Title
Application of Information Theory to Modeling Exploration and Detecting Protein Coevolution

Author
Little, Daniel Ying-Jeh

Publication Date
2013

Peer reviewed|Thesis/dissertation

Application of Information Theory to Modeling Exploration and Detecting Protein Coevolution

by

Daniel Ying-Jeh Little

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Friedrich T. Sommer, Co-chair
Professor Yang Dan, Co-chair
Professor Daniel E. Feldman
Professor Bruno A. Olshausen

Spring 2013

Application of Information Theory to Modeling Exploration and Detecting Protein Coevolution

Abstract

Application of Information Theory to Modeling Exploration and Detecting Protein Coevolution

by

Daniel Ying-Jeh Little

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Friedrich T. Sommer, Co-chair

Professor Yang Dan, Co-chair

In this thesis I introduce novel applications of information theory to two fundamental problems: the modelling of learning-driven exploration and the identification of coevolving protein residues. While sharing a common approach in the use of information theoretic constructs, they each represent significant contributions to their respective fields.

Discovering the structure underlying observed data is a recurring problem in machine learning with important applications in neuroscience. It is also a primary function of the brain. When data can be actively collected in the context of a closed action-perception loop, behavior becomes a critical determinant of learning efficiency. Psychologists studying exploration and curiosity in humans and animals have long argued that learning itself is a *p*rimary motivator of behavior. However, the theoretical basis of learning-driven behavior is not well understood. Previous computational studies of behavior have largely focused on the control problem of maximizing acquisition of rewards and have treated learning the structure of data as a *s*econdary objective. Here, I study exploration in the absence of external reward feedback. Instead, I take the quality of an agent's learned internal model to be the primary objective. In a simple probabilistic framework, I derive a Bayesian estimate for the amount of information about the environment an agent can expect to receive by taking an action, a measure I term the predicted information gain (PIG). I develop exploration strategies that approximately maximize PIG. One strategy based on value-iteration consistently learns faster, across a diverse range of environments, than previously developed reward-free exploration strategies. Psychologists believe the evolutionary advantage of learning-driven exploration lies in the generalized utility of an accurate internal model. Consistent with this hypothesis, I demonstrate that agents that learn more efficiently during exploration are later better able to accomplish a range of goal-directed tasks. I will conclude by discussing how our work elucidates the explorative behaviors of animals and humans, its relationship to other computational models of behavior, and its potential application to experimental design, such as in closed-loop neurophysiology studies.

The structure and function of a protein is dependent on coordinated interactions between its residues. The selective pressures associated with a mutation at one site should therefore depend on the amino acid identity of interacting sites. Mutual information has previously been applied to mul-

tiple sequence alignments as a means of detecting coevolutionary interactions. Here, I introduce a refinement of the mutual information method that: 1) removes a significant, non-coevolutionary bias and 2) accounts for heteroscedasticity. Using a large, non-overlapping database of protein alignments, I demonstrate that predicted coevolving residue-pairs tend to lie in close physical proximity. I introduce coevolution potentials as a novel measure of the propensity for the 20 amino acids to pair amongst predicted coevolutionary interactions. Ionic, hydrogen, and disulfide bond-forming pairs exhibited the highest potentials. Finally, I demonstrate that pairs of catalytic residues have a significantly increased likelihood to be identified as coevolving. These correlations to distinct protein features verify the accuracy of our algorithm and are consistent with a model of coevolution in which selective pressures towards preserving residue interactions act to shape the mutational landscape of a protein by restricting the set of admissible neutral mutations.

To my grandparents for providing the kindle and my parents for providing the spark.

Contents

Acknowledgments

First and foremost I wish to thank my graduate adviser Friedrich Sommer for his invaluable advice, constant support, and reassuring enthusiasm. When I first approached Fritz about the possibility of transitioning into theoretical neuroscience, my formal education in computational methods amounted to a computer programming course in high school. Nevertheless, Fritz encouraged me to follow my interests and welcomed me into his lab. As a teacher, Fritz has introduced me to such an array of subjects in theoretical neuroscience that I know I will never be for want of an interesting problem to tackle. As my curiosity carried me from one question to another, he was always there to explain the related subjects and offer the insights of experience. As a manager, he found the perfect balance between offering me the freedom to explore my interests and develop my own solutions and pressuring me to set and meet definite goals. And as an adviser, he has always been available whether I simply needed a one word answer or I sought an engaging discussion. Even when on the other side of the globe, he was only a Skype call away. My experiences in Fritz's lab have shaped me as a scientist, have filled my tool kit with invaluable skills, and have given me an interesting line of research that I know I will continue to pursue throughout my career.

I am uniquely privileged in having had not one but two great advisers that have shaped my graduate career and allowed me to contribute across distinct fields. It was in Lu Chen's lab that I truly began to learn about the brain and how insights at all levels from the most abstract computational questions to the most minute molecular mechanisms can inform each other. Lu is fearless in tackling any scientific question that piques her interest. Instead of asking, "What questions can I answer with the techniques and skills already present in my lab?" she will ask, "What new techniques or skills do we need to obtain in order to answer the questions we are interested in?" When we wanted to study the interaction between residues of the PDZ domain we thought perhaps an evolutionary perspective could offer some insights. At the time, however, the methods for detecting such evolutionary interaction were not yielding good results. Lu therefore encouraged me to formulate my own approach and to reanalyze the question of coevolution. This culminated in the development of a novel algorithm for identifying coevolving residues. As is often the case in science, in pursuit of a solution to one question we found ourselves facing numerous new questions further and further away from where we had started. Still Lu never hesitated in supporting my research. I hope that I can approach my future research with the same fearlessness as her.

I feel very fortunate to have had the opportunity to work in the Redwood Center for Theoretical Neuroscience. The heart of the Redwood is its members. It attracts individuals willing to ask the big questions and experts capable of diving deep into the details. No subject will fail to spawn an interesting and often heated conversation amongst its members. Both as a student and as a researcher it has been the perfect enriched environment to learn and explore in and I thank everyone there.

I am greatly indebted to Prof. Jocelyn Malamy at the University of Chicago for introducing me to the wonder of scientific experimentation. It was in her lab that realized that I wanted to be a scientist.

I wish to thank my thesis committee Bruno Olshausen, Yang Dan, and Dan Feldman for their support and feedback.

To my dear friends David Melis, Iris Howlett, Michelle Stephenson, Cecil Devers, Adrienne Maxwell, Brett Schofield, and Robert Ell, all of whom have provided me a family away from home, weathered my analysis paralysis, with whom I have gyred and gimbled and shared numerous adventures, I give all of you my love and thanks. A special thanks goes out to Susanna Porter, my oldest and dearest friends. I'll see you soon!

Lastly, I wish to thank my family, the Littles and the Lius and the many others who have been bold enough to join us, for your unwavering love and support. I particularly want to thank my siblings Kim Little-Weinert and David Little for always having my back through the best and worst of times. I know you two will always be there for me, and me for you. To my new nephew Logan Scott Wienert, born April 5, 2013, you were a source of joy and happiness during the stress and exhaustion of preparing this thesis. And finally I have reserved my deepest thanks for my grandparents, Jack and Ellen Little and Tzen-Shen and Chao-Wei Liu, and my parents, Mike and Becky Little. In the great debate over nature versus nurture you have proven that it doesn't have to be one or the other.

Part I

An Information Theoretic Model of Exploration

Chapter 1

Introduction

As scientists, we are intimately familiar with the human desire to explore and learn, to pose questions about the world and seek answers by engaging with it. While our knowledge can be esoteric and our experiments complex, the underlying motivation is a universal human trait. Exploration, along with the curiosity that motivates, is crucial in the cognitive development of children [51, 96, 102] and in the maintenace of cogitive systems for problem solving, thinking and creativity throughout life [67]. It is predictive of academic success [49, 107, 111], positive social relationships [58, 78, 138], and personal growth and well-being [30, 57, 59]. However, despite the important role curiosity-driven behaviors play in cognitive development and positive psychology, they have largely been overlooked by computational models of behavior. In particular, the success of reinforcement learning in describing reward-oriented behaviors has led them to dominate the computational perspective on exploration. The reward-focused exploration of reinforcement learning stands in sharp contrast to the learning-driven exploration esteemed by psychologists. This thesis aims to provide a computational framework for studying learning-driven exploration. In particular, I develop information theoretic and computational principles for guiding explorative behaviors towards efficiently learning about the world. By omitting the rewards structure of classic reinforcement learning and instead looking directly at how behavior effects learning I hope to provide a new perspective of a once underrepresented behavioral motive.

1.1 Outline of thesis

Part I of this thesis consists of five chapters beyond this Introduction. Chapter 2 lays out the mathematical framework for studying explorative behaviors. In particular, it defines the dynamic structure of the worlds to be explored and offers an information theoretic measure of learning. In Chapter 3, I develop PIG(VI) (Value Iterated maximization of Predicted Information Gain) as an information theoretic model of exploration. I demonstrate its efficacy in directing behaviors towards quickly learning about the world and compare it to an array of competitors drawn from the field of reinforcement learning and inspired by findings from the field of Psychology. In Chapter 4, I consider the long-term adaptive benefits of efficient exploration and demonstrate that accurate internal models learned by PIG(VI) offer generalized utility not achieved by a state-of-the-art re-

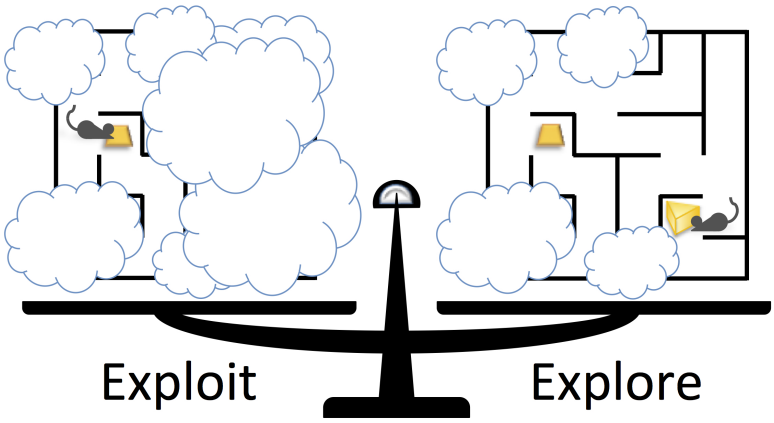

Figure 1.1: The reinforcement learning perspective. When facing an uncertain environment (obscuring clouds), exploitation and exploration must be balanced in order to maximize the acquisition of rewards.

inforcement learning algorithm. In Chapter 5, I begin to relax some of the assumptions of earlier chapters and consider the impact of uncertain priors on exploration. I also demonstrate that the discrete framework employed in this thesis can be utilized to direct exploration of a continuous system. Finally, in Chapter 6, I offer a discussion of the results and of the relationship between my work and previous work in the fields of Psychology and computational modeling of behavior.

In the remainder of this chapter, I will introduce and contrast the reinforcement learning and Psychology perspectives on exploration. In particular, I will contrast their responses to the basic question, "Why do we explore?" My goal is not to draw a fine line between the approaches taken in this thesis and the vast literature on reinforcement learning. Indeed, I will utilize some of the same techniques employed by reinforcement learning algorithms. Instead, my hope is to encourage a new perspective in computational modeling that considers separately the role of learning as a primary motivator of exploration and the question of reward-driven behaviors. Finally, I will conclude this chapter by introducing a second motivation of this thesis framed in terms of closing the action-perception loop. Work in this half of the thesis has been published, in part, in *Frontiers in Neural Circuits* [69].

1.2 The reinforcement learning view of exploration

Reinforcement learning focuses on the control problem of learning a behavioral policy that maximizes the acquisition of rewards. In their seminal book on the subject, Sutton and Barto described it as, "learning what to do–how to map situations to actions–so as to maximize a numerical reward signal" [125]. It was developed, in part, in response to the supervised learning algorithms that dominated the field of machine learning in the early 1980's. In supervised learning, a knowledgeable external teacher informs the learner of the correct action. In many learning situations, however, teachers that can communicate the desired behaviors may not be available and the learner must

learn from its own experiences alone. Reinforcement learning sought to address how an agent could learn to accomplish a task (the acquisition of rewards) when a teacher is not available.

Since its introduction (or reintroduction depending on your historical perspective) in the early 1980's, reinforcement learning algorithms have been successfully applied to an array of problems in robotics [65, 84, 122, 126], operations research [3, 101], and perhaps most famously in gaming [12, 110, 127]. These successes highlight the great progress in machine learning that has been engendered by reinforcement learning. At the same time, recent behavioral and neuroscience findings have significantly expanded the attention on reinforcement learning by introducing it to a wider scientific audience. Reinforcement learning models have been able to replicate behavioral findings in Pavlovian reinforcement [124], social conformity behaviors [64], and decision-making [24]. Furthermore, potential neural correlates of various reinforcement learning signals have been identified in the basal ganglia and cerebral cortex from fMRI and electrophysiology studies [9, 29, 31, 43, 108]. These findings support the hypothesis that the brain may be using something similar to reinforcement learning during goal-directed behaviors.

The modeling of exploratory behavior in reinforcement learning has focused nearly exclusively on the role of exploration in the acquisition of rewards [54, 61, 74, 125, 128]. This function of exploration is typically presented in terms of an exploration-exploitation trade-off. An agent that wishes to optimally gather rewards from an unknown environment must decide whether to focus on exploiting the reward opportunities it has already found or to explore the unknown in search of better reward opportunities (Fig. 1.1). Sutton and Barto introduce this as a distinguishing problem of reinforcement learning:

> One of the challenges that arise in reinforcement learning and not in other kinds of learning is the trade-off between exploration and exploitation. To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions, it has to try actions that it has not selected before. The agent has to *exploit* what it already knows in order to obtain reward, but it also has to *explore* in order to make better action selections in the future. The dilemma is that neither exploration nor exploitation can be pursued exclusively without failing at the task. [125]

Thus, to a reinforcement learner the purpose of exploration is to find rewards. While we can recognize the utility of such reward-driven exploration, this explanation seems insufficient in describing our own personal motivations for exploring. Both as scientist in particular and as human beings in general, we experience the drive to explore as coming from within, as being satisfied not by the accumulation of rewards, but by the very process of interacting with and learning about our world. We are thus faced with contrasting explanations for exploration. On the one hand we understand that exploration is useful in finding rewards but on the other we feel this to be insufficient at explaining our drive to explore. I propose that a reconciliation of these two causes of exploration can be found in the Psychology perspective.

1.3 Proximate and ultimate causes of behavior

If you ask a psychologist, "Why do we explore?" they are likely to respond, "It depends." In fact if you ask them why do we do any behavior, they will likely respond in the same fashion. This is because psychologists have come to distinguish between different causes of behavior. Mayr proposed two categories of explanations for the biology of behaviors: proximate and ultimate [77]. Proximate causes refer to the factors, whether internal or external, that directly control an individual's behaviors. Mayr gives the example of birds migrating south during the winter as an intrinsic physiological response to decreasing day length. He contrasts the proximate cause to an ultimate cause that led to the behavior being incorporated into a species over many thousands of generations. The ultimate cause for why birds migrate south during the winter is because they would starve from loss of food if they didn't. The proximate cause could be described as the behavioral cause and the ultimate as the evolutionary one. Mayr describes it as thus:

> Still another way to express these differences would be to say that proximate causes govern the responses of the individual (and his organs) to immediate factors... while ultimate causes are responsible for the evolution of the particular DNA code of information with which every individual of every species is endowed... [T]he biologist knows that many heated arguments about the "cause" of a certain biological phenomenon could have been avoided if the two opponents had realized that one of them was concerned with the proximate and the other the ultimate cause. [77]

1.4 The proximate cause of exploration

Psychologists have long struggled with identifying the proximate causes of exploration. In supporter of the reward-driven perspective of reinforcement learning, a few early psychologists postulated that explorative behaviors were a learned response to conditioned stimuli [70, 85]. This perspective however quickly grew out of favor. As D. E. Berlyne, a pioneer in the psychology of exploration explained:

> As knowledge accumulated about the conditions that govern exploratory behavior and about how quickly it appears after birth, it seemed less and less likely that this behavior could be derivative of hunger, thirst, sexual appetite, pain, fear of pain, and the like, or that stimuli sought through exploration are welcomed because they have previously accompanied satisfaction of these drives. [16]

Instead, Berlyne suggested that "the most acute motivational problems . . . are those in which the perceptual and intellectual activities are engaged in for their own sake and not simply as aids to handle practical problems" [15].

Psychologists call the internal motivation that promotes explorative behaviors curiosity or interest [70, 99, 118]. Though it has been difficult to define exactly what curiosity is, a consensus has emerged in behavioral psychology that learning represents one of its central components and

a primary drive of exploration [4, 70, 99, 116]. The *Encyclopedia of the Sciences of Learning* offers the following tentative definition:"[C]uriosity is the desire for new information and sensory experiences that motivates exploration of the environment" [1]. While a diverse array of theories of exploration have been developed in the field of Psychology, information, in various guises, has persisted as a central player. Early theories focused on the informational conditions that encouraged exploration. Berlyne identified novelty and complexity as stimulus properties that induced explorative behaviors [16]. Silvia would later add comprehensibility to this list [118]. Piaget postulated that expectation violations were the central instigators of exploration [96]. Later theories, would place information more explicitly at the center of exploration. For example, Lowenstein's Information Gap Theory views curiosity as "arising when attention becomes focused on a gap in one's knowledge" so that the individual "is motivated to obtain the missing information" [71].

1.5 Quantifying information

Thus, the Psychology perspective regards information, not the search for rewards, as the primary drive of exploration. A natural path to incorporate information principles into computational models, and the approach that I take in this thesis, lies in the mathematical foundations of information theory. Interested in the fundamental limits on communication, C.E. Shannon developed information theory to quantify the informational properties of random variables [113]. Many of Shannon's measures, along with subsequent measures developed in the field, may offer explicit mathematical interpretations of some of the theories and central concepts of exploration proposed by psychologist. For example, considering a random variable X with probability distribution p, Shannon defined the entropy of X as:

$$\mathrm{H}(X) = -\sum_{x} p(x) \log_2(p(x))$$

Shannon's entropy quantifies the uncertainty of a random variable by the average number of bits it would take to communicate it, and could be interpreted as an information theoretic analogue to Berlyne's complexity trait [16]. Similarly, self-information, also called surprisal, quantifies the information content of a specific outcome for a random variable [105, 133]:

$$SI(X) = \sum_{x} \log_2(\frac{1}{p(x)})$$

Self-information is highest when an outcome is least expected. It therefor may correspond to the expectation violations of Piaget's incongruity theory of exploration [96]. Finally, the Predicted Information Gain measure which I introduce in this thesis, quantifies the amount of information a learner can expect to obtain from a particular source of data. As such, it would be an appropriate information theoretic analogue of the comprehensibility term introduced by Silvia [118]. While most psychology theories of exploration were developed without an explicit information theoretic model, Lowenstein is unique amongst the for directly mentioning information theory suggesting that, "information theory's entropy coefficient provides a potential measure of the degree of one's information (actually one's ignorance)..." [71].

1.6 The ultimate causes of exploration

If curiosity, and thus information, represents the proximate cause of exploration, what about its ultimate cause? What adaptive benefits might information-driven exploration offer a species. Perhaps the reinforcement learning perspective offers a reasonable answer. Did curiosity evolve because exploration increases the likelihood of finding unknown rewards to exploit? We can easily imagine how this could lead to a survival benefit for curious animals. At the same time, however, recalling the old adage that curiosity killed the cat, we expect such benefits to be partly counterbalanced by the increased chance of stumbling upon unknown predators or natural dangers. While accepting that the search for food, shelter, mating opportunities, etc. is clearly important to fitness, Psychologists believe that the greatest benefit of curiosity-driven exploration lies in the general utility of the gathered information that can be applied adaptively in accordance with varying circumstances and changing needs [55, 97, 98, 103, 104]. The adaptive benefits of exploration are not necessarily immediate; they can be deferred until that time at which the acquired information becomes relevant [92]. The Surplus Resource Theory notes that across those species that provide parental care for their young, the juvenile stages of life are marked by higher levels of exploration and play [21]. It is precisely because the juvenile's survival is disconnected from its actions, its safety and provisioning being supplied by its parents, that it is afforded the opportunity to develop the skills and accumulate the knowledge it will need to survive into adulthood [92].

Thus, the reinforcement learning and Psychology perspectives on explorative behaviors differ both in regards to their proximate causes as well as their ultimate causes. In contrast to the reward-driven exploration of reinforcement learning, little attention has been given to developing computational principles for learning-driven exploration. Furthermore, the reinforcement learning focus on the exploration-exploitation tradeoff has prevented it from sufficiently assessing the generalized utility of information. This thesis aims at filling these gaps.

1.7 Closing the action-perception loop

While the distinction between reward-driven and learning-driven exploration has become central to this thesis, its original motivation lied in a different theoretical question. My exploration of exploration began with asking how closing the action-perception loop would change our models of learning. Machine learning techniques for extracting the structure underlying sensory signals have often focused on passive learning systems that cannot directly affect the sensory input. Closing the action-perception loop offers the learner the opportunity to actively pursue information, that is the opportunity to explore. Learning in closed action-perception loops differs from passive learning both in terms of "what" is being learned as well as "how" it is learned [45]. In particular, in closed action-perception loops:

1. Sensorimotor contingencies must be learned, and

2. Actions must be coordinated to direct the acquisition of data.

Sensorimotor contingencies refer to the causal role actions play on the sensory inputs we receive, such as the way visual scenes change as we shift our gaze or move our head. They must be taken into account to properly attribute changes in sensory signals to their causes. This tight interaction between actions and sensation is reflected in the neuroanatomy where sensory-motor integration has been reported at all levels of the brain [47, 48]. We often take our implicit understanding of sensorimotor contingencies for granted, but in fact they must be learned during the course of development. This is eloquently expressed in the explorative behaviors of young infants (e.g., grasping and manipulating objects during proprioceptive exploration and then bringing them into visual view during intermodal exploration) [86, 89, 106]. Recently, researchers have postulated that the learning of such sensorimotor contigencies is necessary for the emergence of perception [86, 90, 95].

Not only are actions part of "what" we learn during exploration, they are also part of "how" we learn. To discover what is inside an unfamiliar box, a curious child must open it. To learn about the world, scientists perform experiments. Directing the acquisition of data is particularly important for embodied agents whose actuators and sensors are physically confined. Since the most informative data may not always be accessible to a physical sensor, embodiment may constrain an exploring agent and require that it coordinate its actions to retrieve useful data.

In the model of learning-driven exploration I propose here, an agent moving between discrete states in a world has to learn how its actions influence its state transitions. The underlying transition dynamics is governed by a Controllable Markov Chain (CMC). Within this simple framework, various utility functions for guiding exploratory behaviors will be studied, as well as several methods for coordinating actions over time. The different exploratory strategies are compared in their rate of learning and how well they enable agents to perform goal-directed tasks.

Chapter 2

Mathematical framework for embodied active learning

2.1 The Controllable Markov Chain

A *Controllable Markov Chain (CMC)* is a simple discrete-time stochastic control process. It extends the basic Markov chain with the addition of a control variable for switching between different transition distributions in each state [38]. The incorporation of this control variable adds the active component necessary for studying active learning. Formally, a CMC is a 3-tuple $(\mathscr{S}, \mathscr{A}, \Theta)$ where:

- $\mathscr{S}$ is a finite set of *states* (for example, the possible locations of an agent in its world). $N = |\mathscr{S}|$
- $\mathscr{A}$ is a finite set of control values, or *actions*, an agent can choose from. $M = |\mathscr{A}|$
- Θ is a 3-dimensional *CMC kernel* describing the transition probabilities between states for each action. Θ defines the probability an agent moves from an *originating state* s to a *resultant state* s' when it chooses action a as follows:

$$\begin{aligned} &p(s'|a,s;\Theta) = \Theta_{ass'} \\ &\Theta_{as\cdot} \in \Delta_{N-1} \end{aligned} \tag{2.1}$$

Here, Δ_{N-1} denotes the standard $(N-1)$-simplex and is used to constrain Θ to describing legitimate probability distributions:

$$\Delta_{N-1} := \{(x_0, x_1, \cdots, x_{N-1}) \in \mathbb{R}^N | \sum_{i=0}^{N-1} x_i = 1 \text{ and } x_i \geq 0 \; \forall i\}$$

CMCs provide a simple mathematical framework for modeling exploration in embodied action-perception loops. At each time step, an exploring agent is allowed to select any action $a \in \mathscr{A}$ within

its current state. The agent will then transition to a new state with transition probabilities dependent on both on its current state and its selected action. For the scope of this thesis, I will assume the states can be directly observed by the agent, i.e. the system is not hidden. The dynamical structure of a CMC is thus described by the kernel Θ. Accordingly, I take the learning task of the exploring agent to be the formation of an accurate estimate, or *internal model* $\widehat{\Theta}$, of the true CMC kernel that describes its *world* Θ.

Despite its simplicity, the CMC framework captures the two distinguishing features of learning in closed action-perception loops. First, the state transitions are conditioned upon an agent's action choice. A full model of the dynamic structure of the world must account for such sensorimotor contingencies. In other words, the influence of actions on states are a constituent part of "what" is being learned in a CMC. Second, an agent's immediate ability to interact with and observe the world is limited by its current state. This restriction represents an *embodiment* constraint on the agent. Should, in the pursuit of learning, an agent wish to investigate a distant state, it will first have to coordinate its actions to reach that destination. An agent in a CMC, being embodied, can not simply instantaneously relocate itself to any arbitrary state but must act within the limits of its world to reach that state. Through coordination, actions become "how" an agent can compensate for this embodiment constraint.

