologies and Updates.* J Med Chem, 2005. **48**(12): p. 4111-4119.
79. Errington, J., *Regulation of endospore formation in Bacillus subtilis.* Nat Rev Micro, 2003. **1**(2): p. 117-126.
80. Setlow, P., *Spore germination.* Curr Opin Microbiol, 2003. **6**(6): p. 550-556.
81. Setlow, P., *I will survive: DNA protection in bacterial spores.* Trends Microbiol, 2007. **15**(4): p. 172-180.
82. Kawata, T., T. Inoue, and A. Takagi, *Electron microscopy of spore formation germination in Bacillus subtilis.* Japan. J. Microb.,Vol. 7, No. 1, 1963, 1963. **7**: p. 23-41.
83. Hoeniger, J.F.M. and C.L. Headley, *Cytology of Spore Germination in Clostridium pectinovorum.* J. Bacteriol., 1968. **96**(5): p. 1835-1847.

84. Vary, J.C. and H.O. Halvorson, *Kinetics of Germination of Bacillus Spores.* J. Bacteriol., 1965. **89**(5): p. 1340-1347.
85. Ghosh, S. and P. Setlow, *Isolation and Characterization of Superdormant Spores of Bacillus Species.* J. Bacteriol., 2009. **191**(6): p. 1787-1797.
86. Ghosh, S., et al., *Superdormant spores of Bacillus species have elevated wet heat resistance and temperature requirements for heat activation.* J. Bacteriol., 2009. **191**: p. 5584-91.
87. Ghosh, S. and P. Setlow, *The preparation, germination properties and stability of superdormant spores of Bacillus cereus.* Journal of Applied Microbiology, 2010. **108**(2): p. 582-590.
88. Chen, D., S.-s. Huang, and Y.-q. Li, *Real-Time Detection of Kinetic Germination and Heterogeneity of Single Bacillus Spores by Laser Tweezers Raman Spectroscopy.* Anal Chem, 2006. **78**(19): p. 6936-6941.
89. Peng, L., et al., *Elastic and Inelastic Light Scattering from Single Bacterial Spores in an Optical Trap Allows the Monitoring of Spore Germination Dynamics.* Anal Chem, 2009. **81**(10): p. 4035-4042.
90. Kong, L., et al., *Characterization of bacterial spore germination using integrated phase contrast microscopy, Raman spectroscopy, and optical tweezers.* Anal Chem, 2010. **82**(9): p. 3840-3847.
91. Kong, L., et al., *Characterization of bacterial spore germination using phase-contrast and fluorescence microscopy, Raman spectroscopy and optical tweezers.* Nat Protoc, 2011. **6**(5): p. 625-639.
92. Huang, S.-s., et al., *Levels of Ca2+-Dipicolinic Acid in Individual Bacillus Spores Determined Using Microfluidic Raman Tweezers.* J. Bacteriol., 2007. **189**(13): p. 4681-4687.
93. Halmann, M. and A. Keynan, *Stages in germination of spores of Bacillus Licheniformis.* J. Bacteriol., 1962. **84**(6): p. 1187-1193.
94. McCormick, N.G., *Kinetics of Spore Germination.* J. Bacteriol., 1965. **89**(5): p. 1180-1185.
95. Levinson, H.S. and M.T. Hyatt, *Effects of temperature on activation, germination, and outgrowth of Bacillus megaterium spores.* J Bacteriol, 1970. **101**(1): p. 58-64.
96. Stewart, G.S., et al., *Commitment of bacterial spores to germinate. A measure of the trigger reaction.* Biochem J., 1981. **198(1)**: p. 101–-106.
97. Yung, P.T., et al., *Quantification of viable endospores from a Greenland ice core.* FEMS Microbiol Ecol, 2007. **59**(2): p. 300-306.
98. Yang, W.-W. and A. Ponce, *Rapid endospore viability assay of Clostridium sporogenes spores.* International Journal of Food Microbiology, 2009. **133**(3): p. 213-216.
99. Yi, X. and P. Setlow, *Studies of the commitment step in the germination of spores of bacillus species.* J Bacteriol, 2010. **192**(13): p. 3424-3433.
100. Zhang, P., et al., *Factors affecting variability in time between addition of nutrient germinants and rapid dipicolinic acid release during germination of spores of Bacillus species.* J Bacteriol, 2010. **192**(14): p. 3608-3619.
101. Swerdlow, B.M., B. Setlow, and P. Setlow, *Levels of H+ and other monovalent cations in dormant and germinating spores of Bacillus megaterium.* J. Bacteriol., 1981. **148**(1): p. 20-29.