The primary question asked by this thesis is how action policies can optimize the speed and efficiency of learning in embodied action-perception loops as modeled by CMCs. In endeavoring to answer this question, I will have to address the preliminary problem of inference: Given a set of data, how should an agent construct its estimate $\widehat{\Theta}$. For both questions, it will be necessary to quantify the quality of a learned internal model. In the following section, I show how information theory can be used to provide such a quantification.

2.2 Information-theoretic assessment of learning

Following Pfaffelhuber [94], I define *missing information* $\mathrm{I_M}$ as a measure of the inaccuracy of an agent's internal model. To compute $\mathrm{I_M}$, we must first calculate the Kullback-Leibler (KL) divergence of the internal model from the world for each transition distribution:

$$\mathrm{D_{KL}}(\Theta_{as\cdot} \parallel \widehat{\Theta}_{as\cdot}) := \sum_{s'=1}^{N} \Theta_{ass'} \log_2 \left(\frac{\Theta_{ass'}}{\widehat{\Theta}_{ass'}} \right) \tag{2.2}$$

The KL-divergence is an information theoretic measure of the difference between two distributions. Specifically, Eq. 2.2 gives the expected number of extra bits it would take to communicate observations drawn from the true distribution using an encoding scheme optimized for the estimated distribution [25]. It is large when the two distributions differ greatly and zero when they are identical. Next, missing information is defined as the unweighted sum of the KL-divergences:

$$\mathrm{I_M}(\Theta \parallel \widehat{\Theta}) := \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathrm{D_{KL}}(\Theta_{as\cdot} || \widehat{\Theta}_{as\cdot}) \tag{2.3}$$

Table 2.1: Table of Measures

Name used here, Abbreviation (Equation)	Name used in [Reference]	Mathematical expression
Missing Information, $\mathrm{I_M}$ (2.3)	Missing Information [94]	$\sum_{s,a} \mathrm{D_{KL}}(\Theta_{as\cdot} \| \widehat{\Theta}_{as\cdot})$
Information Gain, $\mathrm{I_G}$ (3.1)		$\mathrm{I_M}(\Theta \parallel \widehat{\Theta}) - \mathrm{I_M}(\Theta \parallel \widehat{\Theta}^{a,s\to s^*})$
Predicted Information Gain, PIG (3.2)	Information Gain [83]	$\sum_{s^*} \widehat{\Theta}_{ass^*} \mathrm{D_{KL}}(\widehat{\Theta}^{a,s\to s^*}_{as\cdot} \parallel \widehat{\Theta}_{as\cdot})$
Posterior Expected Information Gain, PEIG (A.1)	KL-Divergence [123]	$\mathrm{D_{KL}}(\widehat{\Theta}^{current}_{as\cdot} \parallel \widehat{\Theta}^{past}_{as\cdot})$
Predicted Mode Change, PMC (3.6)	Probability Gain [83]	$\sum_{s^*} \widehat{\Theta}_{ass^*} \left[\max_{s'} \widehat{\Theta}^{a,s\to s^*}_{ass'} - \max_{s'} \widehat{\Theta}_{ass'} \right]$
Predicted L_1 Change, PLC (3.7)	Impact [83]	$\sum_{s^*} \widehat{\Theta}_{ass^*} \left[\frac{1}{N} \sum_{s'} \left| \widehat{\Theta}^{a,s\to s^*}_{ass'} - \widehat{\Theta}_{ass'} \right| \right]$

I will use missing information to assess learning under different explorative strategies. Steeper decreases in missing information over time represent faster learning and thus more efficient exploration. Since missing information utilizes an unweighted sum, it represents the most parsimonious valuation of information. Specifically, all information is treated as equally important when learning about the world. For easy reference, Table 2.1 compiles the definitions of several terms introduced in this manuscript and includes missing information.

2.3 Bayesian inference learning

As an agent acts in its world, it observes the resulting state transitions and uses these observations to update its internal model $\widehat{\Theta}$. Taking a Bayesian approach, I assume the agent models its world Θ as a random variable $\mathbf{\Theta}$ with an initial *prior distribution* f over the space of possible CMC structures $\mathscr{W} = \Delta_{N-1}^{NM}$. There is no standard nomenclature for tensor random variables and I will therefore use a bold upright theta $\mathbf{\Theta}$ to denote the random variable and a regular upright theta Θ to denote an arbitrary realization of this random variable. As previously introduced, the italicized theta Θ denotes the true value of the transition probabilities which are known to me as the experimenter but not to the exploring agents. Thus, $f(\Theta)$ describes the exploring agent's initial belief that Θ accurately describes its world, i.e. that $\Theta = \Theta$. By Bayes' theorem, an agent can calculate a posterior belief on the structure of its world from its prior and any data $\vec{\mathrm{d}}$ it has collected:

$$f(\Theta|\vec{\mathrm{d}}) = \frac{p(\vec{\mathrm{d}}|\Theta) f(\Theta)}{p(\vec{\mathrm{d}})} \tag{2.4}$$

Bayes' theorem decomposes the posterior distribution of the CMC kernel into the likelihood function of the data, $p(\vec{\mathrm{d}}|\Theta)$, and the prior, $f(\Theta)$. The normalization factor is calculated by integrating the numerator over $\Theta \in \mathscr{W}$:

$$p(\vec{\mathrm{d}}) = \int_{\mathscr{W}} p(\vec{\mathrm{d}}|\Theta) f(\Theta) d\Theta$$

We can now formulate a Bayesian estimate by directly calculating the posterior belief that one will transition to state s' from state s under action a:

$$\begin{aligned}\widehat{\Theta}_{ass'} := p(s'|a,s,\vec{\mathrm{d}}) &= \int_{\mathscr{W}} p(s',\Theta|a,s,\vec{\mathrm{d}})d\Theta \\ &= \int_{\mathscr{W}} p(s'|a,s;\Theta)f(\Theta|\vec{\mathrm{d}})d\Theta \\ &= \int_{\mathscr{W}} \Theta_{ass'} f(\Theta|\vec{\mathrm{d}})d\Theta = \mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}[\boldsymbol{\Theta}_{ass'}] \end{aligned} \tag{2.5}$$

For discrete priors the above integrals would be replaced with summations. Equation (2.5) demonstrates that the Bayesian estimate is simply the expectation of the random variable given the data. While other estimates are possible for inferring world structure, such as Maximum Likelihood, the Bayesian estimate is often employed to avoid over-fitting [75]. Moreover, as the following theorem demonstrates, the Bayesian estimate is optimal under our minimum missing information objective function introduced in Section 2.2.

Theorem 1. *Consider a CMC random variable $\boldsymbol{\Theta}$ modeling the ground truth environment Θ and drawn from a prior distribution f. Given a history of observations $\vec{\mathrm{d}}$, the expected missing information between $\boldsymbol{\Theta}$ and an agent's internal model Φ is minimized by the Bayesian estimate $\Phi = \widehat{\Theta}$. That is:*

$$\widehat{\Theta} := \mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}[\boldsymbol{\Theta}] = \arg\min_{\Phi} E_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\mathrm{I_M}(\boldsymbol{\Theta} \parallel \Phi)\right]$$

Proof. Minimizing missing information is equivalent to independently minimizing the KL-divergence of each transition kernel.

$$\begin{aligned}&\arg\min_{\boldsymbol{\Phi}_{as\cdot}} \mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\mathrm{D_{KL}}\left(\boldsymbol{\Theta}_{as\cdot} \parallel \boldsymbol{\Phi}_{as\cdot}\right)\right] \\ &= \arg\min_{\boldsymbol{\Phi}_{as\cdot}} \mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\sum_{s'} \boldsymbol{\Theta}_{ass'} \log_2\left(\frac{\boldsymbol{\Theta}_{ass'}}{\boldsymbol{\Phi}_{ass'}}\right)\right] \\ &= \arg\min_{\boldsymbol{\Phi}_{as\cdot}} \mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\sum_{s'} \boldsymbol{\Theta}_{ass'} \log_2 \boldsymbol{\Theta}_{ass'} - \boldsymbol{\Theta}_{ass'} \log_2 \boldsymbol{\Phi}_{ass'}\right] \\ &= \arg\min_{\boldsymbol{\Phi}_{as\cdot}} -\mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\sum_{s'} \boldsymbol{\Theta}_{ass'} \log_2 \boldsymbol{\Phi}_{ass'}\right] \\ &= \arg\min_{\boldsymbol{\Phi}_{as\cdot}} -\sum_{s'} \mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\boldsymbol{\Theta}_{ass'}\right] \log_2 \boldsymbol{\Phi}_{ass'} \\ &= \arg\min_{\boldsymbol{\Phi}_{as\cdot}} \mathrm{H}\left[\mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\boldsymbol{\Theta}_{as\cdot}\right]; \boldsymbol{\Phi}_{as\cdot}\right]\end{aligned}$$

Here H denotes cross-entropy [25]. Finally, by Gibb's inequality [25]:

$$\begin{aligned}&\arg\min_{\Phi_{as\cdot}} \mathrm{H}\left[E_{\Theta|\vec{\mathrm{d}}}[\Theta_{as\cdot}]; \Phi_{as\cdot}\right] \\ &= E_{\Theta|\vec{\mathrm{d}}}[\Theta_{as\cdot}] \\ &= \widehat{\Theta}_{as\cdot}\end{aligned}$$

□

The exact analytical form for the Bayesian estimate will depend on the prior distribution. In the next section I will introduce the three classes of CMCs that will be the primary focus of this thesis. I would like to emphasize that the utility of the Bayesian estimate rests on the accuracy of its prior. For now, I will provide the agents with accurate priors that match the generative process by which I created new worlds for the agents to explore. I will then return to the question of learning with uncertain or inaccurate priors in Chapter 5.

2.4 Three test environments for studying exploration

Over the course of exploration, the data an agent accumulates will depend on both its behavioral strategy as well as the structure of its world. I reasoned that studying diverse environments, i.e. CMCs that differ greatly in structure, would allow me to investigate how world structure effects the relative performance of different exploratory strategies and to identify action policies that produce efficient learning under broad conditions. I therefore developed three classes of CMCs that differ greatly in structure to investigate: Dense Worlds, Mazes, and 1-2-3 Worlds. For each class, random CMCs were generated by drawing the transition distributions from a specific generative distribution. These generative distributions were given to the agents as priors for performing Bayesian inference. I will consider each class of CMC in turn.

Dense Worlds

Dense Worlds correspond to complete directed probability graphs with $N = 10$ states and $M = 4$ actions. They are randomly generated from a uniform distribution over all CMCs. They therefore represent very unstructured worlds. Specifically, each transition distribution is independently drawn from a Dirichlet distribution over the standard $(N-1)$-simplex:

$$f(\Theta_{as\cdot}) = \mathrm{Dir}(\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \cdot \prod_{s'} \Theta_{ass'}{}^{\boldsymbol{\alpha}_{s'}-1}$$

The normalizing constant Z brings the area under the distribution to 1:

$$\begin{aligned}&Z(\boldsymbol{\alpha}) := \int_{\Delta_{N-1}} \prod_{s'} \Theta_{ass'}{}^{\boldsymbol{\alpha}_{s'}-1} d\Theta_{as\cdot} = \frac{\prod_{s'} \Gamma(\boldsymbol{\alpha}_{s'})}{\Gamma(\sum_{s'} \boldsymbol{\alpha}_{s'})} \\ &\text{where } \Gamma(x) := \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t\end{aligned}$$

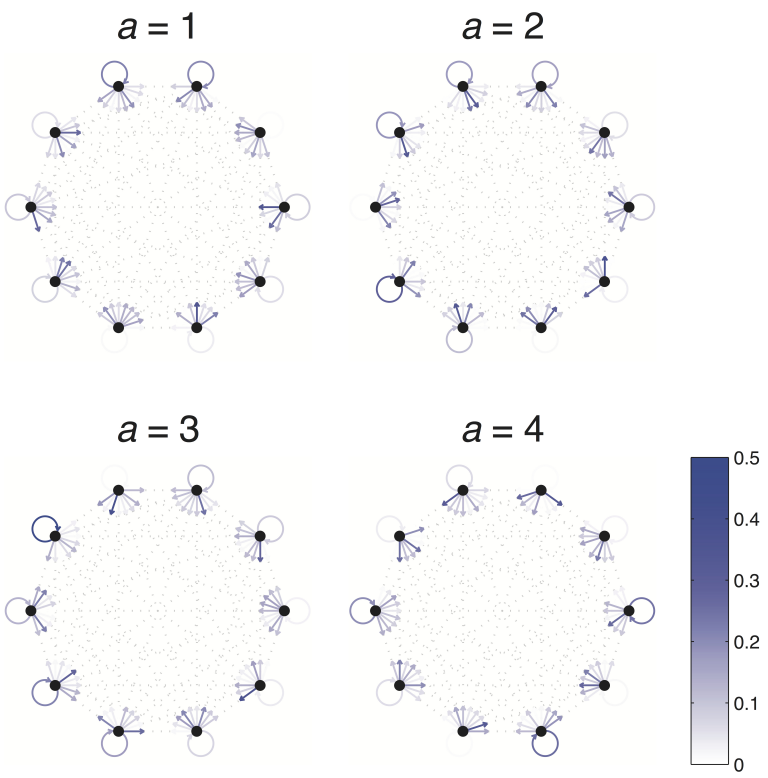

Figure 2.1: Example Dense World. Dense Worlds consist of 4 actions (separately depicted) and 10 states (depicted as nodes of the graphs). The transition probabilities associated with taking a particular action are depicted as arrows pointing from the current state to each of the possible resultant states. Arrow color depicts the likelihood of each transition.

The mean of a Dirichlet distribution takes on a simple form:

$$\int_{\Delta_{N-1}} \Theta_{as\cdot} \frac{\prod_{s'} \Theta_{ass'}^{\alpha_{s'}-1}}{Z(\boldsymbol{\alpha})} d\Theta_{as\cdot} = \frac{\boldsymbol{\alpha}}{\sum_{s'} \alpha'_s}$$

I will assume a symmetric prior setting $\boldsymbol{\alpha}_{s'}$ equal to α for all s'. The vector form of the Dirichlet distribution will nevertheless still be useful in deriving the Bayesian estimate. The parameter α determines how much probability weight is centered at the midpoint of the simplex and is known as the *concentration factor*. For Dense Worlds, I used a concentration factor $\alpha = 1$ which results in a uniform distribution over the simplex. An example Dense World is depicted in Fig. 2.1.

To derive an analytic form for the Bayesian estimate of Dense Worlds, I define the matrix $\boldsymbol{F}$ such that $\boldsymbol{F}_{ass'}$ is a count of the number of times $a, s \to s'$, i.e. a transition from s to s' under action a, has occurred in the data. Since each layer $\widehat{\Theta}_{as\cdot}$ of the CMC kernel is independently distributed,

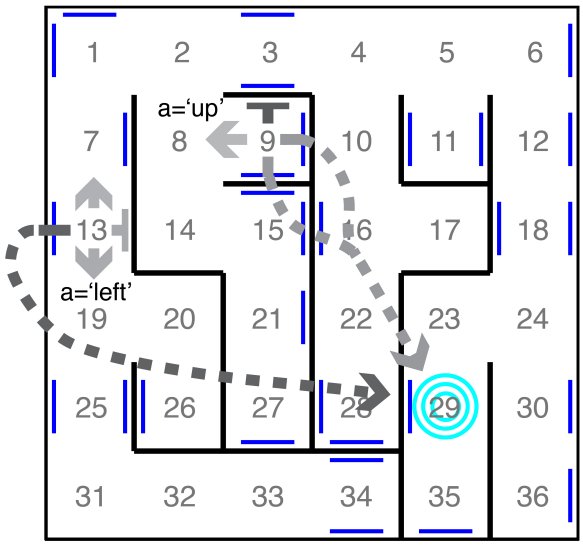

Figure 2.2: Example maze. The 36 states correspond to rooms in a maze. The 4 actions correspond to noisy translations in the cardinal directions. Two transition distributions are depicted, each by a set of 4 arrows emanating from their originating states. Flat-headed arrows represent translations into walls, resulting in staying in the same room. Dashed arrows represent translation into a portal (blue lines) leading to the absorbing state (blue target). The shading of an arrow indicates the probability of the transition (darker color represents higher probability).

its posterior distribution can be computed as follows:

$$f(\Theta|\boldsymbol{F}) = \frac{\prod_{s'} \Theta_{ass'}^{F_{ass'}} \cdot \prod_{s'} \Theta_{ass'}{}^{\alpha-1} / Z(\alpha)}{p(\boldsymbol{F})} = \frac{\prod_{s'} \Theta_{ass'}^{F_{ass'}+\alpha-1}}{Z(\alpha)p(\boldsymbol{F})} = \mathrm{Dir}(\boldsymbol{F}+\alpha)$$

Thus, the posterior distribution is also Dirichlet and the Bayesian estimate $\widehat{\Theta}$ is simply the mean of the distribution:

$$\widehat{\Theta}_{ass'} = \frac{F_{ass'} + \alpha}{\sum_{s^*} F_{ass^*} + \alpha} = \frac{F_{ass'} + 1}{\sum_{s^*} F_{ass^*} + 1} \tag{2.6}$$

In this form, we find that the Bayesian estimate for Dense Worlds is simply the relative frequencies of the observed data with the addition of fictitious counts of size α to each bin. The incorporation of this fictitious observation is referred to as Laplace smoothing and is often performed to avoid over-fitting [75]. The derivation of Laplace smoothing from Bayesian inference over a Dirichlet prior is a well known result [73].

Mazes

In contrast to Dense Worlds, *Mazes* are highly structured and model moving between rooms of a 6-by-6 maze (see Fig. 2.2). The state space in mazes consist of the $N = 36$ rooms in a randomly

generated maze. The $M = 4$ actions correspond to noisy translations in the four cardinal directions. Walking into a wall causes the agent to remain in its current location. 30 transporters are randomly distributed amongst the walls which lead to a randomly chosen absorbing state (concentric rings in Fig. 2.2). While perhaps not typically abundant in mazes, absorbing states, such as at the bottom of a gravity well, are common in real world dynamics. States that are not one step away from the originating state (either directly, through a portal, or against a wall) are assumed to have zero probability of resulting from any action. Transition probabilities for states that are one step away are drawn from a Dirichlet distribution with concentration factor $\alpha = 0.25$, and the highest probability is assigned to the state corresponding to the preferred direction of the action. The agents prior contains no information regarding the cardinal directions associated with each action. The small concentration factor distributes more probability weight in the corners of the simplex resulting in less entropic transitions.

Letting N_s denote the number of states one-step away from state s, the Bayesian estimate for maze transitions follows the derivation for Dense Worlds and is given by:

$$\widehat{\Theta}_{a,s,s'} = \frac{F_{ass'} + \alpha}{N_s \cdot \alpha + \sum_{s^*} F_{ass^*}} \tag{2.7}$$

As with Dense Worlds, the Bayesian estimate (2.7) for mazes is a Laplace smoothed histogram.

1-2-3 Worlds

Finally, *1-2-3 Worlds* differ greatly from both Dense Worlds and Mazes in that their transitions are drawn from a discrete distribution rather than a continuous one (see Fig. 2.3). Since this work is heavily rooted in the Bayesian approach, the consideration of worlds with a different priors was an important addition to understanding the dependency of an exploration strategy on these priors. 1-2-3 Worlds consist of $N = 20$ states and $M = 3$ actions. In a given state, action $a = 1$ moves the agent deterministically to a single target state, $a = 2$ moves the agent with probability 0.5 to one of two target states, and $a = 3$ moves the agent with probability 0.333 to one of 3 potential target states. An absorbing state is formed by universally increasing the likelihood that state 1 is chosen as a target. Explicitly, letting Ω_a be the set of all admissible transition distributions for action a:

$$\Omega_a := \{\Theta \in \mathbb{R}^N | \sum_{s'} \Theta_{s'} = 1 \text{ and } \Theta_{s'} \in \{0, \frac{1}{a}\} \forall s'\}$$

the transition distributions are drawn from the following distribution:

$$p(\Theta_{as\cdot}) = \begin{cases} 0 & \text{if } \Theta_{as\cdot} \notin \Omega_a \\ \frac{1-0.75^a}{\binom{N-1}{a-1}} & \text{else if } \Theta_{as1} = \frac{1}{a} \\ \frac{1-(1-0.75^a)}{\binom{N-1}{a}} & \text{otherwise} \end{cases} \tag{2.8}$$

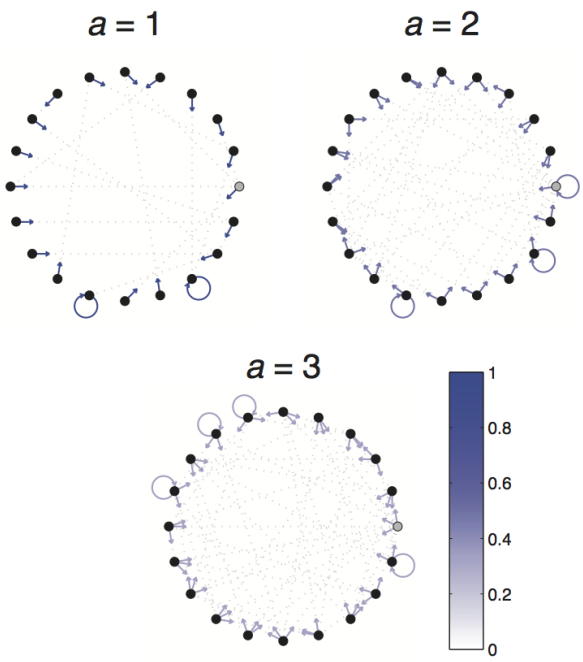

Figure 2.3: Example 1-2-3 World. 1-2-3 Worlds consist of 3 actions (separately depicted) and 20 states (depicted as nodes of the graphs). The transition probabilities associated with taking a particular action are depicted as arrows pointing from the current state to each of the possible resultant states. Arrow color depicts the likelihood of each transition. The absorbing state is depicted in gray.

Bayesian inference in 1-2-3 Worlds differs greatly from Mazes and Dense Worlds because of its discrete prior. If $a, s \to s'$ has been previously observed, then the Bayesian estimate for $\widehat{\Theta}_{ass'}$ is given by:

$$\widehat{\Theta}_{ass'} = \frac{1}{a}$$

If $a, s \to s'$ has not been observed but $a, s \to 1$ has, then the Bayesian estimate is given by:

$$\widehat{\Theta}_{ass'} = \frac{1 - \frac{|\mathscr{S}^*|}{a}}{N - T}$$

Here T is the number of target states that have already been observed. Finally, if neither $a,s \to s'$ nor $a,s \to 1$ have been observed, then the Bayesian estimate is:

$$\widehat{\Theta}_{ass'} = \begin{cases} \frac{1}{a} \cdot \dfrac{1-0.75^a}{1+\left(\binom{a-1}{T}-1\right)\cdot 0.75^a} & \text{if } s' = 1 \\ \dfrac{1-\left(\frac{T}{a}+\widehat{\Theta}_{as1}\right)}{N-T-1} & \text{otherwise} \end{cases}$$

Chapter 3

Information guided exploration

3.1 Assessing the information-theoretic value of planned actions

With the mathematical framework of learning in CMCs laid out, we are now prepared to turn to the central question of how behavior affects the learning process in embodied action-perception loops. The fast reduction of missing information is taken to be the agent's objective during learning-driven exploration (2.3). As discussed in Section 2.3, the Bayesian estimate minimizes the expected missing information and thus solves the inference problem. The control problem of choosing actions to learn quickly nevertheless remains to be solved. Here, I show that Bayesian inference can also be used to predict how much missing information will be removed by an action. I call the decrease in missing information between two internal models the *information gain* ($\mathrm{I_G}$). Letting $\widehat{\Theta}$ be a current model derived from data $\vec{\mathrm{d}}$ and $\widehat{\Theta}^{a,s\to s^*}$ be an updated model after observing a transition from s to s^* under action a, the information gain for this observation is:

$$\mathrm{I_G}(a,s,s^*) := \mathrm{I_M}(\Theta \parallel \widehat{\Theta}) - \mathrm{I_M}(\Theta \parallel \widehat{\Theta}^{a,s\to s^*}) = \sum_{s'} \Theta_{ass'} \log_2 \frac{\widehat{\Theta}^{a,s\to s^*}_{ass'}}{\widehat{\Theta}_{ass'}} \tag{3.1}$$

An exploring agent cannot compute $\mathrm{I_G}$ directly because it depends on the true CMC kernel Θ. It also cannot know the outcome s^* of an action until it has taken it. I therefore again take the Bayesian approach introduced in Section 2.3 and consider the agent to treat Θ and s^* as random variables. Then, by calculating the expected value of $\mathrm{I_G}$, I show in Theorem 2 that an agent can compute an estimate of information gain from its prior belief on Θ and the data it has collected. I term this estimate the *predicted information gain* (PIG).

Theorem 2. *If an agent is in state s and has previously collected data $\vec{\mathrm{d}}$, then the expected information gain for taking action a is given by:*

$$\mathrm{PIG}(a,s) := \mathrm{E}_{s^*,\Theta|\vec{\mathrm{d}}}[\mathrm{I_G}(a,s,s^*)] = \sum_{s^*} \widehat{\Theta}_{ass^*} \mathrm{D_{KL}}(\widehat{\Theta}^{a,s\to s^*}_{as\cdot} \parallel \widehat{\Theta}_{as\cdot}) \tag{3.2}$$

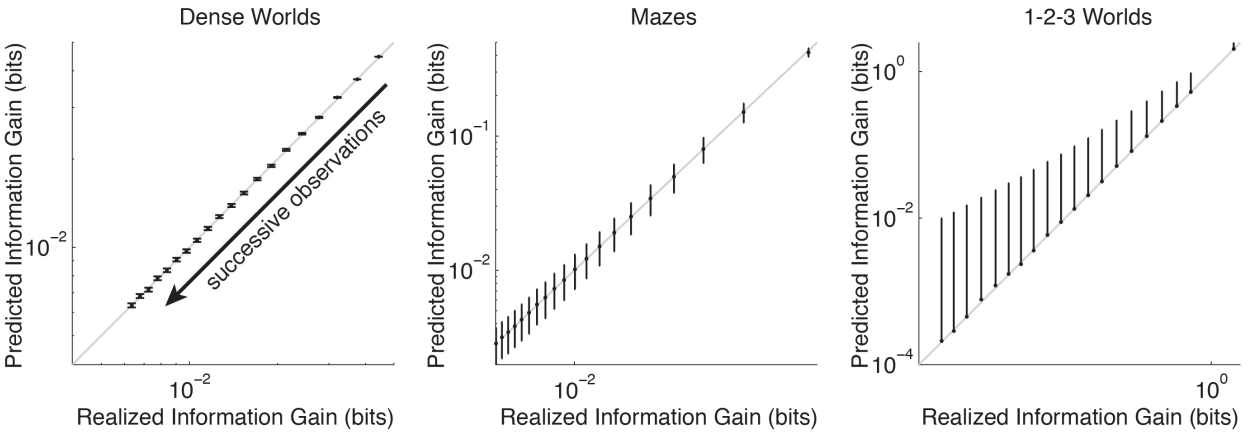

Figure 3.1: Accuracy of predicted information gain. The average predicted information gain is plotted against the average realized information gain. Averages are taken over 200 CMCs, $N \times M$ transition distributions, and 50 trials. Error bars depict standard deviations (only plotted above the mean for 1-2-3 Worlds). The arrow indicates the direction of increasing numbers of observations (top-right =none, bottom-left=19). The unity lines are drawn in gray.

Proof.

$$
\begin{aligned}
\mathrm{E}_{s^*,\boldsymbol{\Theta}|\vec{\mathrm{d}}}[\mathrm{I}_{\mathrm{G}}(a,s,s^*)] &= \mathrm{E}_{s^*,\boldsymbol{\Theta}|\vec{\mathrm{d}}}\left[\sum_{s'} \boldsymbol{\Theta}_{ass'} \log_2\left(\frac{\widehat{\boldsymbol{\Theta}}^{a,s\to s^*}_{ass'}}{\widehat{\boldsymbol{\Theta}}_{ass'}}\right)\right] \\
&= \mathrm{E}_{s^*|\vec{\mathrm{d}}}\left[\sum_{s'} \mathrm{E}_{\boldsymbol{\Theta}|\vec{\mathrm{d}},s^*}[\boldsymbol{\Theta}_{ass'}] \log_2\left(\frac{\widehat{\boldsymbol{\Theta}}^{a,s\to s^*}_{ass'}}{\widehat{\boldsymbol{\Theta}}_{ass'}}\right)\right] \\
&= \mathrm{E}_{s^*|\vec{\mathrm{d}}}\left[\sum_{s'} \widehat{\boldsymbol{\Theta}}^{a,s\to s^*}_{ass'} \log_2\left(\frac{\widehat{\boldsymbol{\Theta}}^{a,s\to s^*}_{ass'}}{\widehat{\boldsymbol{\Theta}}_{ass'}}\right)\right] \\
&= \mathrm{E}_{s^*|\vec{\mathrm{d}}}\left[\mathrm{D}_{\mathrm{KL}}(\widehat{\boldsymbol{\Theta}}^{a,s\to s^*}_{as\cdot} \parallel \widehat{\boldsymbol{\Theta}}_{as\cdot})\right] \\
&= \sum_{s^*} p(s^*|a,s,\vec{\mathrm{d}})\mathrm{D}_{\mathrm{KL}}(\widehat{\boldsymbol{\Theta}}^{a,s\to s^*}_{as\cdot} \parallel \widehat{\boldsymbol{\Theta}}_{as\cdot}) \qquad \text{by (Eq. 2.5)} \\
&= \sum_{s^*} \widehat{\boldsymbol{\Theta}}_{ass^*}\mathrm{D}_{\mathrm{KL}}(\widehat{\boldsymbol{\Theta}}^{a,s\to s^*}_{as\cdot} \parallel \widehat{\boldsymbol{\Theta}}_{as\cdot})
\end{aligned}
$$

□

Interestingly, PIG has a rather intuitive interpretation. In a sense, the agent is imagining the possible outcomes of an action and simply comparing the new model that would result from such an observation to its current model under the missing information measure. Explicitly, it considers the possible outcomes s^* of taking action a in state s. It then determines how each of these results would hypothetically change its internal model. It compares these new hypothetical models,

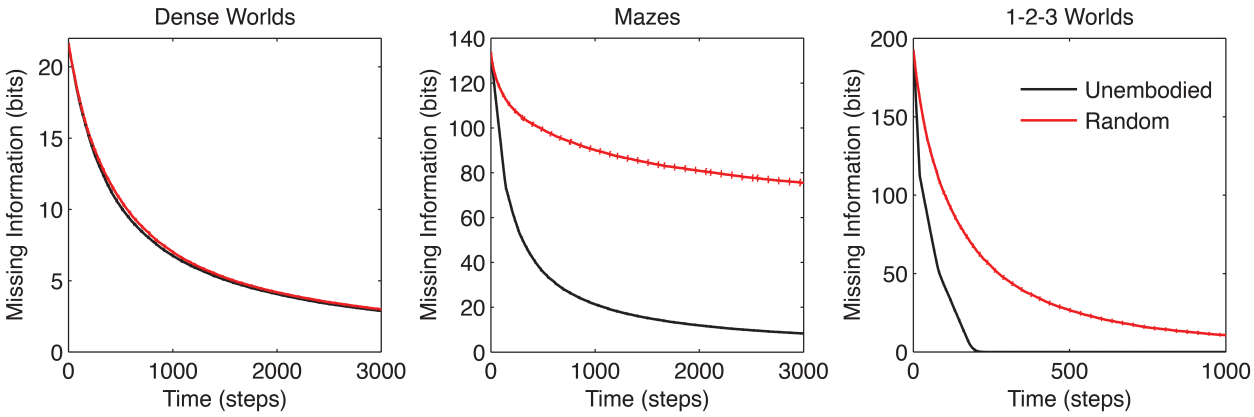

Figure 3.2: Learning curves for control strategies. The average missing information is plotted over exploration time for the unembodied positive control and random action baseline control. Standard errors are plotted as dotted lines above and below learning curves. (n=200)

represented by $\widehat{\Theta}^{a,s\to s^*}_{as\cdot}$, to its current model, $\widehat{\Theta}^{a,s\to s^*}$, by computing the KL-divergence between them. This essentially is calculating the missing information as if the updated hypothetical model were the ground truth. The larger this difference the more information the agent would likely gain if it indeed transitioned to state s^*. Finally, it averages these hypothetical gains according to the likelihood of observing s^* under its current model.

For each class of environments, Fig. 3.1 compares the average PIG with the average realized information gain as successive observation are used to update a Bayesian estimate. In accordance with Theorem 2, in all three environments PIG accurately predicts the average information gain. Thus, theoretically and empirically, PIG represents an accurate estimate of the improvement an agent can expect in its internal model if it takes a planned action in a particular state.

Interestingly, the expression on the RHS of Eq. 3.2 has been previously studied in the field of Psychology where it was introduced ad hoc to describe human behavior during hypothesis testing [63, 83, 87]. To my knowledge, its equality to the predicted gain in information (Theorem 2) is novel. In a later section, I will compare PIG to other measures proposed in the field of Psychology. Again for easy reference, $\mathrm{I_G}$ and PIG have been added to the compilation of terms in Table 2.1.

3.2 Control learners: unembodied and random action

Before introducing and assessing the performance of different explorative strategies, it will be useful to first introduce two control strategies to act as positive and negative benchmarks for performance. A naive strategy would be to select actions uniformly randomly. Such random policies are often employed to encourage exploration in reinforcement learning models. I will therefore use a *random action* strategy as a negative control exhibiting the baseline learning rate of an undirected explorer.

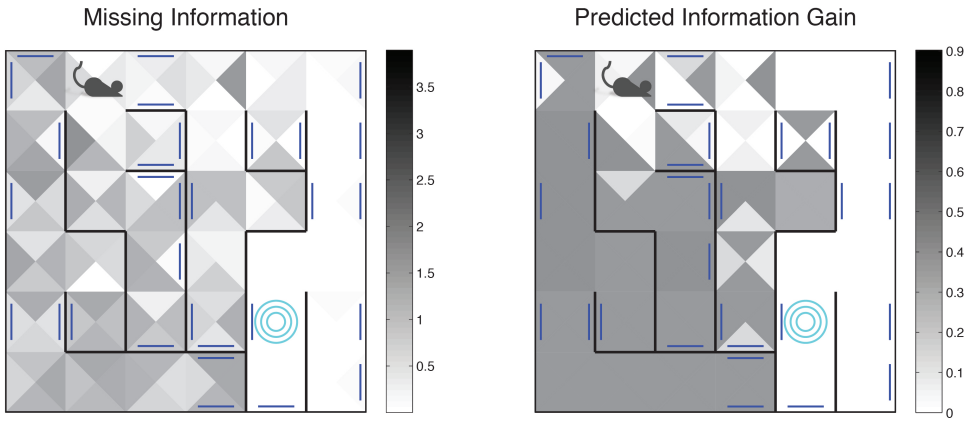

Figure 3.3: The state of knowledge for a random actor that has explored an example maze for 1000 time steps. The missing information (left) or PIG (right) for each action in each state is colorimetrically indicated by the shading of the triangle in that actions direction of highest transition probability.

As a positive control, I developed an *unembodied* agent that achieves an upper bound on expected performance. Unlike an embodied agent, the unembodied control is allowed, at every time step, to choose not only its action but also its state. That is, the unembodied agent is allowed to break the dynamics of the CMC and move itself instantaneously to any state of its choosing before continuing its exploration. For such an agent, optimization of learning decomposes into an independent sampling problem [94]. Since the PIG for each transition distribution decreases monotonically over successive observations (Fig. 3.1), learning by an unembodied agent can be optimized by always sampling from the state and action pair with the highest PIG. Thus, learning can be optimized in a greedy fashion:

$$(a,s)_{\text{Unemb.}} := \arg\max_{(a,s)} \text{PIG}(a,s) \tag{3.3}$$

Comparing the learning performances of the random action and unembodied control (red and black curves respectively in Fig. 3.2) I found a notable difference among the three classes of environments. The performance margin between these two controls is significant in Mazes and 1-2-3 Worlds ($p < 0.001$, Wolcoxon rank-sum test), but not in Dense Worlds ($p > 0.01$). Despite using a naive strategy, the random actor is essentially reaching maximum performance in Dense Worlds, suggesting that exploration of this environment is fairly easy. In contrast, in Mazes and 1-2-3 Worlds, a directed exploration strategy seemed necessary to achieve learning speeds closer to that of the unembodied upper bound.

To illustrate the distribution of information and the failures of random exploration, I have depicted in Fig. 3.3 the missing information and PIG for a random actor that has taken 1000 exploratory steps in an example maze. As is immediately noticeable, the random actor has spent much of its time exploring near the absorbing state. In order to gather high information observations, it will need to more carefully direct its actions.

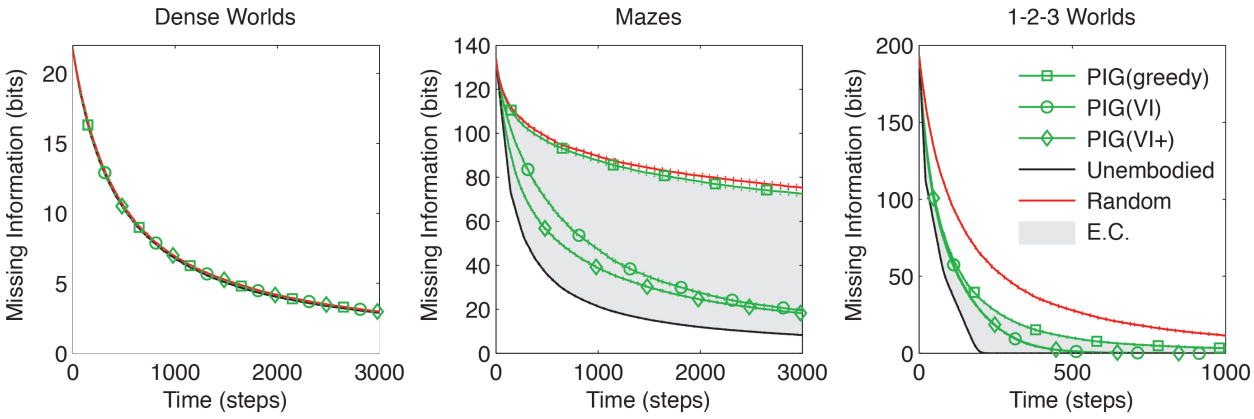

Figure 3.4: Coordinating exploration using predicted information gain. The average missing information is plotted over exploration time for greedy and value-iterated (VI) maximization of PIG. The standard control strategies and the VI+ positive control are also depicted. The grey shaded area represents the embodiment constraint (E.C.). Standard errors are plotted as dotted lines above and below learning curves. (n=200)

3.3 Greedy maximization of PIG by embodied agents

PIG represents a utility function that can be used to guide exploration. Since greedy maximization of PIG is optimal for the unembodied agent, one might expect a similar strategy to be promising for an embodied agent. Unlike the unembodied control, however, the greedy embodied agent, which I denote PIG(greedy), would only be able to select its action, not its state:

$$a_{\text{PIG(greedy)}} := \arg\max_{a} \text{PIG}(a, s) \tag{3.4}$$

The performance comparison between PIG(greedy) (3.4) and the positive control (3.3) is of particular interest because they differ only in that one is embodied while the other is not. As shown in Fig. 3.4 the performance difference is largest in Maze worlds, moderate though significant in 1-2-3 Worlds and smallest in Dense Worlds ($p < 0.001$ for Mazes and 1-2-3 Worlds, $p > 0.001$ for Dense Worlds). To quantify the embodiment constraint faced in each class of CMCs, I defined an *embodiment index* as the relative difference between the areas under the learning curves for PIG(greedy) and the unembodied control. The average embodiment indices for Dense Worlds, Mazes, and 1-2-3 Worlds are 0.02, 2.59, and 1.27, respectively. Finally, whereas PIG(greedy) yielded no improvement over random action in Dense Worlds and Mazes ($p > 0.001$), it significantly improved learning in 1-2-3 Worlds ($p < 0.001$), suggesting that this utility function was most immediately beneficial in the class of CMCs with discrete priors.

3.4 Coordinated maximization of PIG by embodied agents

Greedy maximization of PIG only accounts for the immediately available information gains and fails to account for the effect an action can have on future learning. In particular, when the potential for information gain is concentrated at remote states in the environment, it may be necessary to coordinate actions over time. To see this, one could reexamine the distribution of information for the exploring agent depicted in Fig. 3.3. At its current location, greedy maximization of the PIG would suggest selecting an action that will most likely take the agent back towards its previously explored locations. Instead, one could imagine that the agent should forgo the immediate information gains available to it and choose the action that will most likely take it left, towards the unexplored regions of the maze.

Unfortunately, forward estimation of total future PIG is intractable. I therefore employed a back-propagation approach previously developed in the field of economics called *value-iteration* (VI) [13]. The estimation starts at a distant time point (initialized as $\tau = 0$) in the future with initial values equal to the PIG for each state-action pair:

$$Q_0(a,s) := \mathrm{PIG}(a,s)$$

Then propagating backwards in time, a running total of estimated future value is maintained according to the following update rule:

$$Q_{\tau-1}(a,s) := \mathrm{PIG}(a,s) + \gamma \sum_{s' \in \mathscr{S}} \widehat{\Theta}_{ass'} \cdot V_\tau(s') \tag{3.5}$$
$$\text{where } V_\tau(s) := \max_a Q_\tau(a,s)$$

Here, γ is a discount factor, set to 0.95. Such discount factors are commonly employed in value-iteration algorithms to favor more immediate gains over gains further in the future [13]. I briefly note this particular value of γ was chosen to be consistent with previous literature and a discount factor of $\gamma = 1$ does not qualitatively change any of the subsequent results (data not shown). Nevertheless, as discussed later, discounting may also help in part to account for the decreasing return on information of successive observations (see Fig. 3.1).

Ideally, the true transition dynamics Θ would be used in Eq. 3.5, but since the agent must learn these dynamics, it employs its internal model $\widehat{\Theta}$ instead. Applying the VI algorithm to PIG, we construct a behavioral policy PIG(VI) that coordinates actions over several time steps towards the approximate maximization of expected information gain:

$$a_{\mathrm{PIG(VI)}} := \arg\max_a Q_{-10}(a,s);$$

As shown in Fig. 3.4, the use of VI to coordinate actions yielded the greatest gains in Mazes, with moderate gains also seen in 1-2-3 Worlds. Along with the embodiment indices introduced above, these results support the hypothesis that worlds with high embodiment constraints require agents to coordinate their actions over several time steps to achieve efficient exploration.

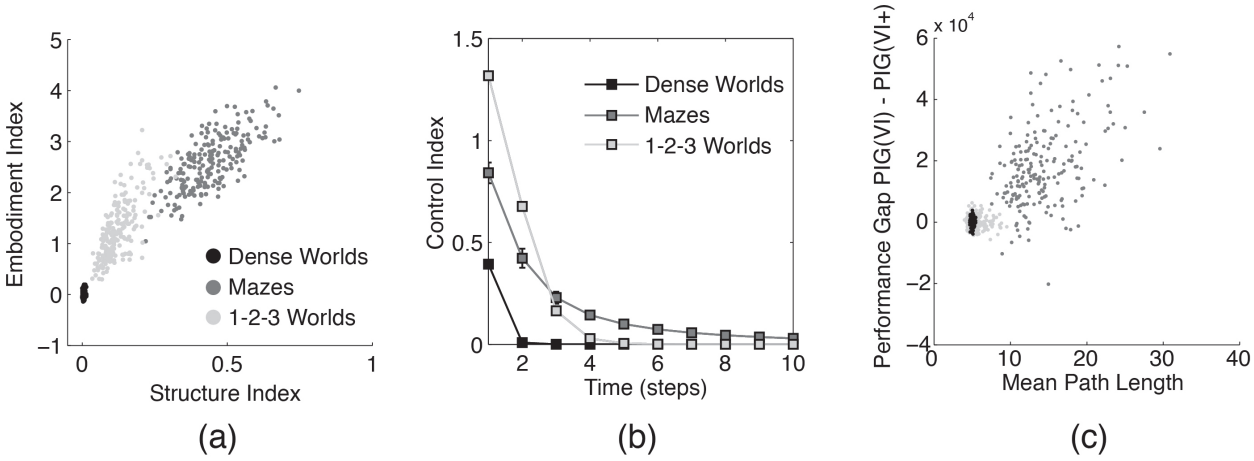

Figure 3.5: Quantifying the structure of the worlds. (a) The embodiment index, defined in Section 3.3, is plotted against the structure index for each of 200 Dense Worlds, Mazes, and 1-2-3 Worlds. (b) For the same CMCs, the average controllability is plotted as a function of the number of time steps the state lies in the future. The error bars depict standard deviations. (c) Again for the same CMCs, the learning performance gap in between PIG(VI) and PIG(VI+) is plotted against the mean path length between any two states.

Bellman showed that VI accurately estimates future gains when the true transition dynamics Θ are known and when the utility function is stationary [13]. Neither of these requirements are met in our case, and PIG(VI) is therefore only an approximation of future gains. Nevertheless, as I will show, its utility is validated by its superior performance when compared to an array of competitor exploration strategies.

While a learning agent cannot use the true dynamics for VI, we can ascertain how much this impairs its exploration by considering a second positive-control PIG(VI+) that is allowed to use the true dynamics for purposes of coordinating its actions. That is, the PIG(VI+) control uses Θ instead of $\widehat{\Theta}$ in Eq. 3.5 above. Learning efficiency under PIG(VI+) only differs from PIG(VI) in Mazes, and this difference is relatively small compared to the gains made over the random or greedy behaviors (Fig. 3.4). Altogether these results suggest that PIG(VI) may be an effective strategy employable by embodied agents for coordinating explorative actions towards learning.

3.5 Structural features of the three worlds and their effects on exploration

To elucidate the interaction between behavioral strategy and the dynamical structure of the explored world, I next considered how structural differences in the three classes of environments correlated with an agents ability to explore. In particular, I developed three measures quantifying the structure of a world: 1) their tendency to draw agents into a biased distribution over states, 2) the amount of control a single action provides an agent over its future states, and 3) the average distance between any two states.

State bias: To assess how strongly a world biases the state distribution of an agent I quantified the unevenness of the equilibrium distribution for a random action policy. The equilibrium distribution Ψ gives the limit likelihood that an agent will be in a particular state at a distant time-point in the future. To quantify the bias of this distribution, I defined a *structure index* (SI) as the relative difference between the entropy of the equilibrium distribution $H(\Psi)$ and the entropy of the uniform distribution $H(U)$:

$$SI(\Psi) := \frac{H(U) - H(\Psi)}{H(U)}$$

where:

$$H(p) := -\sum_{s \in \mathscr{S}} p(s) \log_2(p(s))$$

In Fig. 3.5a, the structure indices for 200 worlds in each class of environment were plotted against their embodiment indices (defined in Section 3.3). As depicted, the embodiment index correlates strongly with the structure index suggesting that state bias represents a significant challenge embodied agents face during exploration.

Controllability: To measure the capacity for an agent to control its state trajectory I computed a control index as the mutual information between a random action a_0 and an agent's state t time steps in the future s_t averaged uniformly over possible starting states s_0:

$$\begin{aligned} CI(t) &= \sum_{s_0 \in \mathscr{S}} \frac{1}{N} \mathrm{MI}[A_0, S_t | s_0] \\ &= \sum_{s_0 \in \mathscr{S}} \frac{1}{N} \left(\sum_{a_0 \in \mathscr{A}, s_t \in \mathscr{S}} p(a_0, s_t | s_0) \log_2 \left(\frac{p(s_t | a_0, s_0)}{p(s_t | s_0)} \right) \right) \end{aligned}$$

As shown in Fig. 3.5b, an action in a Maze or 1-2-3 World has significantly more control over future states than an action in Dense Worlds. One would speculate that the utility of coordinating actions over several time-steps would be limited by the amount of control those actions had on the subsequent states. Consistent with this hypothesis, actions had particularly long-reaching effects on state progression in Mazes, which, of the three classes of CMCs, had shown the greatest gains from VI (Figure 3.4). 1-2-3 Worlds also revealed high controllability, but only over the more immediate future. Accordingly, 1-2-3 Worlds showed intermediate gains from coordinated actions.

Mean Path Length: To assess the size of each CMC within the context of behavior, I calculated the minimum expected path length for moving between any pair of states. To do this, I first determined the action policy that would minimize the expected path length to any target state. I then calculated the expected number of time-steps it would take an agent to navigate to that target state while employing this optimal policy. The average value of this expected path length taken across start and target states was used as a measure of the extent of the CMC (see Appendix A.1 for detailed methods). I had previously found that the three classes of CMCs differed in the relative performance between the PIG(VI) explorer and the PIG(VI+) control. Since these two strategies

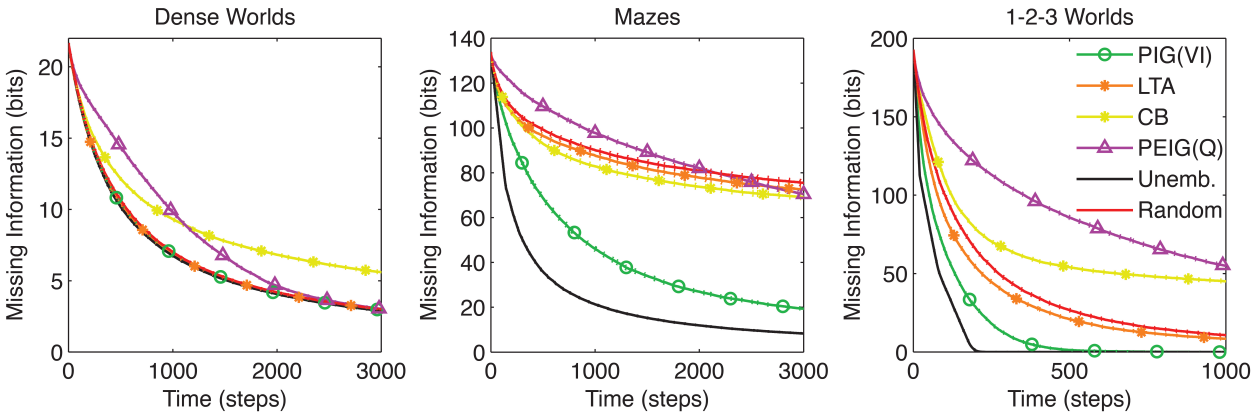

Figure 3.6: Comparison to previous exploration strategies. The average missing information is plotted over time for PIG(VI) agents along with three exploration strategies from the literature: least taken action (LTA) [11, 109, 114], counter-based (CB) [128], and Q-Learning on posterior expected information gain (PEIG(Q)) [123]. The standard control strategies are also shown. Standard errors are plotted as dotted lines above and below learning curves. (n=200)

differ only in that the former uses the agent's internal model to coordinate its actions while the latter is allowed to use the true world dynamics, I wondered if the performance gap between the two (the area between their two learning curves) could be related to the path length to a potential source of information. Indeed, comparing this performance gap to the mean path length for each world, I found a strong correlation, as shown in Fig. 3.5c. This suggests that an accurate internal model is more necessary for effectively coordinating actions during exploration of more spatially extended worlds. Finally, one might notice that in Mazes the mean path lengths are typically larger than 10 time steps, the planning horizon used in Value Iteration. Ten was chosen simply as a round number and it may be surprising that it works as well as it does in such spatially extended worlds. I believe two factors may contribute to this. First, it is likely that states of high informational value will be close together. Coordinating actions towards a nearby state of high value will therefore likely bring the agent closer to other states of potentially higher value. Second and, I suspect, more importantly, since the mean path length is an average, a VI planner can direct its action towards a high information state under the possibility that it might reach that state within 10 time steps even if the expected path length to that location is significantly longer.