102. Kong, L., et al., *Monitoring the Kinetics of Uptake of a Nucleic Acid Dye during the Germination of Single Spores of Bacillus Species.* Anal Chem, 2010. **82**: p. 8717-24.
103. Woese, C.R., J.C. Vary, and H.O. Halvorson, *A kinetic model for bacterial spore germination.* Proc Natl Acad Sci U S A, 1968. **59**(3): p. 869-875.
104. Moir, A. and D.A. Smith, *The Genetics of Bacterial Spore Germination.* Annu. Rev. Microbiol., 1990. **44**(1): p. 531-.
105. Cabrera-Martinez, R.-M., et al., *Effects of Overexpression of Nutrient Receptors on Germination of Spores of Bacillus subtilis.* J. Bacteriol., 2003. **185**(8): p. 2457-2464.
106. Moir, A., B.M. Corfe, and J. Behravan, *Spore germination.* Cell Mol Life Sci., 2002. **59(3)**: p. 403-409.
107. Paidhungat, M. and P. Setlow, *Isolation and Characterization of Mutations in Bacillus subtilis That Allow Spore Germination in the Novel Germinant D-Alanine.* J. Bacteriol., 1999. **181**(11): p. 3341-3350.
108. Atluri, S., et al., *Cooperativity Between Different Nutrient Receptors in Germination of Spores of Bacillus subtilis and Reduction of This Cooperativity by Alterations in the GerB Receptor.* J. Bacteriol., 2006. **188**(1): p. 28-36.
109. Hill, A.V., *The Combinations of Haemoglobin with Oxygen and with Carbon Monoxide. I.* Biochem J, 1913. **7**(5): p. 471-480.
110. Koshland, D.E., G. Némethy, and D. Filmer, *Comparison of experimental binding data and theoretical models in proteins containing subunits.* Biochemistry, 1966. **5**(1): p. 365-385.
111. Monod, J., J. Wyman, and J.P. Changeux, *On the nature of allosteric transitions: a plausible model.* J Mol Biol., 1965. **12**: p. 88-118.
112. Weiss, J., *The Hill equation revisited: uses and misuses.* FASEB J., 1997. **11**(11): p. 835-841.
113. Taniguchi, Y., et al., *Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells.* Science, 2010. **329**(5991): p. 533-538.
114. Friedman, N., L. Cai, and X.S. Xie, *Linking stochastic dynamics to population distribution: an analytical framework of gene expression.* Phys Rev Lett, 2006. **97**(16): p. 168302.
115. Shahrezaei, V. and P.S. Swain, *Analytical distributions for stochastic gene expression.* Proc Natl Acad Sci U S A, 2008. **105**(45): p. 17256-17261.
116. Cai, L., N. Friedman, and X.S. Xie, *Stochastic protein expression in individual cells at the single molecule level.* Nature, 2006. **440**(7082): p. 358-362.
117. Paidhungat, M. and P. Setlow, *Localization of a Germinant Receptor Protein (GerBA) to the Inner Membrane of Bacillus subtilis Spores.* J. Bacteriol., 2001. **183**(13): p. 3982-3990.
118. Griffiths, K.K., et al., *Germination proteins in the inner membrane of dormant Bacillus subtilis spores colocalize in a discrete cluster.* Molecular Microbiology, 2011. **81**(4): p. 1061-1077.
119. Wang, S.T., et al., *The forespore line of gene expression in Bacillus subtilis.* J Mol Biol, 2006. **358**(1): p. 16-37.
120. Hazelbauer, G.L., J.J. Falke, and J.S. Parkinson, *Bacterial chemoreceptors: high-performance signaling in networked arrays.* Trends in Biochemical Sciences, 2008. **33**(1): p. 9-19.

121. Briegel, A., et al., *Universal architecture of bacterial chemoreceptor arrays.* Proceedings of the National Academy of Sciences, 2009. **106**(40): p. 17181-17186.
122. Greenfield, D., et al., *Self-Organization of the Escherichia coli Chemotaxis Network Imaged with Super-Resolution Light Microscopy.* PLoS Biol, 2009. **7**(6): p. e1000137.
123. Bai, F., et al., *Conformational Spread as a Mechanism for Cooperativity in the Bacterial Flagellar Switch.* Science, 2010. **327**(5966): p. 685-689.
124. Battogtokh, D., et al., *An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa.* Proceedings of the National Academy of Sciences, 2002. **99**(26): p. 16904-16909.
125. Klinke, D., *An empirical Bayesian approach for model-based inference of cellular signaling networks.* BMC Bioinformatics, 2009. **10**(1): p. 371.
126. Metropolis, N., et al., *Equation of State Calculations by Fast Computing Machines.* The Journal of Chemical Physics, 1953. **21**: p. 1087-1087.
127. Hastings, W.K., *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika, 1970. **57**(1): p. 97-109.
128. Pelczar, P.L., et al., *Role of GerD in Germination of Bacillus subtilis Spores.* J. Bacteriol., 2007. **189**(3): p. 1090-1098.
129. Paredes-Sabja, D., P. Setlow, and M.R. Sarker, *Germination of spores of Bacillales and Clostridiales species: mechanisms and proteins involved.* Trends Microbiol, 2011. **19**(2): p. 85-94.
130. Wang, G., et al., *Germination of individual Bacillus subtilis spores with alterations in the GerD and SpoVA proteins important in spore germination.* J Bacteriol, 2011. **193**: p. 2301-11.
131. Shoemaker, B.A., J.J. Portman, and P.G. Wolynes, *Speeding molecular recognition by using the folding funnel: The fly-casting mechanism.* Proc Natl Acad Sci U S A, 2000. **97**(16): p. 8868-8873.
132. Krishnaiah, P.R. and M.M. Rao, *Remarks on a Multivariate Gamma Distribution.* The American Mathematical Monthly, 1961. **68**(4): p. 342-346.