Taken together, these results begin to reveal the structural features of a world that determine the constraints faced by embodied agents and the capacity for those agents to overcome or compensate for these constraints.

3.6 Comparison to previous explorative strategies

Many models of exploration have been previously developed in the field of reinforcement learning (RL). As discussed in the introduction, these models usually focus on the indirect role of explo-

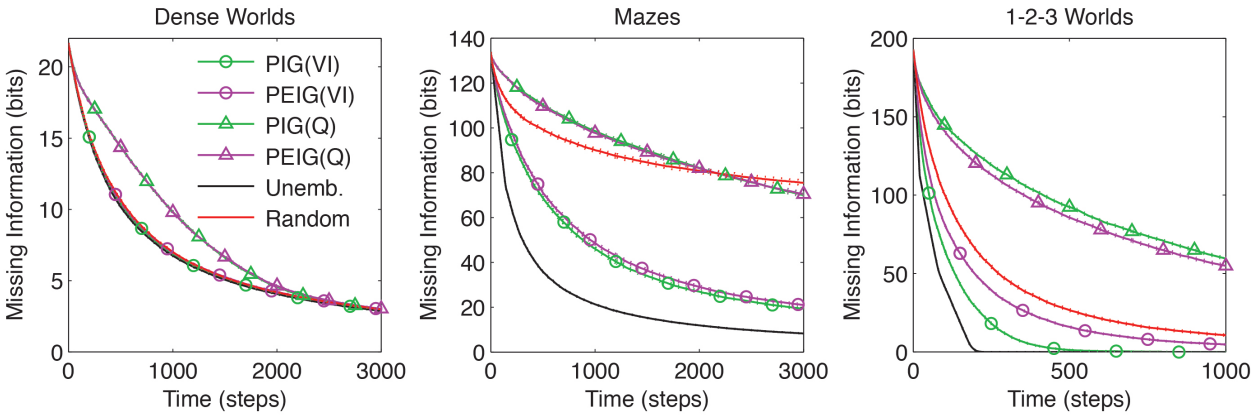

Figure 3.7: Comparison between different features of current and previous exploration strategies. The average missing information is plotted over time for agents that apply either VI (circles) or Q-learning (triangles) towards maximization of either PIG (green) or PEIG (magenta). Standard control strategies are also shown. Standard errors are plotted as dotted lines above and below learning curves. (n=200)

ration in reward acquisition rather than its direct role in learning world structure. Indeed, many RL models of exploration are explicitly guided by these external rewards. Such models would not be applicable under the CMC framework as there are no external rewards. Several other RL exploration principles however do not require external rewards and can be implemented in the CMC framework. In this section, I compare these various methods to PIG(VI) under the missing information learning objective. Random action is perhaps the most commonly employed exploration strategy in RL. As I have already demonstrated, random action is only efficient for exploring Dense Worlds. The following directed exploration strategies have also been developed in the RL literature (their learning curves are plotted in Fig. 3.6):

Least Taken Action (LTA): Under LTA, an agent will always choose the action that it has performed least often in the current state [11, 109, 114]. Like random action, LTA yields uniform sampling of actions in each state. Across worlds, LTA fails to significantly improve on the learning rates seen under random action ($p > 0.001$ for all three environments).

Counter-Based Exploration (CB): Whereas LTA actively samples actions uniformly, CB attempts to induce a uniform sampling across states. To do this, it maintains a count of the occurrences of each state, and chooses its action to minimize the expected count of the resultant state [128]. CB performs even worse than random action in Dense Worlds and 1-2-3 Worlds ($p < 0.001$). It does outperform random actions in Mazes but falls far short of the performance seen by PIG(VI) ($p < 0.001$).

Q-learning on Surprise (PEIG(Q)): Storck *et al.* [123] developed Surprise as a measure to quantify past changes in an agent's internal model which they used to guide exploration under a Q-learning algorithm [125]. Interestingly, it can be shown that Surprise as employed by Storck *et al.* is equivalent to the posterior expected information gain (PEIG), a posterior analogue to our

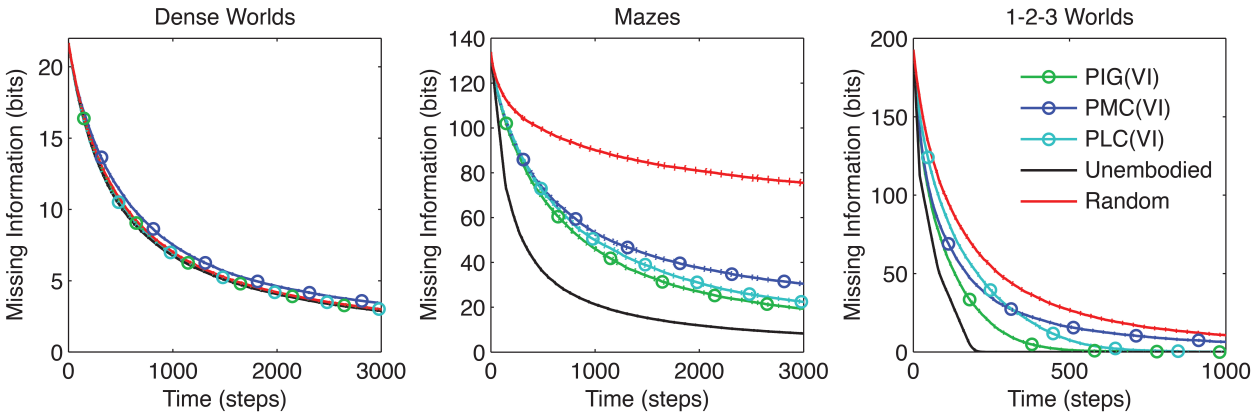

Figure 3.8: Comparison between utility functions. The average missing information is plotted over time for agents that employ VI to maximize long-term gains in the three objective function, PIG, PMC, or PLC. The standard control strategies are also shown. (n=200)

PIG utility function (see Appendix A.2 and Table 2.1). Q-learning is a model-free approach to maximizing long-term gains of a utility function [125]. Implementing their strategy, we found that, as with CB, PEIG(Q) generally performed even worse than random action.

The results in Fig. 3.6 show that PIG(VI) outperforms the previous explorative strategies at learning in structured worlds. It is important to note that all of these RL strategies were originally developed to encourage exploration for the sake of improving reward-acquisition, and their poor performance under the learning objective does not conflict with their previously demonstrated utility within the reinforcement learning framework.

The limited number of competitors that can be pitted against the PIG(VI) strategy reflects the dearth of consideration given to learning-driven, reward-free exploration in reinforcement learning. PEIG(Q) is unique amongst previous studies in its attempt to employ information theoretic constructs to guide exploration, but as was demonstrated it was a poor contender. Given the similarities between PIG(VI) and PEIG(Q), I wanted to investigate whether their differences in performance resulted from the choice of utility function (PIG versus PEIG) or from their method of coordinating actions (VI versus Q-learning). I therefore considered all permutations of these two components of the exploration strategies. In Fig. 3.7, we see that across all three classes of CMCs the method of coordinating actions strongly effects performance, with the model-based VI outperforming the model-free Q-learning in each class. Furthermore we find a significant performance boost in 1-2-3 Worlds from predicting future information gains (PIG) rather than simply estimating past information gains (PEIG). These permutations have not been tried before in the reinforcement literature and represent new findings. In addition, it should be noted that PEIG, being dependent on past observations, is ill-defined at the start of exploration and thus must be seeded with an initial value. These initial values add additional free parameters to the PEIG strategies. In these experiements I generously seeded PEIG with the expect information gain of the first observation. Increasing these intial seeds did not qualitatively change these results, while decreasing them led to significant

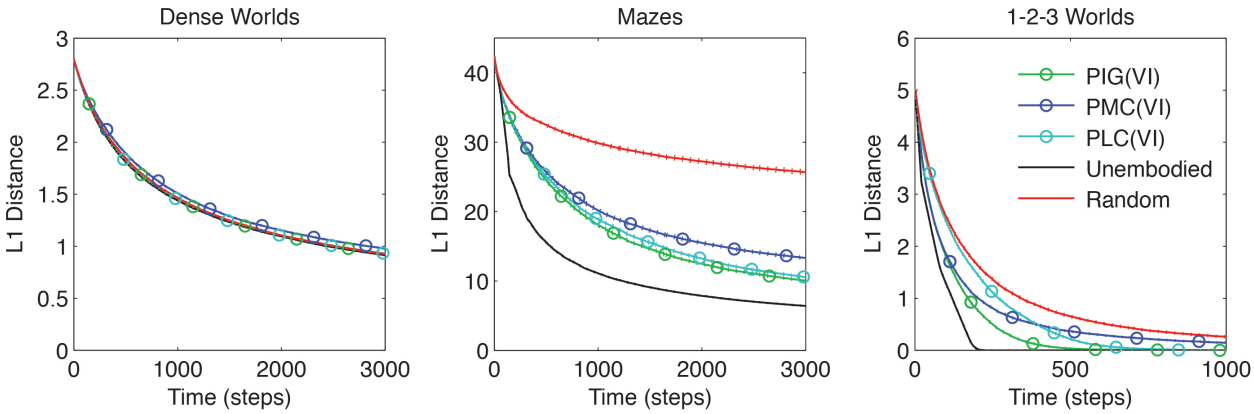

Figure 3.9: Comparison between utility functions under L1 objective. The average L1 distance is plotted over time for agents that coordinate actions using VI to maximize long-term gains in PIG, PMC, or PLC. Standard control strategies are also shown. Standard errors are plotted as dotted lines above and below learning curves. (n=200)

decreases in the performance of all PEIG strategies (data not shown).

3.7 Comparison to utility functions from Psychology

The previous sections introduced PIG as a useful means of directing efficient exploration in CMCs. Interestingly, independent findings by Oaksford and Chater in the field of Psychology have suggested that the maximization of a measure similar to PIG can be used to explain human behavior during hypothesis testing [87]. These results were not derived in the framework of a closed action-perception loops and did not consider sequences of actions. Additionally, they did not derive the expression of PIG from first principles and simply introduced it ad hoc. Along with PIG, they also introduced several other ad hoc measures and could not distinguish between the predictive power of PIG and two of these alternatives. Inspired by these results, I investigated these two other measures. Like PIG, both are measures of the difference between the current and hypothetical future internal models:

Predicted mode change (PMC) approximates the height difference between the modes of the current and future internal models [10, 83]:

$$\mathrm{PMC}(a,s) = \sum_{s^*} \widehat{\Theta}_{ass^*} \left[\max_{s'} \widehat{\Theta}^{a,s\to s^*}_{ass'} - \max_{s'} \widehat{\Theta}_{ass'} \right] \tag{3.6}$$

Predicted L1 change (PLC) approximates the average L1 distance between the current and future internal models [63]:

$$\mathrm{PLC}(a,s) = \sum_{s^*} \widehat{\Theta}_{ass^*} \left[\frac{1}{N} \sum_{s'} \left| \widehat{\Theta}^{a,s\to s^*}_{ass'} - \widehat{\Theta}_{ass'} \right| \right] \tag{3.7}$$

Both measures have been added to Table 2.1. Following the approach I took for PIG, I tested agents that approximately maximized PMC or PLC using VI. As Fig. 3.8 reveals, PIG(VI) proved again to be the best performer overall. In particular, PIG(VI) significantly outperforms PMC(VI) in all three environments, and PLC(VI) in 1-2-3 Worlds ($p < 0.001$). Nevertheless, PMC and PLC achieved significant improvements over the baseline control in Mazes and 1-2-3 Worlds, highlighting the benefit of coordinated actions across different utility functions. Interestingly, when performance was measured by an L1 distance instead of missing information, PIG(VI) still outperformed PMC(VI) and PLC(VI) in 1-2-3 Worlds (see Fig. 3.9).

Chapter 4

Ultimate cause of exploration

4.1 Generalized utility of exploration

Up until now, I have focused on learning as the proximate (i.e. behavioral) cause of exploration [4, 70, 99]. Accordingly, I have considered reduction of missing information to be the primary objective of exploration from a behavioral viewpoint. In this chapter, I explore the possible ultimate (i.e. evolutionary) causes of exploration. That is, I ask what adaptive benefit might learning-driven exploration offer an individual. As discussed in the Introduction, the reinforcement learning perspective would suggest that exploration increases the chances that an individual will find food, mates, shelter, etc. leading to its survival/reproductive success. In contrast, Psychologists suggest the adaptive benefits of learning-driven exploration lie in the general utility of possessing an accurate internal model of the world [55,97,98,103,104]. The Psychologist Stephen Kaplan nicely presented this theory in his essay "Cognitive Maps, Human Needs and the Designed Environment":

> Humans, like other animals, operate in a spatial world.... To behave effectively with respect to extended space, especially when different places are interesting at different times, for different reasons, and never for sure, requires an organized approach.... It would be necessary to have an overall conception of the layout of the spatial environment, and of the distribution of the assets and the dangers.... It thus appears that a well structured memory, a cognitive map of the spatial environment, would be essential for survival under circumstances of this kind. A cognitive map is, however, an outcome of experience, and there is no assurance that random or unmotivated experience would lead to a cognitive map that is either extensive or well structured.... If the quality of cognitive map were related to the probability of survival, then those who survived would have been those who loved to explore, who craved to know, whose restlessness and eagerness for new sights constantly led them to map-extending experiences. [56]

To investigate this hypothesis, I wanted to compare the general utility of the internal models, or cognitive maps as Kaplan might call them, learned by the various exploration methods. I therefore assessed the ability of the explorers to apply their internal models towards solving an array of goal-

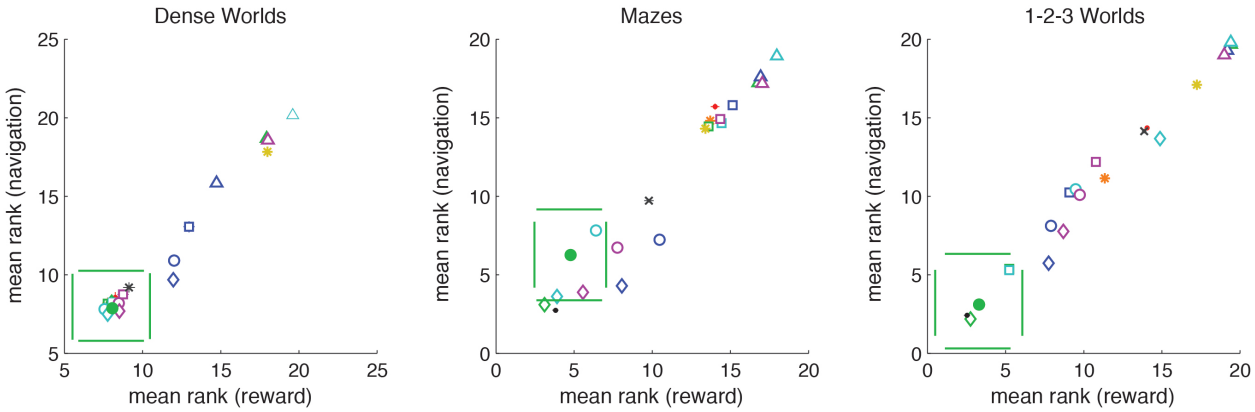

Figure 4.1: Demonstration of generalized utility. For each world (n=200), explorative strategies are ranked for average performance on the navigational tasks (averaged across N start states and N target states) and the reward tasks (averaged across N start states and 10 randomly generated reward distributions). The average ranks are plotted with standard deviations. PIG(VI) is depicted as a filled green circle. Strategies lying outside the pair of horizontal green lines differ significantly from PIG(VI) in navigational performance. Strategies lying outside the pair of vertical green lines differ significantly from PIG(VI) in reward performance ($p < 0.0001$). The different utility functions and heuristics are distinguished by color: PIG(green), PEIG(magenta), PMC(dark-blue), PLC(cyan), LTA(orange), CB(yellow). The different coordination methods are distinguished by symbol: Greedy(squares), VI(circles), VI+(diamonds), Heuristic Strategies(asterisks). The two standard controls are depicted as points as follows: Unembodied(black), Random(red). The BOSS reinforcement learner is depicted by a black cross.

directed tasks. It should be noted that these studies were performed without any changes to the exploration strategies employed by the agent. To do this, I interrupted an agent's exploration at several benchmark time points. I then asked, given its present internal model, how the agent would solve a particular task. I would then allow it to continue exploring and would assess its solutions off-line. I compared the solutions each explorer provided to the optimal solution derived from the true transition dynamics. I considered two types of tasks, navigation and reward acquisition:

*N*avigation: Given a starting state, the agent has to quickly navigate to a target state.

*R*eward Acquisition: Given a starting state, the agent has to gather as much reward as possible over 100 time steps. Reward values are drawn from a normal distribution and randomly assigned to every state in the CMC. The agent is told the reward value of each state.

For each task, I calculated the behavioral strategy that would optimize performance under the internal model. As a positive control, I also calculated a true optimal policy that maximizes performance given the true CMC kernel. The difference in realized performance between the agent's policy and the control was used as a measure of navigational or reward loss. For detailed methods, please see Appendix A.3.

Fig. 4.1 depicts the average rank in the navigational and reward tasks for the different explorative strategies. In all environments, for both navigation and reward acquisition, PIG(VI) always grouped with the top performers ($p > 0.001$), excepting positive controls. PIG(VI) was the only strategy to do so. Thus, the explorative strategy that optimized learning under the missing information objective function also prepared the agent for accomplishing arbitrary goal-directed tasks.

This assessment of the generalized utility of efficient exploration differs from the standard reinforcement learning paradigm in that it tests an agent across multiple tasks. The agent therefore cannot simply learn habitual sensorimotor responses specific to a single task. Though most reinforcement learning studies consider only a stationary, unchanging reward structure, I wanted to compare PIG(VI) to reward-driven exploration. BOSS is a state-of-the-art model-based reinforcement learning algorithm [5]. To implement reward-driven exploration I trained a BOSS reinforcement-learner to navigate to internally chosen target-states. After reaching its target, the BOSS agent would randomly select a new target, updating its model reward structure accordingly. I then assessed the internal model formed by the BOSS explorer under the same navigational and reward acquisition tasks. As can be seen in Fig. 4.1, BOSS (black asterisk) did not perform as well as PIG(VI) at either class of objectives despite being trained specifically on the navigation task.

4.2 Direct competition between exploring agents

As a final test of the generalized utility of efficient exploration, I placed the different exploring agents head-to-head in a game of tag. I first allowed a pair of agents to explore a common maze for 1000 time-steps. I then placed the agent designated as "It" and the agent designated as the target in opposite corners of the maze. These roles could also be thought of as predator and prey respectively. "It" was then given 25 time-steps to try and tag its target. Planning in a game of tag is difficult, especially when one must anticipate the movements of the opponents. Since I was not interested in introducing complex new methods to approximate the opponent's actions, particularly as these approximations could unfairly favor one explorative strategy over another, I chose instead to allow both "It" and the target to know directly the plans of their opponent. Both competitors however would have to use their own internal models to devise their strategy. The difference between being caught and escaping could rest in the accuracy of an agent's assessments on the likely outcomes of its and its opponent's actions. Given the long run-times of these experiments, I decided to focus on comparing my PIG(VI) to the PEIG(Q) and PEIG(VI) competitors, the former because it is the most closely related exploration strategy to mine previously developed in the literature and the latter because of its close performance to PIG(VI) in exploring mazes. I also included the two controls, random action and the unembodied positive control, in these experiments.

In Fig. 4.2, I show the distribution of time-to-tag lengths for the various competitions across 50 different mazes. For each maze, both the PIG(VI) explorer and its opponent were given one opportunity to be "It" and one opportunity to be the target. Consistent with the theory that efficient explorers are better able to solve complex tasks, PIG(VI) outperformed all other strategies (except the unembodied positive control) taking less time to tag the opponent when "It" and avoiding being tagged for longer when not "It". The next closest contender was the PEIG(VI) strategy introduced in this manuscript. The closeness between PIG(VI) and PEIG(VI) is consistent with their similar

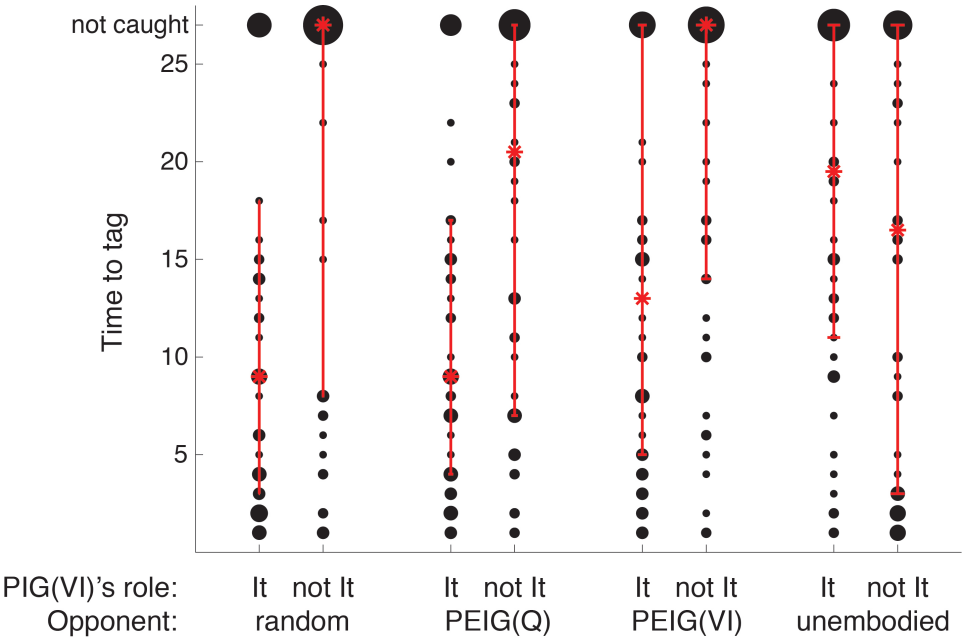

Figure 4.2: Performance in one-on-tag competitions. PIG(VI) was pitted against four competitor strategies. Each column depicts the distribution of times taken for It to tag its opponents across 50 mazes. The number of mazes in which a tag was achieved in the indicated time is proportional to the area of each circle. The results are paired based on PIG(VI)'s opponent. In each pair, the first column represents the condition where PIG(VI) is It and the second column represents the condition where PIG(VI) is the target. Asterisks denote median tag time and error bars indicates lower and upper quartiles.

performance under the learning objective. It would be interesting to compete these two strategies in a world with a discrete prior, like 1-2-3 Worlds, but this experiment fell outside the scope of this thesis.

Chapter 5

Exploring with inaccurate priors

The optimality of the Bayesian estimate (Theorem 1) and the estimation of information gain (Theorem 2) both require an accurate prior over the transition kernels. For biological agents, such priors could have been learned from earlier exploration of related environments or may represent hardwired beliefs optimized by evolutionary pressures. Alternatively, an agent could attempt to simultaneously learn a prior while exploring its environment. Finally, biological agents may not always have access to an accurate prior for an environment and it remains an open question how well an exploring agent could fair in light of an inappropriate prior. In this chapter, I begin to address some of these issues. I begin by demonstrating that maximum-likelihood estimation is sufficient for an agent to accurately estimate its prior when exploring Dense Worlds and Mazes. I then show how the exploration strategy developed in this thesis can be applied to learning the dynamics of a countinuous system. The representation of a continuous system by a discrete internal model is a significant departure from the accurate priors of my earlier experiments and demonstrates the utility of the information theoretic approach even for complex dynamic systems.

5.1 Learning the prior distribution

In the previous experiments, the agent's were given an accurate prior with which to perform Bayesian inference and direct exploration. In particular, in Mazes and Dense Worlds, the agents were given the concentration factor α of the Dirichlet distribution from which the transition distributions were drawn. I wanted to see if this information was necessary to obtain efficient exploration. Specifically, I wanted to see if the priors could be learned over the course of exploration while still being used to direct an agents actions. A Bayesian approach to learning the prior, for example by using a hyper-prior distribution, proved intractable. I therefore considered the Maximum Likelihood Estimate (MLE) of the prior. Given a set of data, the MLE $\widehat{\alpha}_{\text{MLE}}$ identifies the value of α for which the likelihood of the data, or equivalently the log-likelihood, is largest:

$$\widehat{\alpha}_{\text{MLE}} = \arg\max_{\widehat{\alpha}} \; p(\vec{\text{d}}|\alpha = \widehat{\alpha}) = \arg\max_{\widehat{\alpha}} \; log\left(p(\vec{\text{d}}|\alpha = \widehat{\alpha})\right)$$

Letting $\vec{\mathrm{d}}_{as\cdot}$ be the data for transitions originating in state s and given action a and once again letting $\boldsymbol{F}_{ass'}$ be a count of the number of times s' occured in $\vec{\mathrm{d}}_{as\cdot}$, the likelihood of the data is:

$$\begin{aligned}
p(\vec{\mathrm{d}}|\alpha=\widehat{\alpha}) &= \prod_{s\in\mathscr{S},a\in\mathscr{A}} p(\vec{\mathrm{d}}_{as\cdot}|\widehat{\alpha}) \\
&= \prod_{s\in\mathscr{S},a\in\mathscr{A}} \int_{\Delta_{N_s-1}} p(\vec{\mathrm{d}}_{as\cdot}|\Theta_{as\cdot}) f(\Theta_{as\cdot}|\widehat{\alpha}) d\Theta_{as\cdot} \\
&= \prod_{s\in\mathscr{S},a\in\mathscr{A}} \int_{\Delta_{N_s-1}} \prod_{s'\in\mathscr{S}} \Theta_{ass'}^{\boldsymbol{F}_{ass'}} \frac{\prod_{s'}\Theta_{ass'}^{\widehat{\alpha}-1}}{Z(\widehat{\alpha})} d\Theta_{as\cdot} \\
&= \prod_{s\in\mathscr{S},a\in\mathscr{A}} \frac{1}{Z(\widehat{\alpha})} \int_{\Delta_{N_s-1}} \prod_{s'\in\mathscr{S}} \Theta ass'^{\boldsymbol{F}_{ass'}+\widehat{\alpha}-1} d\Theta_{as\cdot} \\
&= \prod_{s\in\mathscr{S},a\in\mathscr{A}} \frac{Z(\boldsymbol{F}_{as\cdot}+\widehat{\alpha})}{Z(\widehat{\alpha})}
\end{aligned}$$

Recalling that:

$$Z(\boldsymbol{x}) = \frac{\prod_i \Gamma(x_i)}{\Gamma(\sum_i x_i)}$$

We can derive the log-likelihood of the data:

$$\begin{aligned}
\mathscr{L}(\vec{\mathrm{d}}) &= \log(p(\vec{\mathrm{d}}|\widehat{\alpha}=\widehat{\widehat{\alpha}}) \\
&= log\left(\prod_{s\in\mathscr{S},a\in\mathscr{A}} \frac{Z(\boldsymbol{F}_{as\cdot}+\widehat{\alpha})}{Z(\widehat{\alpha})}\right) \\
&= \sum_{s\in\mathscr{S},a\in\mathscr{A}} \log\left(\frac{\prod_{s'\in\mathscr{S}}\Gamma(\boldsymbol{F}_{ass'}+\widehat{\alpha})}{\Gamma(N_s\widehat{\alpha}+\sum_{s'\in\mathscr{S}}\boldsymbol{F}_{ass'})}\frac{\Gamma(N_s\widehat{\alpha})}{\prod_{s'\in\mathscr{S}}\Gamma(\widehat{\alpha})}\right) \\
&= \sum_{s\in\mathscr{S},a\in\mathscr{A}} \left[\sum_{s'\in\mathscr{S}} \log\Gamma(\boldsymbol{F}_{ass'}+\widehat{\alpha}) - \log\Gamma(N_s\widehat{\alpha}+\sum_{s'\in\mathscr{S}}\boldsymbol{F}_{ass'}) + \log\Gamma(N_s\widehat{\alpha}) - N_s\log\Gamma(\widehat{\alpha})\right]
\end{aligned} \tag{5.1}$$

Given an uncertain prior, an exploring agent can formulate an MLE of α by maximizing Eq. 5.1. This was done using Quasi-Newton maximization as implemented by the minFunc program freely distributed by Mike Scmidt (http://www.di.ens.fr/ mschmidt/Software/minFunc.html). To avoid divergent estimates, the inferred $\widehat{\alpha}$ was constrained to a maximum value of 20. For the experiments described in Fig. 2.1, PIG(VI) explorers determined an MLE of α after every step of exploration and updated their predicted information gain using their new estimates of the prior. As shown in Figs. 2.1 a and b, this approach was sufficient to allow the exploring agents to quickly and accurately predict the concentration factor. Accordingly, these agents also quickly recovered efficient learning of the transition dynamics, and after about 20 exploration steps their internal models were

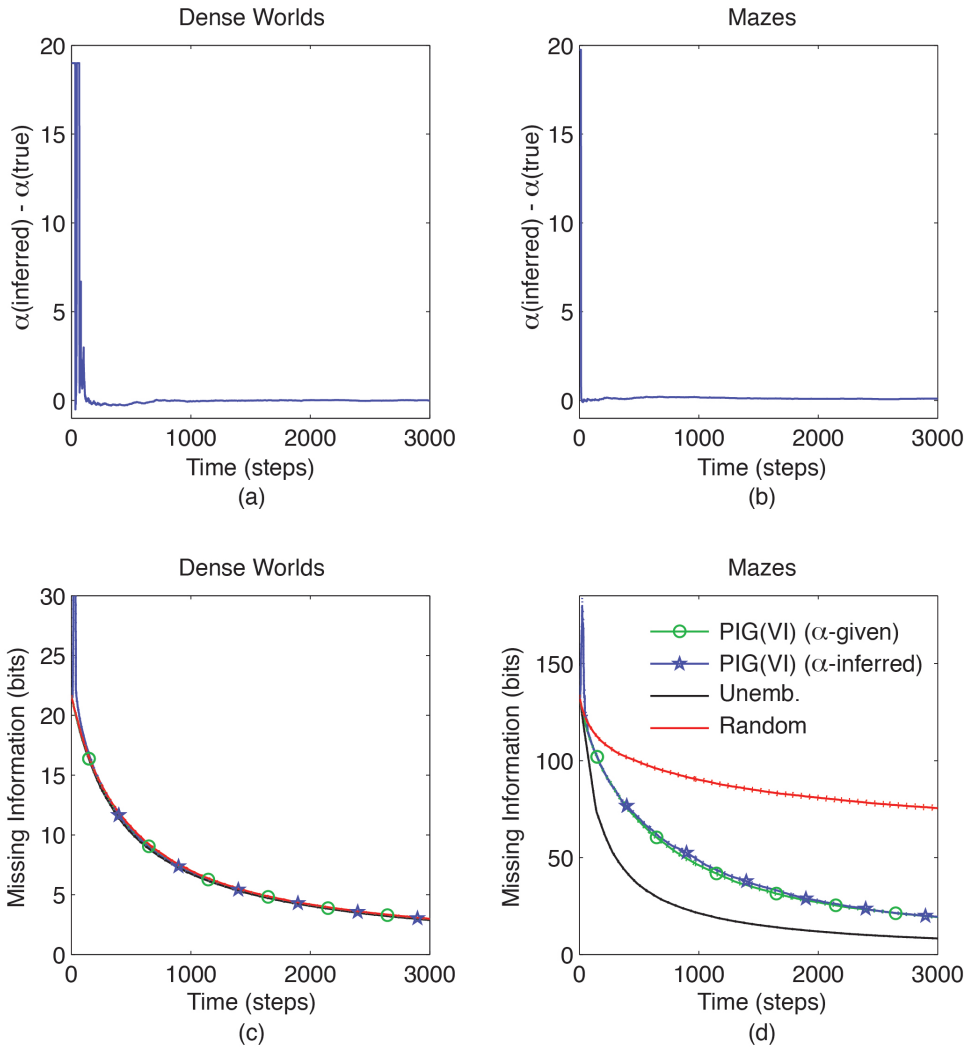

Figure 5.1: Inferring the concentration parameter during learning. (a,b) The mean error in inferred concentration parameter over time is plotted for Dense Worlds and Mazes. (c,d) The missing information over time is plotted for a PIG(VI) explorer updating its internal model using the true (green with circles) or inferred (purple with stars) concentration factor. Standard control explorers (with α given) have been included. Dotted lines above and below learning curves depict standard errors.

as accurate as those learned by explorers given accurate priors (Fig. 2.1 c and d). Thus, even when agent's were required to infer the concentration parameter, the PIG(VI) explorer was still able to quickly learn an accurate internal model.

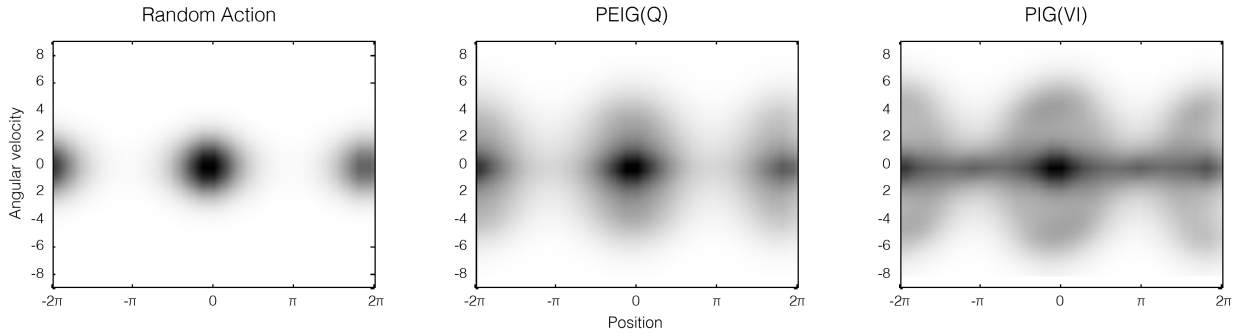

Figure 5.2: Pendulum exploration. Graphs depict the distribution of states visited during exploration of the continuous dynamics for a pendulum. The employed exploration strategy is noted above each graph. Darker areas represent more frequently visited states. Three angles (-2π, 0, and 2π) all correspond to the same downward position. Each graph is thus showing in repetition two cycles through the angles.

5.2 Exploring continuous dynamics

The discrete structure of CMCs may pose a significant limitation on the types of dynamics that can be simulated in this framework. While discrete dynamic systems have been greatly studied, many real world problem that an individual may wish to explore are more appropriately described by continuous dynamics. Having an exploring agent model its continuous world as a discrete one would represents a significant departure from the accurate priors considered in the previous experiments. I wanted to test whether the PIG(VI) strategy could be used to explore a continuous system despite these potential limitations of the discrete framework and chose the idealized pendulum as a test case. The state of the idealized pendulum is given by an angle θ and angular velocity ω. Actions correspond to applying torque τ in either direction on the pendulum. The pendulum's continuous dynamics are given by the following system of differential equations:

$$\frac{d\theta}{dt} = \omega$$
$$\frac{d\omega}{dt} = \frac{\tau}{I} - \frac{g}{L}\sin\theta - \frac{c}{Lm}\omega$$

Where I is the moment of inertia, g is the standard gravity, L the length of the pendulum arm, c the friction constant, and m the mass. While exploring this continuous system, I had my agents construct an internal model of the dynamics as a CMC. The set of states for the system were defined by discretizing the angle and angular velocity into a total of 288 states. The agent was allowed at discrete time-points to apply torque at preassigned strengths (either strong, weak, or none) in either the positive or negative directions making for a total of 5 actions. A small amount of Gaussian noise was added to the torque applied at each step. While the agent was modeling the system as a CMC, the actual dynamics, and thus the sensory feedback it received, were governed by the continuous dynamics. Thus the agent's prior for the system, for which I again used a Dirichlet distribution,

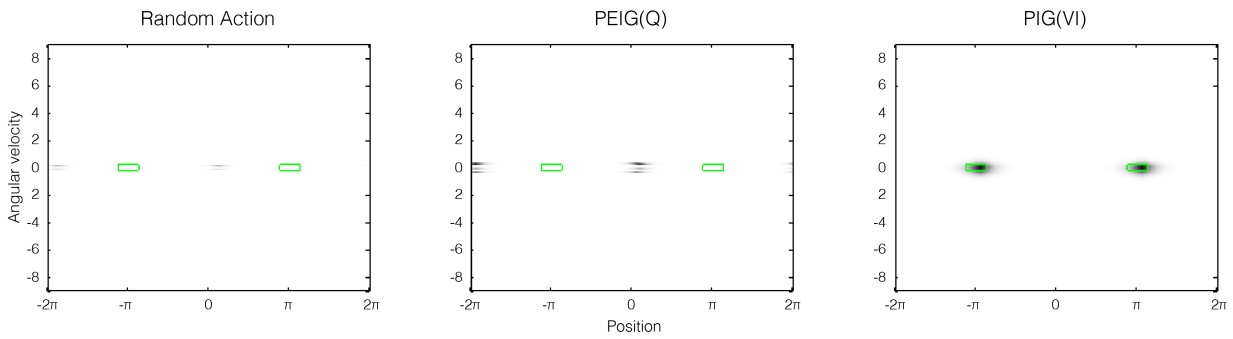

Figure 5.3: Reaching the inverted state. Graphs depict the distribution of states visited during the goal-directed task of reaching the inverted state at $\pm\pi$. Goal-directed action policies were derived from internal models learned under the exploration strategy noted above each graph. Darker areas represent more frequently visited states. The green box denotes the target. Again, each graph shows two cycles through the angles.

very poorly described the system. I tested three strategies under this paradigm: random action as a naive control and prominent strategy from reinforcement learning, PIG(VI) as my candidate strategy, and PEIG(Q) as the most closely related competitor develop in reinforcement learning. Since there is no true CMC describing this system I could not calculate the missing information of the explorers. I could however look qualitatively at the resultant behavior of exploration. The density plot in Fig. 5.2 depicts the distribution of states covered by the different strategies during exploration of an idealized pendulum. Notice that 0 radians (corresponding to a hanging position) represents an absorbing state of the system, while π radians (corresponding to an upright pendulum and equivalent to $-\pi$ radians) represents a critical state. Of the tested strategies, only PIG(VI) seemed to effectively reach and explore the states of this upright position.

While I could not quantify the missing information of the system, I could ask about the utility of the learned internal model. Following the methods described in Section 4.1, after exploration, I tasked the agents to apply their internal models to design a strategy for reaching the inverted states at $\theta = \pi$. The density plot in Fig. 5.3 shows the distribution of states reached under this goal-directed strategy. The green rectangles denote the target states. Again, of the different strategies tested, PIG(VI) was the only one whose internal model was a sufficient approximation of the continuous dynamics to allow the agent to reach the inverted state. Altogether, these results show that the CMC framework, despite its limitation, is capable of modeling exploration in even complex dynamics. It also further demonstrates the utility of the PIG(VI) algorithm for exploration.

Chapter 6

Discussion

In this thesis I introduced a parsimonious mathematical framework for studying learning-driven exploration by embodied agents based on information theory, Bayesian inference, and controllable Markov chains (CMCs). I compared agents that utilized different exploration strategies towards optimizing learning. To understand how learning performance depends on the structure of the world, three classes of environments were considered that challenge the learning agents in different ways. I found that fast learning could be achieved in all environments by an exploration strategy that coordinated actions towards long-term maximization of predicted information gain, PIG(VI).

6.1 Caveats

A potential limitation of my approach is that the VI algorithm is only optimal if the utility function is stationary (i.e. unchanging) [13]. Any utility function, including PIG, that attempts to capture learning progress will necessarily change over time. This caveat may be partially alleviated by the fact that PIG changes only for the sampled distributions. Furthermore, PIG decreases in a monotonic fashion (see Fig. 3.1) which can potentially be captured by the discount factor of VI. Interesting future work may lie in accounting for such monotonic decreases in estimates of future information gains either through direct estimation or through improved approximations perhaps by a guided choice of discounting mechanism. The problem of accounting for diminishing returns on utility has been previously approached in the field of optimal foraging theory. Modeling the foraging behaviors of animals, optimal foraging theory considers an animals decision of when it should leave its present feeding area, or patch, in which it has been consuming the available food and expend energy to seek out a new, undiminished patch [72]. Charnov's Marginal Value Theorem, a pivotal finding in the field, suggests that the decision to transition should be made once the expected utility of the current patch decreases to the average expected utility across all patches accounting for transition costs [23]. Extending this work to my information theoretic approach in CMCs may provide the necessary insights to address the challenges of diminishing returns on information gain.

Furthermore, the VI algorithm scales linearly with the size of the state space, and the calculation of PIG can scale linearly with the square of the size of the state space. This means as we

consider larger CMCs, these approaches will become more computationally expensive to perform. For large worlds, clever methods for approximating these approaches or for sparsifying their representation may be necessary. An explicit model of memory may also be necessary to fully capture the limitation on computational complexity biological organisms face. A wealth of literature from machine learning and related fields may offer insights in approaching these challenge which I reserve for future work.

6.2 Related work in Reinforcement Learning

CMCs are closely related to Markov Decision Processes (MDPs) commonly studied in Reinforcement Learning. MDPs differ from CMCs in that they explicitly include a stationary reward function associated with each transition [38, 125]. RL research of exploration usually focuses on its role in balancing exploitative behaviors during reward maximization. Several approaches for inducing exploratory behavior in RL agents have been developed. One very common approach is the use of heuristic strategies such as random action, least taken action, and counter-based algorithms. While such strategies may be useful in gathering unchanging external rewards, my results show that they are inefficient for learning the dynamics of structured worlds.

Other RL approaches involve reward-driven exploration. In the absence of external rewards, exploration could still be induced under reward-driven strategies by having the agent work through a series of internally chosen reward problems. This is essentially how the described BOSS agent operates. It was nevertheless insufficient to reach the performance accomplished by PIG(VI).

In addition, several RL studies have investigated intrinsically motivated learning. For example, Singh *et al.* [119] have demonstrated that RL guided by saliency, an intrinsic motivation derived from changes in stimulus intensity, can promote the learning of reusable skills. As described in Section 3.6, Storck *et al.* introduced the combination of Q-learning and PEIG as an intrinsic motivator of learning [123]. In their study, PEIG(Q) outperformed random action only over long time scales. At shorter time scales, random action performed better. Interestingly, I found exactly the same trend, initially slow learning with eventual catching-up, when I applied PEIG(Q) to exploration in my test environments (Fig. 6). One might consider the approach taken in this thesis to be one of modeling exploration as intrinsically motivated by learning. The significant departure from reinforcement learning in the present work lies in changing the perspective on exploration from one where reward is the immediate objective to one where information is. A fundamental challenge to extending reinforcement learning to the learning-driven perspective lies in the inherent way information values change as an agent gathers observations. Barto and Sutton noted in there introduction to reinforcement learning that "the reward function must necessarily be unalterable by the agent" [125]. This requirement will need to be addressed if information gain is to be mapped into the reinforcement learning framework.

6.3 Between learning-driven and reward-driven exploration

While curiosity, as a value for learning, is believed to be the primary drive of explorative behaviors, other factors, including external rewards, may play a role either in motivating exploration directly

or in shaping the development of curiosity [4,70,99,116]. In this manuscript, I wished to focus on a pure learning-based exploration strategy and therefore chose to take an unweighted sum of missing information as a parsimonious objective function (2.3). Two points, however, should be noted in considering the extension of this work to previous work in the literature. First, my objective function considers only the learning of the transition dynamics governing a CMC as this fully describes such a world. If we incorporate additional features into this framework, such as rewards in MDPs, those features too could be learned and assessed under the missing information objective function. Towards this goal, interesting insights may come from comparing my work with the multi-armed bandits literature. Multi-armed bandits are a special class of single state MDPs [40]. By considering only a single state, multi-armed bandits remove the embodiment constraint of multi-state CMCs and MDPs. Thus, CMCs and multi-armed bandits represent complimentary special cases of MDPs. That is, a CMC is an MDP without reward structure, while a multi-armed bandit is an MDP without transition kernels. Recent research has attempted to decouple the exploration and exploitation components of optimal control in multi-armed bandits [2, 19]. These studies aim at minimizing, through exploration, a construct termed regret, the expected reward forgone by a recommended strategy. Regret is similar to the navigational and reward acquisition loss values I calculated for ranking the explorers under goal-directed tasks. Importantly, while my work considered a wide array of goal-directed tasks, these multi-armed bandit approaches typically consider only learning a single fixed reward structure. Understanding these differences will be important if one wishes to shift attention from the unbiased information theoretic view I take to a directed task-dependent view. Identifying a means, perhaps through information theory, of quantifying the uncertainty regarding which strategy will optimize a task, will be an important extension bridging these two approaches.

The idea of directed information brings us to our second consideration in extending this work to previous literature. Psychologists have found that curiosity, or interest, can vary greatly both between and within individuals [115,117]. While one should be careful to not conflate the valuation of an extrinsic reward with the emotion of interest, it is possible such valuations could act to influence the development of interests. By transitioning away from the non-selective measure of missing information towards a weighted objective function that values certain information over others, we may begin to bridge the learning-driven and reward-driven approaches to exploration. One interesting proposal, put forth by Vergassola *et al.* suggests that information regarding a reward often falls off with distance as an organism moves away from the source of the reward [134]. Accordingly, a greedy local maximization of information regarding the reward may simultaneously bring the individual closer to the desired reward. The resultant "infotaxis" strategy is closely related to the PIG(greedy) strategy but is applied only to the single question of where a particular reward is located.

6.4 Related work in Psychology

In the Psychology literature, PIG, as well as PMC and PLC, were directly introduced as measures of the expected difference between a current and future belief [10, 63, 83, 87]. Here, I showed that PIG equals the expected change in missing information (Theorem 2). Analogous theorems do not

hold for PMC or PLC. For example, PLC is not equivalent to the expected change in L1 distance with respect to the true world. This might explain why PIG(VI) outperformed PLC(VI) even under an L1 measure of learning.

In this thesis, I applied PIG, PMC, and PLC to the problem of learning a full model of the world. In contrast, the mentioned psychology studies focused specifically on hypothesis testing and did not consider sequences of actions or embodied action-perception loops. These studies revealed that human behavior during hypothesis testing can be modeled as maximizing PIG, suggesting that PIG may have biological significance [83, 87]. However, their results could not distinguish between the different utility functions (PIG, PMC and PLC) [83]. The finding that 1-2-3 Worlds give rise to large differences between the three utility functions may help identify new behavioral tasks for disambiguating the role of these measures in human behavior.

To model bottom-up visual saliency and predict gaze attention, Itti and Baldi recently developed an information theoretic measure closely related to PEIG [8,52,53]. In this model, a Bayesian learner maintains a probabilistic belief structure over the low-level features of a video. Attention is believed to be attracted to locations in the visual scene that exhibit high Surprise. Several potential extensions of this work are suggested by my results. First, it may be useful to model the active nature of data acquisition during visual scene analysis. In Itti and Baldi's model, all features are updated at all points in the visual scene regardless of current gaze location or gaze trajectory. Differences in acuity between the fovea and periphery however suggest that gaze location will have a significant effect on which low-level features can be transmitted by the retina [137]. Second, my comparison between PIG and PEIG (Fig. 6) suggests that predicting future changes may be more efficient than focusing attention only on those locations where change has occurred in the past. A model that anticipates Surprise, as PIG anticipates information gain, may be better able to explain some aspects of human attention. For example, if a moving object disappears behind an obstruction, viewers may anticipate the reemergence of the object and attend that location. Finally, a subtle difference between our two approaches lies in the choice of random variable for which we measure information gain. Recall that in my framework, missing information is calculated for the transition kernel describing the world. At the same time, this transition kernel itself is considered by the agent to be a random variable with a distribution updated from the prior according to the data. If one takes the true distribution over the transition kernel to be a delta function, one could attempt to assess missing information in the context of this hyper distribution rather than at the descriptive level of the transition kernels themselves. That is, instead of considering the information measures of $\widehat{\Theta}$ and Θ one could consider the same measures applied to $f(\Theta|_{\cdot})$ and $\delta_{\Theta}(\Theta)$. The Surprise construct of Itti and Baldi uses this latter approach [8]. In my framework, exploration guided by predicted information gain with respect to the prior distributions is less efficient at exploration and offers less generalized utility when compared to exploration guided by PIG with respect to the descriptive distributions (data not shown). Incorporating these insights into new models of visual saliency and attention could be an interesting course of future research.

6.5 Information-theoretic models of behavior

Recently information-theoretic concepts have become more popular in computational models of behavior. These approaches can be grouped under three guiding principles. The first principle uses information theory to quantify the complexity of a behavioral policy, with high complexity considered undesirable. Tishby and Polani for example, considered RL maximization of rewards under such complexity constraints [131].

The second principle is to maximize a measure called *p*redictive information which quantifies the amount of information a known (or past) variable contains regarding an unknown (or future) variable [6, 121, 130]. Predictive information has also been referred to as *e*xcess entropy [28] and should not be confused with predicted information gain (PIG). When the controls of a simulated robot were adjusted such that the predictive information between successive sensory inputs was maximized, Ay *et al.* found that the robot began to exhibit complex and interesting explorative behaviors [6]. This objective selects for behaviors that cause the sensory inputs to change often but to remain predictable from previous inputs, and we can therefore describe the resulting exploration as stimulation-driven. Stimulation-driven exploration generally benefits from a good internal model but on its own, does not drive fast learning. It is therefore more suitable later in exploration, after a learning-driven strategy, such as PIG(VI), has had a chance to form an accurate model. PIG, in contrast, is most useful in the early stages when the internal model is still deficient. These complimentary properties of predictive information and PIG lead us to hypothesize that a simple additive combination of the two objectives may naturally lead to a smooth transitioning from learning-driven exploration to stimulation-driven exploration, a transition that may indeed be present in human behavior (see Section 6.6).

Epsilon machines introduced by Crutchfield [27] and the information bottleneck approach introduced by Tishby *et al.* [130] combine these first two principles of maximizing predictive information and constraining complexity. In particular maximizing the information between a compressed internal variable and the future state progression subject to a constraint on the complexity of generating the internal variable from sensory inputs. Recently, Still extended the information bottleneck method to incorporate actions [121].

Finally, the third information-theoretic principle of behavior is the minimization of free-energy, an information-theoretic bound on surprise. Friston put forth this Free-Energy (FE) hypothesis as a unified variational principle for governing both the inference of an internal model and the control of actions [37]. Under this principle, agents should act to minimize the number of states they visit. This stands in stark contrast to both learning-driven and stimulation-driven exploration. A learning-driven explorer will seek out novel states where missing information is high, while a stimulation-driven explorer actively seeks to maintain high variation in its sensory inputs. Still, reduced state entropy may be valuable in dangerous environments where few states permit survival. The balance between cautionary and exploratory behaviors would be an interesting topic for future research.

6.6 Towards a general theory of exploration

With the work of Berlyne [16], psychologists began to dissect the different motivations that drive exploration. A distinction between play (or diversive exploration) and investigation (or specific exploration) grew out of two competing theories of exploration. As reviewed by Hutt [50], "curiosity"-theory proposed that exploration is a consummatory response to curiosity-inducing stimuli [14, 81]. In contrast, "boredom"-theory held that exploration was an instrumental response for stimulus change [41, 82]. Hutt suggested that the two theories may be capturing distinct behavioral modes, with "curiosity"-theory underlying investigatory exploration and "boredom"-theory underlying play. In children, exploration often occurs in two stages, inspection to understand what is perceived, followed by play to maintain changing stimulation [51]. These distinctions nicely correspond to the differences between my approach and the predictive information approach of Ay *et al.* [6] and Still [121]. In particular, I hypothesize that my approach corresponds to curiosity-driven investigation, while predictive information a la Ay *et al.* and Still may correspond with play. Furthermore, the proposed method of additively combining these two principles (Section 4.4), may naturally capture the transition between investigation and play seen in children.

For curiosity-driven exploration, there are many varied theories [70]. Early theories viewed curiosity as a drive to maintain a specific level of arousal. These were followed by theories interpreting curiosity as a response to intermediate levels of incongruence between expectations and perceptions, and later by theories interpreting curiosity as a motivation to master one's environment. Loewenstein developed an Information Gap Theory and suggested that curiosity is an aversive reaction to missing information [70]. More recently, Silvia proposed that curiosity is motivated by two traits, complexity and comprehensibility [116]. For Silvia complexity is broadly defined, and includes novelty, ambiguity, obscurity, mystery, etc. Comprehensibility is simply an appraisal of how well something can be understood. It is interesting how well these two traits match information-theoretic concepts, complexity being captured by entropy, and comprehensibility by information gain [94]. Indeed, predicted information gain might be able to explain the dual aspects of curiosity-driven exploration proposed by Silvia. PIG is bounded by entropy and thus high values require high complexity. At the same time, PIG equals the expected decrease in missing information and thus may be equivalent to expected comprehensibility.

All told, these results add to a bigger picture of exploration in which the theories for its different aspects fit together like pieces of a puzzle. This invites future work for integrating these pieces into a more comprehensive theory of exploration and ultimately of autonomous behavior.

6.7 Conclusion

The fundamentally new perspective I took in the thesis allowed me to re-examine the computational principles of exploration free of the prior assumptions propounded by reinforcement learning. By separating considerations of the proximate and ultimate causes of exploration, I've introduced new methods for evaluating the effectiveness and utility of learning-driven exploration. I believe these new performance benchmarks represent in themselves a significant contribution to the field. Furthermore, they have allowed me to develop a novel and effective new strategy for exploration,

PIG(VI). The question of how to direct behavior during embodied learning towards the formation of accurate internal models and the development of adaptable skills I believe will be central to future work in machine learning, neuroscience, and robotics.

Part II

Using information theory to identify coevolving protein residues

Chapter 7

Introduction

A complete understanding of protein evolution will require full characterization of the many factors that determine the selective forces acting on each amino acid of a protein. Although it has long been hypothesized that the residues within a protein interact and influence each other's evolution, models of protein evolution, for simplicity and lack of sufficient data, have traditionally assumed that residues evolve independently of each other. However, the increasing power of bioinformatics and the increasing availability of genomic data offer a new opportunity to search for specific signals of coevolution.

The covarion (concomitantly variable codon) hypothesis, put forth by Fitch and Markowitz [35], postulated that, at any point during the evolution of a protein, only a small fraction of its residues are free to vary. As the freely varying sites mutate, however, interacting sites can switch between variable and invariant states. While Fitch and Markowitz emphasized this binary switching, they acknowledged that more subtle changes in selective pressures might occur. For example, in response to a mutation at a neighboring site, a residue might switch from varying among one set of amino acids to varying among another set. To encompass this broader conceptualization of coevolution, the covarion hypothesis can be restated as: at any point during the evolution of a protein, only a small fraction of possible mutations are admissible, but as one site changes, it can alter the selective forces associated with other sites, thus altering the set of mutations that are selectively admissible at those site. This form of coevolutionary interaction could be recognized within a protein as residue pairs in which the variability at one site is dependent upon the amino acid state of the other.

Mutual information (*MI*) is a statistical measure of the codependency between two random variables. By considering the final amino acid states of a protein's residues, after a span of evolution, as discrete random variables, *MI* becomes a natural method for detecting codependencies between them. Using multiple sequence alignments (MSAs) to estimate the amino acid distribution at each site, *MI* quantifies how much uncertainty in the amino acid state at one site can be removed by knowledge of the amino acid state at another site.

The application of *MI* to sequence alignments was first introduced by Korber *et al.* as a means of identifying covarying sites in a viral peptide [66]. This approach was later extended to general proteins as a measure of coevolution [39]. Without refinement, however, *MI* yielded limited suc-

cess and several attempts have been made to improve the measure [32,42,46,129,139]. Wollenberg and Atchely, for example, used parametric bootstrap simulations to model the effect of phylogenetic relationships on *MI* in the absence of coevolution [139]. Their approach, however, could not separate this global phylogenetic influence from the specific coevolutionary signal between a pair of sites [139]. Tillier and Lui attempted to capture biases acting on each site of a protein through an analysis of the total amount of interdependencies each site had across all other sites [129]. They, however, did not characterize the correlation between *MI* and their measure of this bias. Their method of removing this bias from *MI* may, therefore, have been suboptimal and may have hindered the accuracy of their algorithm. These and the other researchers have emphasized the need to quantify and effectively remove the poorly understood biases that are hindering the efficacy of *MI* as a measure of coevolution [32, 42, 46, 76, 129, 139].

Since the true coevolutionary history of a protein cannot be experimentally determined, measures of coevolution cannot currently be directly tested. This complicates the validation of any measure and necessitates the use of indirect evidence. A correlation between predicted coevolving residue-pairs and protein structure is the most common evidence offered to support the accuracy of an algorithm [32, 36, 42, 44, 46, 60, 66, 88, 91, 129, 132, 135, 139, 140]. Indeed, many researchers who develop algorithms for quantifying covariability between sites abandon coevolution as their primary goal and instead focus on the algorithm's potential as a tool for structure prediction, in particular contact prediction [33, 44, 88, 112]. Still, the correlation that these algorithms yield with protein structure is likely mediated by their capacity to accurately measure coevolution combined with an inherent tendency for physically close residues to interact evolutionarily.

Demonstrating that a measure's predicted coevolving residues are further correlated to additional relevant protein features aside from structure can, by an argument of parsimony, greatly increase the support for that measure as it limits the range of potential non-coevolutionary explanations. Towards this end, researchers occasionally offer examples of coevolving residues that they consider to be functionally relevant or near functionally relevant sites [42,132,135,140]. Such correlations should, however, be evaluated carefully and with consideration of two factors. First, site-specific biases, such as conservation, may artificially conflate the coevolutionary measure of functionally relevant residues. Second, the appropriate controls are rarely given to demonstrate that the highlighted examples represent a true trend. Once a correlation is shown to be statistically significant and not the result of artefactual biases, it not only supports the accuracy of a measure but also provides insight into the nature of coevolution.

In this thesis, I offer a refinement of *MI* as a measure of coevolution that removes a strong non-coevolutionary influence and accounts for differences in within-site variability. I demonstrate a high correlation between the predicted coevolving residues and protein structure, which even extends to quaternary structures. I also demonstrate a significant trend for those residues that are annotated as participating directly in a protein's catalytic activity to coevolve with each other. Going beyond these two more commonly considered correlations, I offer a novel measure of the propensity for each pair of the 20 amino acids to be found at coevolving sites, which I term their coevolution potentials. I found that amino acid pairs known to interact in bond formation exhibited the strongest coevolution potentials, providing a unique correlation for my measure with the known biochemistry of proteins that had not previously been explored. I concluded by demonstrat-

ing directly that my measure surpasses previous methods in its degree of structural correlation, a standard comparison for evaluating measures of coevolution [32, 36, 141]. Work in this part of the thesis has been published in *PLoS ONE* [68].

Chapter 8

Developing an information theoretic measure of coevolution

8.1 Multiple sequence alignments

All protein alignments were obtained from the PFAM database (http://pfam.sanger.ac.uk/, Pfam 21.0) [34]. In total, 1592 PFAM full alignments were utilized. These full alignments were chosen based only on the criteria that they contained at least 500 sequences and at least two sites with fewer than 20% gaps. 1240 of these alignments had solved crystal structures available from the Protein Data Bank (http://www.pdb.org/) [17, 18].

8.2 Mutual information as a biased measure of coevolution

To develop a statistical framework for measuring coevolution, I began by modeling the propensity for each amino acid to evolve at a site in a protein as a discrete random variable with 20 possible outcomes representing the 20 amino acids. To look for interdependencies between two sites (i.e. two random variables), I considered the mutual information, MI, between them. MI is a statistical quantity that measures the codependency of two random variables by examining how much less entropy (i.e. more order) there is in their joint distribution than would be expected if the two distributions were completely independent. If the propensity for a particular amino acid to evolve at one site is completely independent from the amino acid state of the other site, mutual information would be zero. If, however, the propensity for a particular amino acid to evolve at one site is completely determined by the amino acid state at the other site, then the two single distributions and their joint distribution will have entropy equal to the mutual information.

Given a multiple sequence alignment (MSA), let p_i be the vector of length 20 whose entries are the frequencies of the 20 amino acids amongst all the sequences at position i ignoring gaps. Next let $p_{i,j}$ be the 20-by-20 matrix whose entries are the joint distribution of each ordered amino acid pair at positions i and j. Entropy, H_i, is a measure of the uncertainty associated with p_i and is

given by:

$$H_i = -\sum_{x \in \mathscr{A}} p_i(x) \log_2 p_i(x)$$

Here, $\mathscr{A} = \{A, C, D, \ldots, Y\}$ is the set of the 20 amino acids. H_i has a minimum value of 0, i.e. no uncertainty, when all sequences in the MSA have the same amino acid at position i, and it increases as the amino acid frequencies become more evenly distributed with a maximal value when all 20 amino acids are equally represented. The joint entropy, $H_{i,j}$, between p_i and p_j is simply the entropy of the joint distribution $p_{i,j}$ and is given by:

$$H_{i,j} = -\sum_{x,y \in \mathscr{A}} p_{i,j}(x,y) \log_2 p_{i,j}(x,y)$$

If p_i and p_j are completely independent, then $H_{i,j} = H_i + H_j$. As p_i and p_j become more codependent, $H_{i,j}$ decreases and is minimized when the amino acid at i completely determines what amino acid must occur at j.

Finally, the mutual information of p_i and p_j, $MI_{i,j}$, is a statistical quantification of the interdependency between them and is given by:

$$MI_{i,j} = H_i + H_j - H_{i,j}$$

MI can be interpreted as the increase, due to codependency, in the certainty of the joint outcome over the expected certainty assuming complete independence. Gaps in an MSA can bias which phylogenies are represented at a site and decrease the sample size for estimating frequencies. For this reason, if any pair of sites had more than 20% of the sequences in the MSA gapped at either positions then no MI score, nor any of the derived measures of coevolution, were calculated for that pair. Such gapped pairs were thus left untested for any coevolutionary relationship.

Due to the phylogenetic relationships between the sequences of an MSA, the assumption that these sequences evolved independently from one another is false. Indeed, for any pair of proteins in an MSA, there will have been a most recent common ancestor. While the mutations that became stabilized in the lineages of these proteins following the branching of this common ancestor represent independent evolutionary events, those mutations that stabilized prior to this branch point would only be a single evolutionary event even though they would be treated by the *MI* analysis as independent events. This treatment of a single event as multiple independent events gives rise to a phylogenetic bias that increases the mutual information among the residues. By independently mixing the amino acids at each site among the sequences of an MSA, we can calculate random mutual information (*RI*) scores in which all coevolutionary signals and phylogenetic biases have been removed. As an example, I plotted the *MI* scores for each pair of amino acid sites in the PFAM full alignment of 5612 PDZ domains against their average *RI* scores from 300 randomizations (Figure 8.1A; Pfam ID: PF00595 [34]). The PDZ domain is commonly found in scaffolding proteins where it serves as a binding site for specific peptide sequences in target proteins. 80-90 amino acids in length, its small size makes it amenable towards easily visualizing the coevolutionary pairs identified by my algorithm. One would expect most residues to have strong interactions

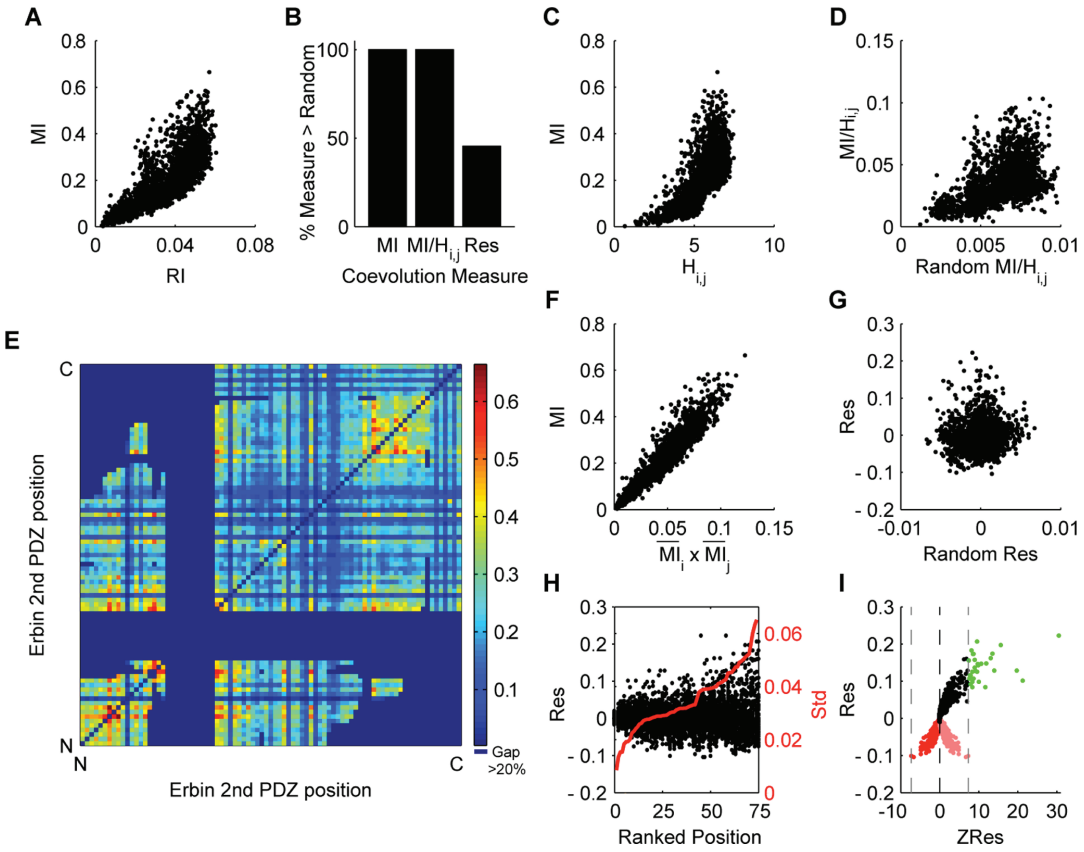

Figure 8.1: Measuring coevolution without biases. (A) *MI* scores are correlated to random information scores (*RI*) in which all coevolutionary and phylogenetic relationships have been removed by random perturbations. (B) The percentage of tested residue pairs that have coevolution measures higher than their average random measure (standard deviations are plotted but are too small to be visualized). Phylogenetic biases induce high *MI* and $MI/H_{i,j}$ scores, which are unobtainable from randomized results. (C) *MI* is correlated to $H_{i,j}$. (D) $MI/H_{i,j}$ is correlated to its randomized values. $MI/H_{i,j}$ is therefore still subject to non-phylogenetic biases. (E) A colorimetric representation of *MI* scores between pairs of residues in the 2nd PDZ domain of the Human Erbin protein. The striated appearance highlights a large variation in basal *MI* values between sites. Residue positions are aligned from the N-terminus to the C-terminus. Red = high *MI*, Blue = low *MI*, Darkest Blue = untested ($> 20\%$ gaps). (F) *MI* is correlated to $\overline{MI}_i \cdot \overline{MI}_j$. (G) *Res* is not correlated with its randomized values. (H) Positions are ranked in order of increasing variance in *Res* scores (red line indicates deviation of *Res* scores) and the distribution of *Res* scores are plotted. (I) *ZRes* scores are calculated as the product of the z-scores of a *Res* value relative to its distribution across each site. Light red points represent residue pairs where both z-scores were negative. The *ZRes* score for such sites are taken as the negative of the product of the z-scores (dark red points). The negative of the lower bound of *ZRes* (gray lines) is a cutoff for choosing coevolving residues (green points).

with only a few evolutionarily closely-coupled sites [129], but the *MI* scores were almost always higher than RI scores (Figure 8.1B; less than 1 residue pair out of all 2193 pairs per randomization). This suggests that the high mutual information scores obtained were likely a result of phylogenetic relationships within the MSA and not true signals of coevolution. Furthermore, *MI* proved to be significantly correlated to *RI* (R = 0.7892) despite having removed the coevolutionary and phylogenetic interactions. This suggests that MI is further subject to additional non-phylogenetic biases, which I collectively termed the stochastic bias.

8.3 Derived coevolutionary measures

To remove the biases associated with *MI*, I first calculated the average *MI* for each position:

$$\overline{MI}_i = \frac{1}{n_i} \sum_{j \neq i} MI_{i,j}$$

Where n_i is the number of positions j for which an *MI* score was calculated between i and j. Plotting $MI_{i,j}$ against $\overline{MI}_i \cdot \overline{MI}_j$, I found a strong linear relationship (Figure 8.1F). Since each site would be expected to coevolve with only a few other sites, their average *MI* would not be expected to contain much coevolutionary signal. Yet this correlation persisted even when $MI_{i,j}$ or the top 5 *MI* values for each site were removed (data not shown). $\overline{MI}_i \cdot \overline{MI}_j$ is therefore a confounding variable which potentially contains the phylogenetic or stochastic biases of *MI*. To remove the influence of this non-coevolutionary variable from *MI*, I calculated the linear least squares regression of $MI_{i,j}$ against $\overline{MI}_i \cdot \overline{MI}_j$ and took the residual of each $MI_{i,j}$ over this line of best fit as a new measure of coevolution, $Res_{i,j}$.

As shown in Figure 8.1G, *Res* no longer correlated with randomized results (R=0.0863), suggesting that it successfully removed the stochastic bias. Furthermore, about 50% of all residue pairs exhibited random scores higher than the measured *Res* values suggesting that the phylogenetic biases driving up global coevolution signals may have also been attenuated (Figure 8.1B). I however noticed that the variation in the residuals still displayed heteroscedasity: increased variation with increasing MI (Figure 8.1F). To examine how differences in variation might be influencing the *Res* scores, I plotted the distribution of *Res* scores for each site, sorting the sites by increasing variance (Figure 8.1H). While average *Res* values tended to be similar across all sites, the variation at each site differed dramatically. A plot of the standard deviation in *Res* scores for each site against the entropy of that site revealed that the two are correlated, suggesting that sites with more variation in amino acid composition (i.e. more entropy) have an increased tendency to vary in their *Res* value ($R = 0.4516$, $p < 0.0001$; Supplemental Figure B.1). Without correction, more variable sites would have a wider distribution of *Res* values and thus an increased chance to surpass any chosen threshold. To adjust for these differences in variation, I compared the *Res* score for a particular pair of sites to the distribution of *Res* scores for each of those sites considered separately. Specifically, I calculated z-score, $Z_i(j)$, for $Res_{i,j}$ relative to the distribution of *Res* scores across partners of i,

$Res_{i,\cdot}$:

$$Z_i(j) = \frac{Res_{i,j} - \mu(Res_{i,\cdot})}{\sigma(Res_{i,\cdot})}$$

Here, $\mu(Res_i,\cdot)$ and $\sigma(Res_i,\cdot)$ represent the mean and standard deviation of $Res_{i,\cdot}$ respectively. Finally, to normalize for the variability at both i and j, I calculated a final measure, *ZRes*:

$$ZRes_{i,j} = \begin{cases} Z_i(j) \cdot Z_j(i) & \text{if } Z_i(j) > 0 \text{ or } Z_j(i) > 0 \\ -Z_i(j) \cdot Z_j(i) & \text{otherwise} \end{cases}$$

Thus *ZRes* is a normalized measure of the position of a site-pair's *Res* score relative to the distribution of *Res* scores for those sites. The split function was used to address the situation where both z-scores are negative. This was problematic since their multiplication would then become positive (Figure 8.1I, light-red). I therefore interpreted only position pairs where both $Z_i(j)$ and $Z_j(i)$ were positive as potentially coevolving. Since the z-scores distribute around zero, and since there can not be a negative coevolutionary interaction, I was able to derive a natural threshold for detecting coevolving sites. Specifically, by letting *ZLB* (*ZRes* lower bound) be the most negative value obtained by *ZRes*, $-ZLB$ was a natural cutoff threshold for selecting predicted coevolving residues:

$$ZLB_{i,j} = \min\{ZRes_{i,j} \text{ s.t. } Z_i(j) \leq 0 \text{ or } Z_j(i) \leq 0\}$$

Those site pairs whose *ZRes* value exceeded the $-ZLB$ cutoff were identified as coevolving (Figure 8.1I, green).

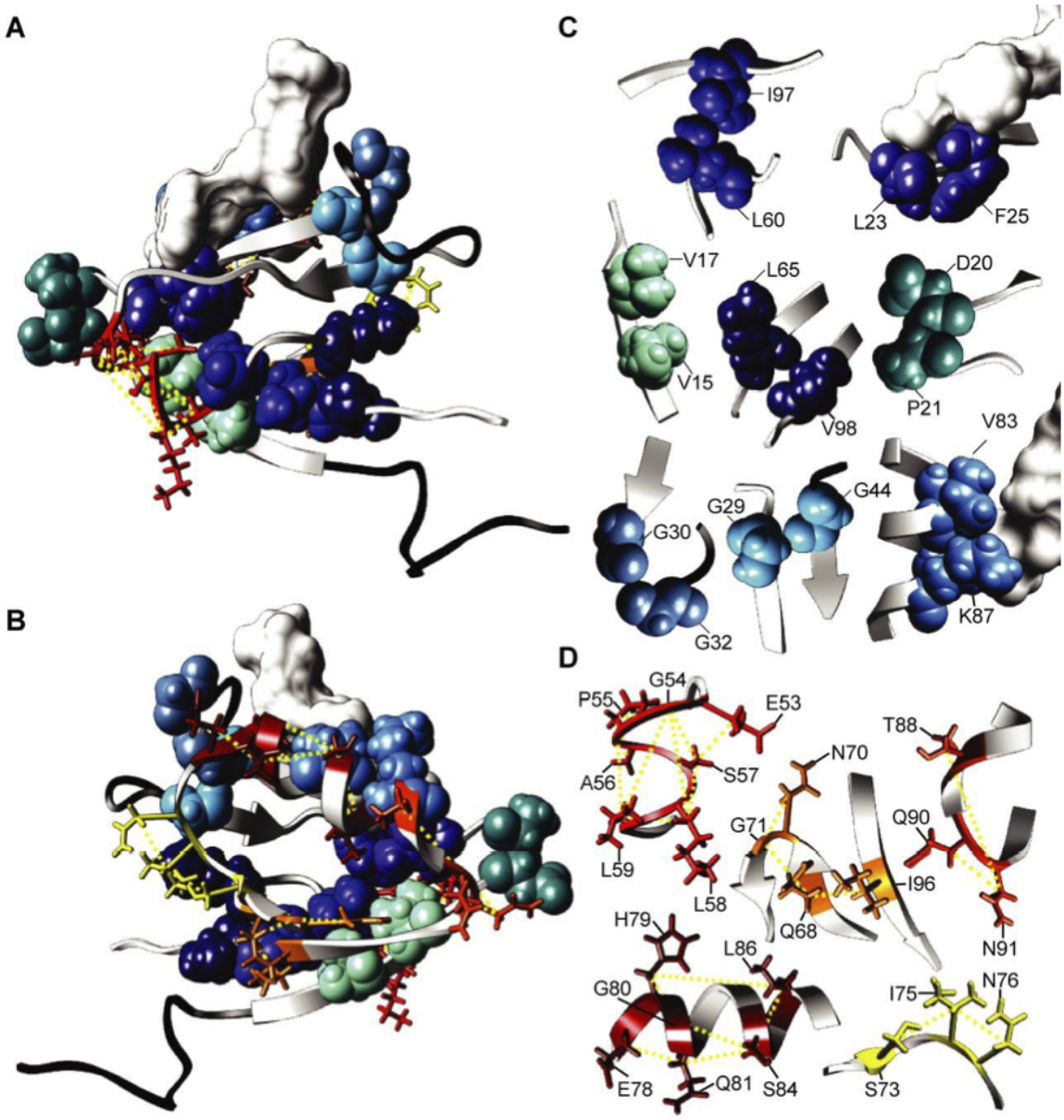

Figure 8.2: Coevolving residues in the 2nd PDZ domain of Human Erbin. (A) The structure of 2nd PDZ domain of Human Erbin with peptide ligand. Coevolving networks of at least 3 residues are depicted as balls-and-sticks in shades of red with dashed yellow lines connecting the coevolving pairs. Isolated pairs of coevolving residues are depicted as spheres in shades of blue. The molecular surface of the peptide ligand is depicted in white. Black ribbons represent untested residues ($>$ 20% gaps). (B) Backside of A. (C) Isolated pairs of coevolving residues. (D) Networks of 3 or more coevolving residues.

Chapter 9

Identified coevolving sites correlate with protein structure, biochemical interactions, and catalytic function

9.1 Identified coevolving sites in PDZ domains

The structure of the 2nd PDZ domain of the Human Erbin protein has been solved and shown to be similar in general topology to other PDZ representatives [120] (PDB ID: 1N7T [17, 18]). To examine spatial relationship between coevolving residues identified by my algorithm, I mapped all residue pairs with *ZRes* scores higher than the $-ZLB$ cutoff onto the structure of the Erbin 2nd PDZ domain (Figures 8.2A&B; visualizations done with UCSF Chimera [93]). Isolated pairs of residues that were identified as coevolving with each other and no other sites are depicted as space-filled spheres, each pair a different shade of blue (Figure 8.2C). Networks of three or more residues connected by coevolutionary interactions are depicted in ball-and-stick form with dashed yellow lines connecting the β carbons of the coevolving pairs (Figure 8.2D). In total I identified 30 coevolving pairs falling into 13 networks and involving 39 unique residues, nearly half of the tested residues.

The close physical proximity between each coevolving residue pair is quite striking. I plotted the distribution of distances between pairs of coevolving residues in comparison to the distribution for all pairs of tested residues (Figure 9.1) and found that the interacting residues were significantly closer together ($p < 1x10-16$, 2-sample Kolmogorov-Smirnov (K-S test); median distances: 2.88 $\AA$ (coevolving), 11.30 $\AA$ (all)). I interpret this strong correlation between my measure and physical structure as arising from the tendency for coevolving residues to be close to each other. These results suggest that the *ZRes* measure is indeed picking up a signal of coevolution. Interestingly, while many of the coevolving residues where found to lie in the same secondary structure (e.g. Val-83 and Lys-87 which align on one side of the only α-helix; Figure 8.2C), several examples were also found of residue interacting between secondary structures (e.g. Gln-68 and Ile-96 interacting between the 4th and 6th β-sheets; Figure 8.2D).

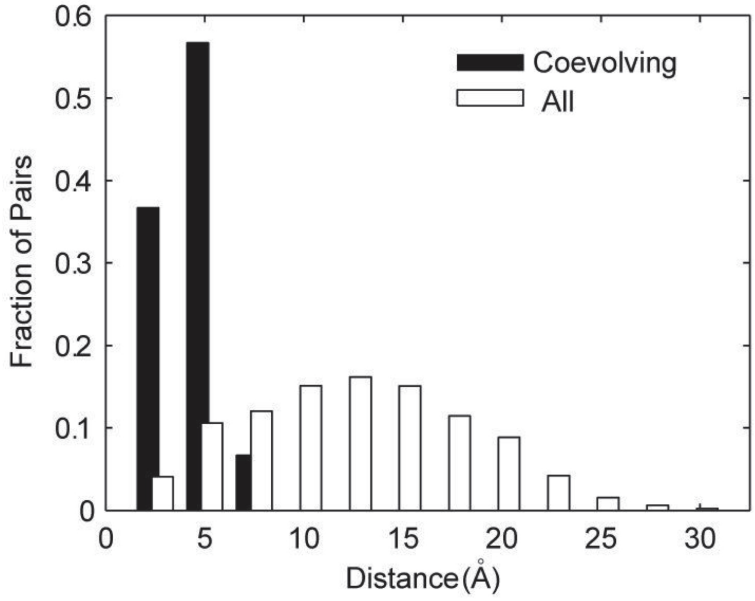

Figure 9.1: Distribution of distances between coevolving residues of PDZ domains. The fraction of coevolving (black bars) or all (white bars) residue pairs that lie within the specified interval of physical distance from each other is depicted.

9.2 Coevolution in 1592 Pfam families

The PFAM website (http://pfam.sanger.ac.uk/) maintains a database of alignments of well-characterized protein families and domains [34]. In order to test the generality of my results from the PDZ alignment across a larger set of proteins, I downloaded 1592 PFAM full alignments chosen based on the criteria that they contained at least 500 sequences and at least one pair of sites with less than 20% gaps.

Applying the *ZRes* algorithm to all 1592 alignments, I identified 126,085 coevolving residue pairs (out of 18,073,342) with *ZRes* scores above the $-ZLB$ cutoff. On average, 57.1% $\pm$ 19.6% of the tested residues for each protein family were identified as coevolving with at least one other residue. In comparison, the $MI/H_{i,j}$ measure developed by Gloor *et al.* yielded coverage of only 11.2% $\pm$ 7.5% (58,267 pairs; utilizing the 4 standard scores cutoff suggested by Gloor *et al.* [135]).

To test whether the identified coevolving residues correlated with physical structure, I obtained the structural data for representative members of 1240 of the 1592 PFAM alignments [17]. Figure 9.2A shows the distribution of the distances between the 86,084 identified coevolving residue pairs present in the representative structures. For comparison, the distribution of distances between all 12,203,471 tested pairs of residues in the 1240 crystal structures is also shown. Indeed, the coevolving residues were significantly closer together ($p < 1x10-307$, K-S test) with a median distance of 4.3$\AA$ as compared to a median of 19.2$\AA$ for all tested residue pairs. 56% of the identified coevolving residues were within 6 $\AA$ of each other, indicative of direct physical contact [42].

In comparison, only 7% of all tested residue pairs were in a similar range of contact. Furthermore, to test whether these results could have arisen from a bias in the *ZRes* measure towards selecting a specific set of sites that as a population tended to be close together, I examined the set of all sites identified as coevolving with at least one other site. The median distance between pairs of sites amongst this set (19.3 $\AA$) was no different than the total distribution for all tested pairs of sites nor was the percentage of site pairs in contact (7%). This demonstrates that the correlation between *ZRes* and physical structure is specifically dependent on the pairing of identified coevolving residues and not the result of single-site biases. I therefore interpret these results as emerging from the accuracy of my algorithm at identifying coevolving residues paired with the tendency for direct structural interactions to strongly influence residue coevolution.

To further explore correlations between coevolving residues and structural interactions, I next considered secondary structure. Of the 86,084 coevolving residue pairs, 14,653 (17.0%) were found to lie in a common α-helix or β-sheet. In comparison, only 3.8% of all residue pairs were identified as lying in a common α-helix or β-sheet, suggesting that residues interacting within a secondary structure have an increased tendency to influence each other's evolution. In the PDZ domain, coevolutionary interactions had tended to space out to align along the same side of the α-helix or β-sheets. To test the generality of this observation, I considered all coevolving pairs of residues where both residues lied in the same α-helix (Figure 9.2B) or the same β-sheet (Figure 9.2C) and determined their primary sequence separation. The results are given as a fraction of the total number of residue pairs that were located within a common secondary structure of the respective type and separated by the given primary distance. Residues within an α-helix exhibited a strong peak at 3 and 4 amino acids primary distance, coincident with the first turn of an α-helix (3.6 amino acids, first dashed line in Figure 9.2B). The propensity to coevolve quickly died off for primary distances past 4 amino acids, probably because subsequent helix turns become further and further away from each other in the molecular structure. Still a subtle peak can be seen every 3-4 amino acids consistent with the approximate 3.6 amino acids per turn characteristic of α-helices [26]. Even though the correlation for β-sheets was not as strong, it did exhibit a strong peak for residues that were separated by only a single amino acid (i.e. the closest residues to align on the same side of a β-sheet; Figure 9.2C).

I next tested whether coevolving residues that were distant in primary sequence were still close in tertiary structure. I therefore calculated the median physical distance between residues for a spectrum of minimum primary distance separations (Figure 9.2D). Even at a minimum of 30 amino acids primary distance, coevolving sites were significantly closer in physical distance (median: 9.8 $\AA$) than the total distribution of sites for that separation (median: 22.5 $\AA$; $p < 10^{-307}$, K-S test; Figure 9.2D). Similar statistical significance was obtained for all minimum primary distances from 1 to 30 ($p < 10^{-307}$, separate K-S tests for each minimum primary distance). For increasing minimum primary distance thresholds from 1 through 6, a moderate decrease in the difference between the median coevolving distances and the median for all sites was observed (Figure 9.2D, dashed line). This is perhaps due to the significance of secondary structural relationships in this range of primary sequence separation. Past a minimum primary distance of 6, however, the differences between the coevolving sites and all sites become constant suggesting that the tendency towards coevolution is indifferent to the degree of primary sequence separation beyond those separations

strongly correlated to interactions within a secondary structure.

Finally, I examined the influence of sequence length and alignment size on the accuracy of the *ZRes* algorithm. I approximated accuracy in identifying coevolving residues by accuracy in contact prediction (the percentage of identified coevolving residue pairs separated by at most 6 $\AA$). Across alignments, the total number of tested residue pairs that contacted each other scaled with the protein's effective sequence length (the square-root of the number of tested residue pairs; Figure B.2A). This led to a strong correlation between the percentage of tested residue pairs that were in contact and the reciprocal of effective sequence length (R = 0.8428; Figure B.2B). Thus, one might expect that the ability to preferentially identify those residue pairs in contact as coevolving over those not in contact would decrease with increases in effective sequence length. However, the robustness of my previous results led me to speculate that the use of the $-ZLB$ selection threshold potentially adjusted for this bias. Indeed, the contact accuracy for identified coevolving residue pairs was much less correlated to the reciprocal of effective sequence length than were the percentages of all tested residue pairs contacting (R = 0.1976; Figure B.2C), though there was still a slight overall gain in performance for shorter proteins. This suggests that the *ZRes* effectively compensated for the decreased representation of coevolving residue pairs (which should increase linearly with protein length) relative to the total number of tested residue pairs (which increased quadratically with protein length). Finally, I also found a subtle but significant positive correlation between the contact accuracy for identified coevolving residue pairs and the number of sequences in an alignment, suggesting that larger alignments yielded increased accuracy (R = 0.1003, $p < 0.001$; Figure B.2D). These correlations to contact prediction accuracy most likely reflect a corresponding correlation to coevolution prediction accuracy.

9.3 Coevolution potentials

Having applied the *ZRes* algorithm to a large set of proteins, I next wanted to search for possible trends in the amino acid compositions of coevolving sites. I therefore developed a measure of the propensity for strongly coevolving sites to be composed of each of the 210 possible pairings of the 20 amino acids, which I termed the coevolution potentials between the amino acids. For each pair of coevolving sites (with $ZRes \geq -ZLB$), I calculated the frequency of each amino acid pair amongst the sequences of the corresponding MSA. I then weighted these frequencies by the *ZRes* score between those sites. These weighted values were calculated for all coevolving pairs and then summed. To account for biases resulting from differences in the frequency of occurrence for each amino acid, I determined the statistically expected outcome for repeating this calculation using randomly selected residue pairs weighted by the *ZRes* values of the original coevolving pairs. My final coevolution potentials represent the standard score for the coevolving amino acid pairs relative to their expected values and variance under the random process (Figure 9.3A).

The 11 highest coevolution potentials (in decreasing order) were found to be between: Asp-Arg, Cys-Cys, Glu-Arg, Glu-Lys, Asp-Lys, His-His, Asp-His, His-Thr, His-Tyr, His-Glu, His-Ser (Table S1). The high coevolution potentials of the acid-base amino acid pairs (Asp-Arg, Glu-Arg, Glu-Lys, Asp-Lys) suggest that coevolutionary forces act to maintain balanced ionic charges or specific ionic interactions. Similarly, the series of pairings with histidine highlights the impor-

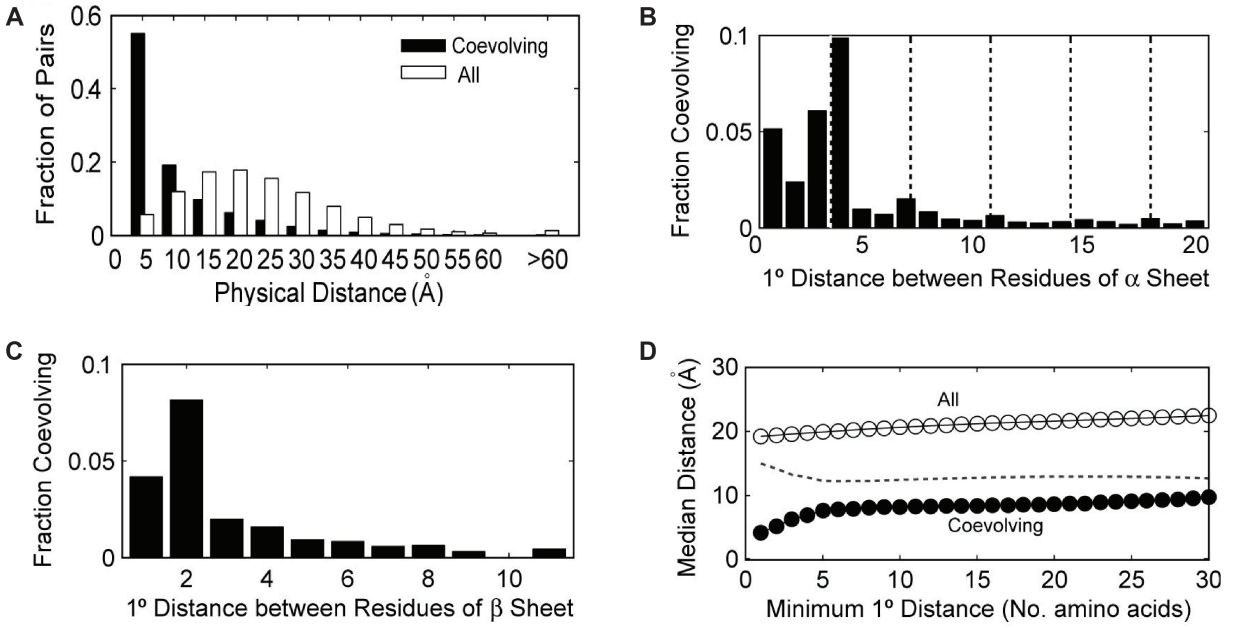

Figure 9.2: Coevolving residues correlate with structure. (A) The fraction of coevolving (black bars) or all (white bars) residue pairs that lie within the specified interval of physical distance from each other across 1592 Pfam families. (B) The fraction of residue pairs lying within the same α-helix and having the specified primary sequence separation that are coevolving. Neighboring residues have a primary (1°) distance of 1. Multiples of 3.6 have been superimposed onto the plot (dashed lines) to indicate typical spacing between turns of an α-helix. (C) The fraction of residue pairs lying within the same β-sheet and having the specified primary sequence separation that are coevolving. (D) The median distance of coevolving (closed circles) or all (open circles) residue pairs with the indicated minimum primary sequence separation. The dotted line depicts the difference between all and coevolving median distances.

tance of maintaining acceptor[A]-donor[D] interactions in side-chain hydrogen bonds (His[A/D]-His[A/D], Asp[A]-His[D], His[A/D]-Thr[A/D], His[A/D]-Tyr[A/D], His[D]-Glu[A], His[A/D]-Ser[A/D]) [7]. Interestingly, as noted, histidine along with serine, tyrosine, and threonine represent a class of amino acids whose side chains can act both as hydrogen donors and accepters [7]. I speculate that these amino acid pairs represent an evolutionary 'pivot-point' around which acceptors and donors can reverse roles. I also note that histidine is unique in its ability to act both as an acid and a base at physiological pHs suggesting that it may play a similar role in the evolutionary transitions between different acid-base pairs. Finally, coevolutionary pressures selecting against the reactive thiol group of cysteine may explain the high coevolution potential of the Cys-Cys pair.

The known importance of ionic interactions, hydrogen-bonds, and disulfide bonds in protein structure also offer a biochemical explanation for the correlation between physical structure and the

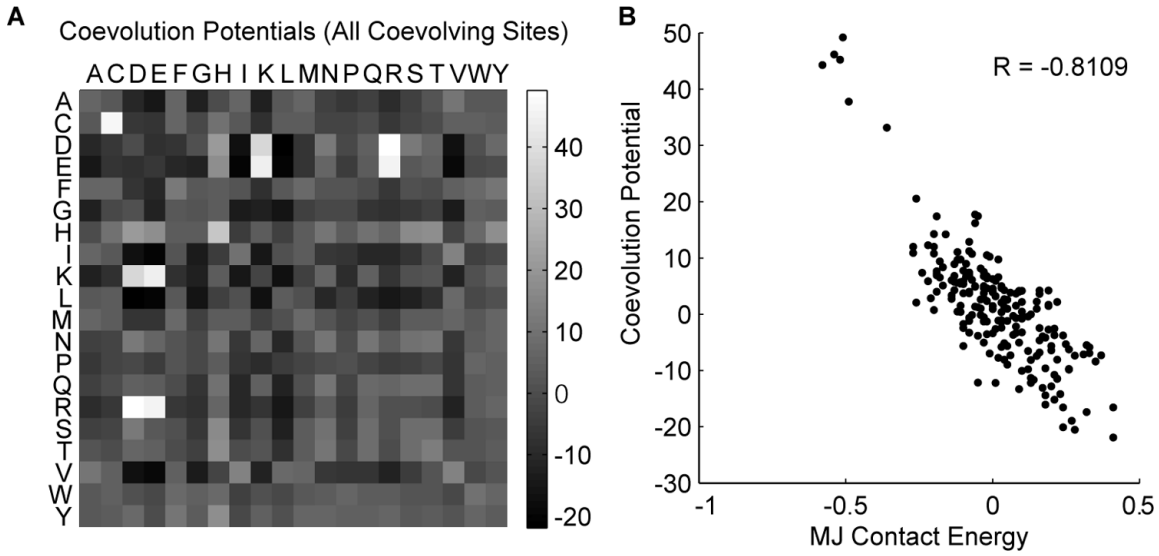

Figure 9.3: Coevolution potentials between the amino acids. (A) Coevolution potentials calculated using all identified coevolving sites. (B) Coevolution potentials are correlated with the MJ contact energies.

ZRes coevolution scores. Indeed the coevolution potentials showed high correlation to Miyazawa and Jernigan's contact energies, which describe the potential for amino acid pairs to be in physical contact with each other (MJ; $R = -0.8109$, Figure 9.3B) [112]. It is possible that the high coevolution potentials for these biochemically interacting amino acid pairs is actually a result of their correlation to physical proximity rather than an explanation for it. To test this possibility, I recalculated the coevolutionary potentials but only considered those pairs of sites that were already known to be within 6 $\AA$ of each other in the representative structure. Since these contacting coevolution potentials were normalized by the expected results for randomly selected contacting site-pairs, they represent the tendency for each amino acid pair to be found at a coevolving sites above and beyond the biases due to physical proximity. The results show that even once physical proximity has been removed as a bias in the potentials, acid-base, cysteine-cysteine, and hydrogen bond acceptor-donor pairs still dominate the coevolutionary interactions (Figure B.3A). Indeed the contacting coevolution potentials still strongly correlate with the MJ contact energies ($R = -0.7394$, Figure B.3B). I interpret these results as suggesting that a common form of coevolution arises from selective pressures to maintain important biochemical bonds, which are inherently short-range interactions. Such selective pressure would help to explain the tendency for coevolving sites to be close to each other.

While the correlation between the coevolution potentials and the MJ contact energies is consistent with my earlier findings that coevolving residues tend to be close together, many of the coevolving residues were not close in their representative structures. To investigate the amino acid compositions of these distant coevolving sites, I again recalculated the coevolution potentials considering only those residue pairs that were greater than 6 $\AA$ apart in their representative structures

(Figure B.3C). Surprisingly, even when considering only residue pairs that were greater 6 $\AA$ apart, the high coevolution potentials between acid-base pairs and the cysteine-cysteine pair remained high and still correlated to MJ contact energies ($R = -0.6601$, Figure B.3D). Thus, while these residues may be distant in the representative structures, their high coevolution score suggests that they may nonetheless still be close together in a different context such as different protein conformations, different representative structures, or contacts between copies of the protein in multi-protein complexes. I examined this last possibility in the following section.

My inability to separate distant coevolving residues out from those that interact at close-range makes it difficult to address the question of which amino acid pairs are common in long-range coevolutionary interactions. Nevertheless, the distant coevolution potentials did exhibit an increased ranking for pairs of aromatic amino acids in preference over several of the hydrogen-bond forming pairs identified by the earlier potentials: His-His (rank 6), Trp-Tyr (rank 7), Phe-Tyr (rank 8), and Trp-Trp (rank 10). It is unclear to me why these aromatic amino acid pairs were particularly represented among the distant coevolving residues.

9.4 Inter-molecular coevolution

In examining the coevolving residues of chorismate synthase I happened upon a surprising finding. Chorismate synthase is a homotetramerizing protein important in the synthesis of aromatic compounds in bacteria, and its crystal structure has been solved (PDB ID: 1UM0) [80]. Examining the distribution of distances between residues within a single chain of chorismate synthase (chain A in the representative crystal structure), I had found, as usual, that the coevolving residue pairs were significantly closer together than all tested residue pairs ($p < 1x10^{-}48$, K-S test; median distances: 5.78 $\AA$ (coevolving), 23.63 $\AA$ (all); Figure B.4). However, many of the strongly coevolving sites still seemed to be separated by a large physical distances. Interestingly, when I began mapping the strongest coevolving sites onto the crystal structure of the chorismate synthase tetramer, I found that many of these distant coevolving pairs were actually directly apposed to each other across the dimer interfaces (Figure 9.4A-C). Amongst the top 50 *ZRes* scoring residue pairs, 34 residue pairs (68%) were found to be contacting each other (6 $\AA$ apart) within a single molecule of chorismate synthase (chain A). Of the 16 pairs that were not in intra-molecular contact, 9 were found to be in contact between molecules of the tetramer (Figure 9.4A-C) and an additional pair was found to form a planar ring at the interface of the four chains (Lys-232 and Leu-349; Figure 9.4D). Many of these coevolving residues were predicted by UCSF Chimera to form inter-molecular hydrogen bonds (data not shown) [93]. Taken together with the previous results, this suggests that residues may coevolve to maintain structural interactions both within and between protein molecules.

To further test this hypothesis, I identified 532 alignments whose representative crystal structure contained multiple copies of the corresponding peptide. Since formation of protein crystals inherently imposes a multimerization of the peptides, I restricted my analysis to only those chains in the structure identified as being part of a biologically relevant assembly (REMARK 350 in PDB files) [17]. Plotting the joint histogram of intra-molecular and inter-molecular distances for the coevolving sites normalized to the joint histogram for all tested sites, I found that the coevolving sites were disproportionately represented amongst sites that were physically close either within a

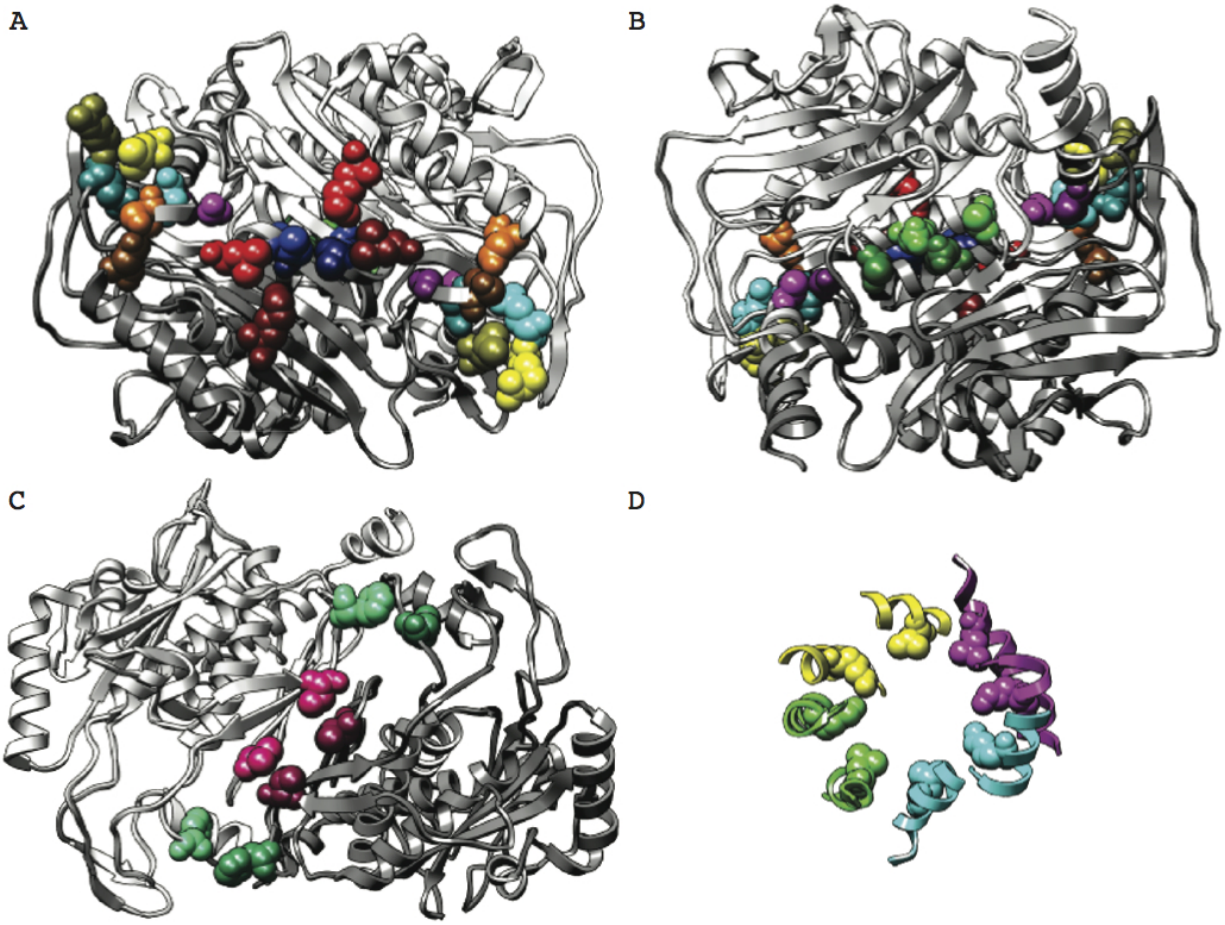

Figure 9.4: Inter-molecular interactions between coevolving residues of the chorismate synthase tetramer. (A-C) Coevolving residues are highlighted by the same hue. Light residues are from chain A. Dark residues are from chain D (panels A and B) or chain C (panel C). (B) The back side of the structure depicted in panel A. (D) A pair of coevolving residues forming a planar ring at the center of the tetramer. Each molecule of chorismate synthase is depicted in a different color.

protein or between interacting copies of the protein (Figure 9.5). Of all 9207 residues pairs that were within 6 $\AA$ of each other in inter-molecular distance, over 10% (1167 pairs) of them were identified as coevolving. In comparison, only 0.7% of all site-pairs (distant or close) were selected as coevolving. These results clearly demonstrate the importance of inter-molecular interactions in the coevolution of residues.

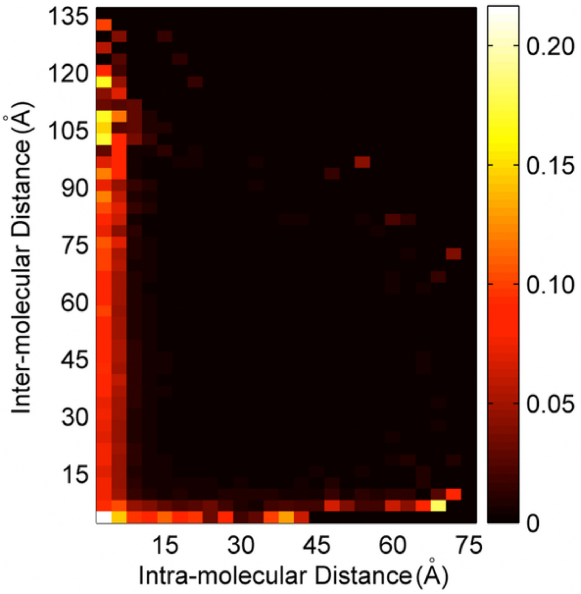

Figure 9.5: Joint distribution of intra-molecular and inter-molecular distances between coevolving residues. 532 protein and domain alignments whose representative PDB structures contained multiple copies of the corresponding peptide were used for the analysis. The color of each cell depicts the fraction of all residues pairs lying within the specified intervals of intra-molecular and inter-molecular distances that are coevolving. Coevolving pairs are particularly prevalent amongst residues pairs that lie in close physical proximity to each other either intra-molecularly or inter-molecularly.

9.5 Coevolution of catalyic sites

I next examined whether catalytic sites, being direct participants in the functional role of enzymatic proteins, exhibited specific coevolutionary tendencies. Two lines of evidence have commonly been offered to support the hypothesis that catalytic sites elicit or require strong coevolutionary interactions: 1) examples of catalytic sites coevolving with other (not necessarily catalytic) sites are highlighted, or 2) a prevalence of non-catalytic coevolving sites within 10 $\AA$ of a protein's active sites is demonstrated [32, 42, 132, 135, 140]. Statistical support verifying that these trends surpass random expectations, however, is often not offered. Furthermore, care should be given towards considering what biases in an algorithm might inappropriately increase coevolutionary measures for catalytic sites. For example, since low entropy is correlated with high conservation, the normalization of MI by $H_{i,j}$ introduced by Gloor *et al.* might bias the measure towards selecting evolutionarily conserved sites [42, 135].

The Catalytic Site Atlas (CSA) provides information on which residues in a PDB structure are implicated in the direct catalytic activity of an enzyme [141]. Of the 1240 representative crystal structures utilized in this study, a total of 645 catalytic sites in 257 proteins had been identified in the CSA. Using the *ZRes* method, I found that 61.6% (397) of these sites were identified as

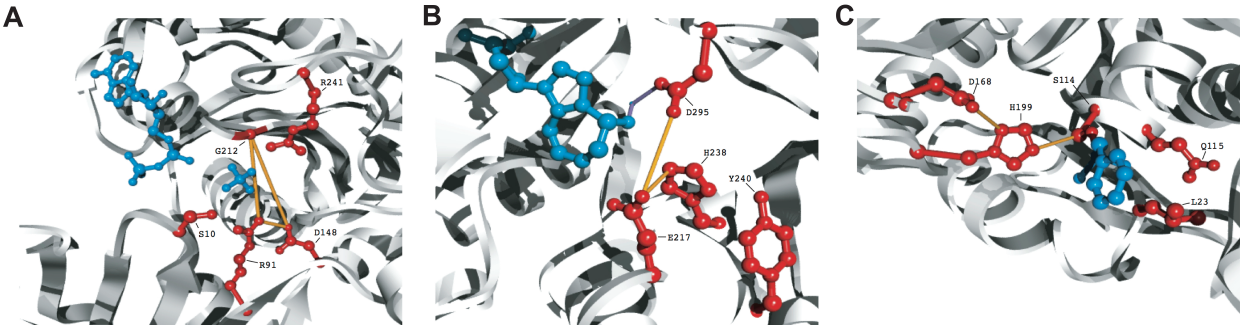

Figure 9.6: Coevolution between catalytic sites. All catalytic sites annotated by the CSA [100] and tested for coevolution (i.e.20% gaps) are depicted in red. The protein backbones are depicted as a white ribbon. Coevolving catalytic residues are connected by orange lines. (A) The nucleotide binding site of Methanosarcina thermophila acetate kinase (PF00871, PDB 1g99) [22]. The bound ADP molecule and a sulfate ion are depicted in blue. (B) Active site of murine adenosine deaminase (PF00962, PDB 1a4l) [136]. The inhibitor, pentostatin, and a coordinating Zn2+ ion are depicted in blue. The coordinating interactions with Zn2+ is depicted as purple lines [136]. (C) Active site of Pseudomonas fluorescens carboxylesterase (PF02230, PDB 1aur) [62]. The inhibitor, phenylmethylsulfonyl fluoride, is covalently bound to Ser114 and its phenylmethylsulfonyl moiety is depicted in blue.

coevolving with at least one other site, and 97.8% (631) were within 10$\AA$ of a coevolving pair of sites. While these may seem like a large representations of the catalytic sites and are comparable to previous reports [42, 46, 60, 129, 135], they were not larger than the portion of all sites that had coevolving partners (61.2%) nor larger than the portion of all sites within 10$\AA$ of a coevolving pair (98.8%). I therefore conclude that, while functionally relevant sites are indeed amongst the coevolving sites, they have no increased propensity to be coevolving over other sites.

While functional sites are no more likely to have coevolving partners than random sites, I wondered whether functional sites tend to coevolve specifically with each other. Of the 257 PDB structures with CSA entries, 175 had at least two catalytic sites annotated and were used for the subsequent analysis. I found that 61 of these PDB structures contained at least one pair of catalytic sites identified as coevolving with each other. In total, there were 90 such coevolving pairs of catalytic sites, representing 11% of all possible catalytic site pairs (793). To determine whether this propensity for catalytic sites to coevolve with one another was significant, for each of the 175 crystal structures I selected a number of random sites equal to the number of catalytic sites and asked how many random site pairs were coevolving. Over 2000 randomizations, the average total number of coevolving random pairs was only 6.5 ± 2.7 (0.8%), significantly fewer than the number of identified coevolving catalytic sites (the probability of finding at least 90 coevolving sites given a normal fit of the random results, log transformed to satisfy normalicy, was less than 10^{-16}). When random pairs were chosen only amongst those sites that were contacting each other, only 56.7 ± 7.2 (7.2%) were identified as coevolving, showing that the tendency for catalytic sites to coevolve was not due to their potential tendency to be located near each other at active sites

($p < 10^{-16}$). Three example proteins containing coevolving catalytic sites have been depicted in Figure 9.6.

Chapter 10

Comparison to previous algorithms

To compare the performance of the *ZRes* algorithm to previously published methods, I considered several measures that attempt to detect residue coevolution by quantifying the covariability between sites. I had chosen to utilize an *MI*-based approach because *MI* is well established in Information Theory as a measure of codependency. Other methods for quantifying the covariability, however, have been adapted towards coevolution detection. The Observed Minus Expected Squared (OMES) approach developed by Kass and Horovitz utilized a χ^2 goodness-of-fit test to identify site pairs at which the observed distribution of amino-acid pairs diverged significantly from expectation [36, 60]. The McLachlan Based Substitution Correlation (McBASC) approach developed by Göbel *et al.* looked for correlations in the degrees of divergence for paired substitutions at two sites [36, 44, 88]. Furthermore, a recent report from Dunn *et al.* independently developed a measure of coevolution (MIp) analogous to our *Res* measure [32]. A subtle difference lies in how $\overline{MI}_i \cdot \overline{MI}_j$ is removed from the *MI* score. Dunn *et al.* utilized an insightful mathematical proof, to estimate the relationship between *MI* and $\overline{MI}_i \cdot \overline{MI}_j$. I, on the other hand, directly calculated the residuals of the linear regression of the measure on the bias. Dunn *et al.* however, did not account for the differences in within-site variability addressed by the *ZRes* measure [32].

To compare the *ZRes* algorithm to these previously developed methods, I used contact prediction accuracy as an approximate correlate of coevolution prediction accuracy. Since none of these algorithms utilize structural data (including primary sequence order) and since none of them are based on known signals for contact prediction, any correlation with structural data should arise from their ability to recognize coevolving sites combined with a tendency for coevolving sites to be close together (or for close residues to be coevolving). Contact prediction therefore is a reasonable approximation of algorithm accuracy. In order to make the comparisons, each measure was used to rank all tested site pairs for each analyzed protein family, and the percentage of the top ranking site pairs contacting in their representative structures were calculated. Our ZRes measure out-performed both OMES and McBASC ($p < 10^{16}$, Friedman's nonparametric two-way ANOVA; Figure 10.1A). Furthermore, whereas MIp and *Res* performed equally well, they both under-performed *ZRes*, showing that accounting for heteroscedasticity significantly improved the measure ($p < 10^{16}$; Figures 9.6A and B). Since shorter protein sequences have a large fraction of residue pairs in contact with each other (Figure B.2B), I repeated the analysis adjusting for sequence length

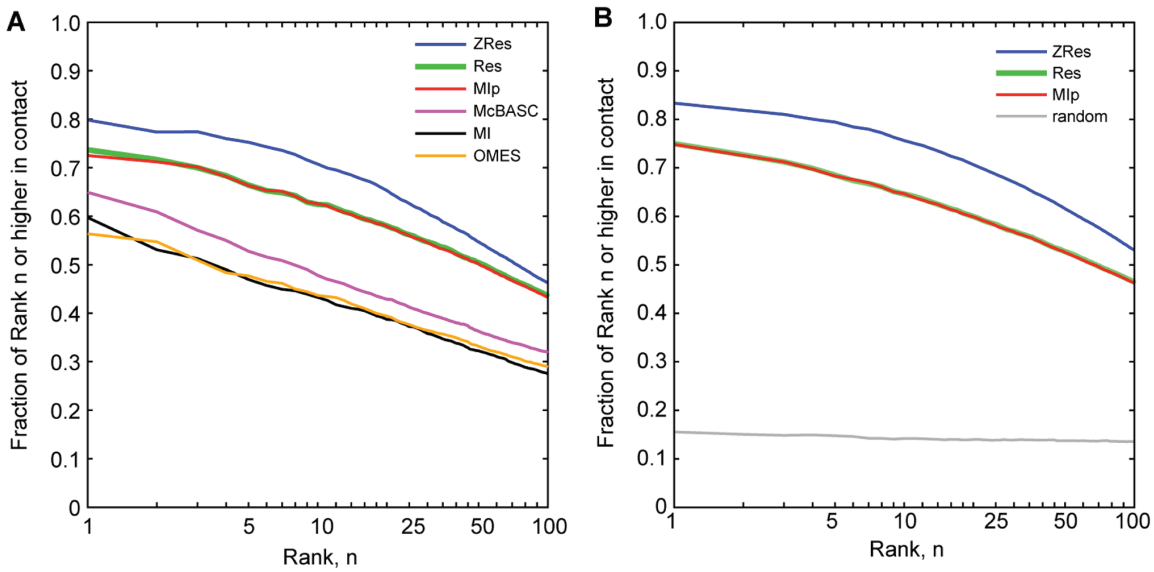

Figure 10.1: Comparison of *ZRes* to other measures of coevolution. (A) To ease processing load, calculations were limited to the 424 alignments with representative structures for which the product of the protein sequence length and alignment size was less than or equal to 100,000. Following the analysis performed previously [129], all residue pairs were ranked from highest to lowest *ZRes* score. For ranks 1 up to 100, the fraction of residue pairs at or higher than each rank lying within 6 $\AA$ of each other was calculated. The average of this contact accuracy across all alignments was then plotted (blue). The process was repeated with the *Res* (green), OMES (brown), McBASC (magenta), MIp (red), and MI (black) measures. (B) as in A, but utilizing all 1240 alignments with representative crystal structure. The results from one randomization of residue pair rankings are plotted in black. Statistical significance was assessed by Friedman's nonparametric 2-way ANOVA for measure effects on selectivity after factoring out rank effects. All pair-wise comparison in both A and B were significant except between MIp and Res.

by normalizing the number of top scoring site pairs chosen for each protein family by the length of the protein sequence (Figure B.5). Again, ZRes performed significantly better than all other measures ($p < 0.05$ for 1% protein sequence length down to $p < 10^5$ for 32% protein sequence length, K-S test).

Chapter 11

Discussion

Since it is presently infeasible to directly test whether two residues in a protein interacted evolutionarily, researchers must defer to correlation as evidence for the accuracy of an algorithm. Even when such correlations exist and are shown to be statistically significant, care must be taken in considering whether a non-coevolutionary variable or bias in the measure might be the underlying source of the correlation. Once such non-coevolutionary explanations have been ruled out, these correlations not only validate an algorithm but also provide insights into the nature of coevolution.

The coevolving residues found by my algorithm show a strong correlation with physical structure, namely coevolving residues tend to be in close proximity to each other. Since the algorithm uses no information on structural data, not even the primary sequence order, this correlation suggests that residues that lie in close physical proximity are more likely to influence the selective pressure acting on each other. While previous algorithms have demonstrated high correlation to structure, they achieve this with only a limited number of residue pairs. While retaining high selectivity, my algorithm on average identified over half of all residues measured as coevolving with at least one other residue, a level of coverage not previously reported. For example, a recent study by Yeang *et al.* searched 2000 protein families to find 2000 coevolving residue pairs, 50% of which were in contact [140]. In comparison, applying my algorithm to 1592 protein families, I isolated over 50 times as many coevolving pairs (126,085) and yet retained the same level, if not higher, of contact accuracy (59%), demonstrating the increased sensitivity of my algorithm.

The high performance of the *ZRes* algorithm at predicting residue contacts may in the future offer a means of improving protein structure prediction algorithms. Indeed several methods for combining coevolutionary measures in structural predictions have been previously described and would be interesting to pursue in future studies [33, 79, 112].

The calculation of coevolution potentials between the 20 amino acids offers new insights into the role of biochemical interactions in evolution. The results suggest that bond forming residue pairs may commonly face particularly strong coevolutionary selective pressure, probably towards maintaining these bonds. Although selective pressure typically suggest conservation, one should bear in mind that coevolution requires variation. Thus the capacity for similar bonds to be formed by different amino acid pairs may provide a means to maintain necessary physical interactions while tolerating variation. The predominance among coevolving residues of acid-base pairs could

also suggest that coevolutionary selective pressures act to maintain a balance of ionic charges. The high coevolution potential of the cysteine-cysteine pair may suggest a similar balancing pressure to protect against the high reactivity of the cysteine thiol group.

The coevolution potentials for "distant" residues highlighted the importance of context in investigating algorithms for detecting coevolution. While one could explain the coevolution of distant residues of opposite charges as maintaining a global balance of ionic charge, the persistence of cysteine-cysteine pairs among the highest coevolution potentials would be hard to explain if such residues were indeed distant. More likely they are only distant in one context but are close in another. The physical interaction of these residues may be revealed if we consider their structures from a different context such as looking at different representative structures within an alignment or at different conformational states of the protein. As one example (Figure 9.6), I showed that the structural correlations between seemingly distant coevolving sites can be revealed upon consideration of inter-molecular distances within a protein complex.

It is often expected that coevolving residues tend to play fundamental roles in the function of a protein. Researchers therefore often highlight those predicted coevolving residues that are known to play important roles in protein function. However, they rarely offer statistical support for the hypothesis that coevolving residues have a propensity for being directly involved in protein function. I have shown that catalytic sites, as determined by the CSA, do not have an increased propensity to coevolve in general. I did, however, reveal an increased tendency for these sites to coevolve specifically with each other. Thus, catalytic sites selectively coevolve more strongly with other catalytic sites. Since this correlation to coevolution was identified only for pairs of catalytic sites and was not present when considering catalytic sites one at a time, it is not likely to arise from site-specific biases. These findings underscore the importance of residue coordination in realizing and maintaining an optimal enzymatic activity.

To explain the competing roles of selective pressure and variation, both necessary for coevolution, I offer a coevolutionary extension of the Neutral Model of Evolution offered by Kimura [79], and King and Jukes [20]. I hypothesize that coevolutionary change predominantly occurs through the genetic drift of neutral mutations at interacting sites, but the set of neutral mutations available to those sites is largely restricted to maintaining structural and biochemical interactions. When multiple means of retaining such interactions are available (e.g. multiple ways of forming similar bonds), these selective forces would not be so constraining that they prevent any variation at the sites. As nearly-neutral mutations stabilize, the interactions between each residue change, altering the set of subsequently available neutral mutations. Given that variability is important in the detection of coevolution, those residue pairs that most strongly cooperate in defining the shape of a protein's mutational landscape without severely restricting it will exhibit the strongest coevolutionary signal. This might further explain why catalytic sites do not exhibit a general increase in tendency to coevolve. Perhaps many functional sites are too constrained to allow any variation, and thus do not allow any covariation.

Appendix A

Supplemental proofs and methods for Part I

A.1 Derivation of Mean Path Length

To optimize navigation to a target state s^*, we consider modified transition probabilities:

$$p_{navigation}(s'|a,s) = \begin{cases} \Theta_{ass'} & \text{if } s \neq s^* \\ 1 & \text{if } s = s' = s^* \\ 0 & \text{otherwise} \end{cases}$$

A navigational utility function is then defined as:

$$U_{navigation}(s) = \begin{cases} -1 & \text{if } s \neq s^* \\ 0 & \text{otherwise} \end{cases}$$

An optimal policy π is derived through value-iteration as follows:

$$\begin{aligned} Q_0(a,s) &:= U_{navigation}(s) \\ Q_{\tau-1}(a,s) &:= U_{navigation}(s) + \sum_{s' \in \mathscr{S}} p_{navigation}(s'|a,s) \cdot V_\tau(s') \\ &\text{where } V_\tau(s) := \max_a Q_\tau(a,s) \end{aligned}$$

Value-iteration is continued until V converges, and the optimal policy is then defined as:

$$\pi(s) = \arg\max_a Q_{convergence}(a,s)$$

The expected path length to target s^* is then calculated as:

$$\mathrm{E}[steps\ to\ s^*] = \sum_s -\frac{1}{N} V_{convergence}(s)$$

The mean path length is then taken to be the average of the expected path length over the N possible target states.

A.2 Derivation of PEIG

Theorem 1. *Surprise, as employed by Storck et al. [123], is equal to the posterior expected information gain. That is, if an agent is in state s and has previously collected data $\vec{\mathrm{d}}$, then the expected information gain for taking action a and observing resultant state s^* is given by:*

$$\text{Surprise}(a,s,s') := \mathrm{D}_{\mathrm{KL}}(\widehat{\Theta}^{\vec{\mathrm{d}}\cup s^*}_{as\cdot} \parallel \widehat{\Theta}^{\vec{\mathrm{d}}}_{as\cdot}) = \mathrm{E}_{\Theta|\vec{\mathrm{d}}\cup s^*}[\mathrm{I}_{\mathrm{G}}(a,s,s^*)] \tag{A.1}$$

Proof.

$$\begin{aligned}
\mathrm{E}_{\Theta|\vec{\mathrm{d}}\cup s^*}[\mathrm{I}_{\mathrm{G}}(a,s,s^*)] &= \mathrm{E}_{\Theta|\vec{\mathrm{d}}\cup s^*}\left[\sum_{s'}\Theta_{ass'}\log_2\left(\frac{\widehat{\Theta}^{\vec{\mathrm{d}}\cup s^*}_{ass'}}{\widehat{\Theta}^{\vec{\mathrm{d}}}_{ass'}}\right)\right] \\
&= \sum_{s'}\mathrm{E}_{\Theta|\vec{\mathrm{d}}\cup s^*}[\Theta_{ass'}]\log_2\left(\frac{\widehat{\Theta}^{\vec{\mathrm{d}}\cup s^*}_{ass'}}{\widehat{\Theta}^{\vec{\mathrm{d}}}_{ass'}}\right) \\
&= \sum_{s'}\widehat{\Theta}^{\vec{\mathrm{d}}\cup s^*}_{ass'}\log_2\left(\frac{\widehat{\Theta}^{a,s\to s^*}_{ass'}}{\widehat{\Theta}^{\vec{\mathrm{d}}}_{ass'}}\right) \\
&= \mathrm{D}_{\mathrm{KL}}(\widehat{\Theta}^{\vec{\mathrm{d}}\cup s^*}_{as\cdot} \parallel \widehat{\Theta}^{\vec{\mathrm{d}}}_{as\cdot})
\end{aligned}$$

□

A.3 Methods for assessing performance in goal-directed tasks

To assess the general utility of an agent's internal model, the agent is first allowed to explore for a fixed number of times steps. After exploring, the agent is asked, for each goal-directed task, to choose a fixed policy that optimizes performance under its learned model:

Navigation: To optimize navigation to a target state s^* under internal model $\widehat{\Theta}$, I took an approach analogous to the method for calculating the mean path length of a world (see Appendix 5.4). We first consider modified transition probabilities:

$$p_{navigation}(s'|a,s;\widehat{\Theta}) = \begin{cases} \widehat{\Theta}_{ass'} & \text{if } s \neq s^* \\ 1 & \text{if } s = s' = s^* \\ 0 & \text{otherwise} \end{cases}$$

A navigational utility function is then defined as:

$$U_{navigation}(s) = \begin{cases} -1 & \text{if } s \neq s^* \\ 0 & \text{otherwise} \end{cases}$$

An optimal policy $\pi_{\widehat{\Theta}}$ is derived through value-iteration as follows:

$$\begin{aligned}
&Q_0(a,s) := U_{navigation}(s) \\
&Q_{\tau-1}(a,s) := U_{navigation}(s) + \sum_{s' \in \mathscr{S}} p_{navigation}(s'|a,s;\widehat{\Theta}) \cdot V_\tau(s') \\
&\qquad \text{where } V_\tau(s) := \max_a Q_\tau(a,s)
\end{aligned}$$

This process is iterated a number a times, $\tau_{convergence} > 1000$, sufficient to allow Q to converge to within a small fixed margin. An optimal policy is then defined as:

$$\pi_{\widehat{\Theta}}(s) = \arg\max_a Q_{-\tau_{convergence}}(a,s)$$

The realized performance of $\pi_{\widehat{\Theta}}$ is assessed as the expected number of time steps, capped at 20, it would take an agent employing $\pi_{\widehat{\Theta}}$ to reach the target state. A true optimal policy is calculated as above except using Θ instead of $\widehat{\Theta}$. For each world and each exploration strategy, navigation is assessed after $t \in \{25, 50, 75, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000\}$ exploration time steps and compared to the true optimal strategy. Performance difference from true optimal is calculated is averaged over the tested exploration lengths, all starting states, and all target states. The different explorative strategies are then ranked in performance.

Reward Acquisition: Policies in reward acquisition tasks are derived as above for navigational tasks except as follows:

$$\begin{aligned}
&p_{reward}(s'|a,s;\widehat{\Theta}) = \widehat{\Theta}_{ass'} \\
&U_{reward}(s) \sim Uniform([-1,1]) \\
&\pi_{\widehat{\Theta}}(s) = \arg\max_a Q_{-100}(a,s)
\end{aligned}$$

Realized performance is assessed as the expected total rewards accumulated by an agent employing $\pi_{\widehat{\Theta}}$ over 100 time steps.

Appendix B

Supplemental figures for Part II

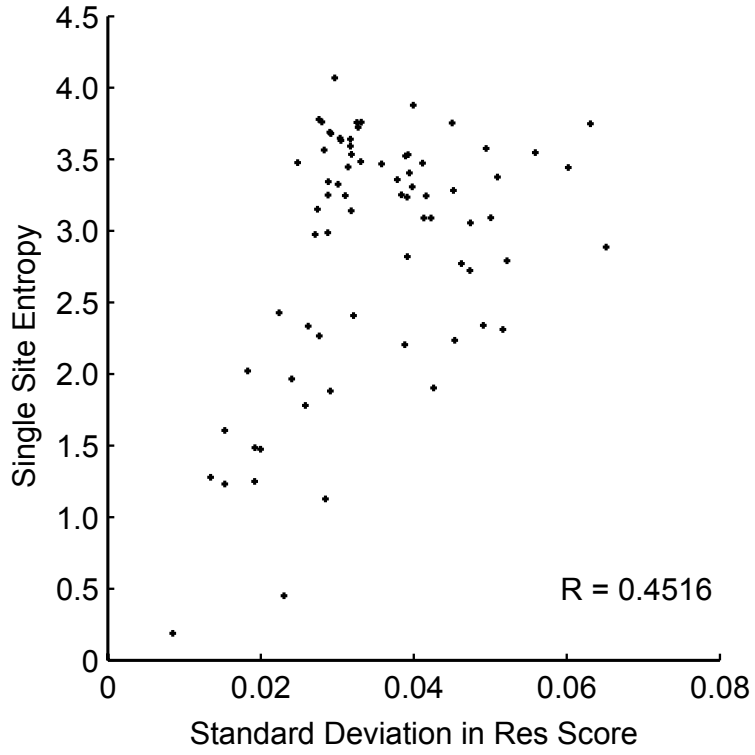

Figure B.1: Within-site *Res* score variability correlates with entropy. The entropy of each site is plotted against the standard deviation of *Res* scores at that site. The positive correlation suggests that sites with higher variation in amino acid composition are more likely to exhibit spuriously high *Res* scores.

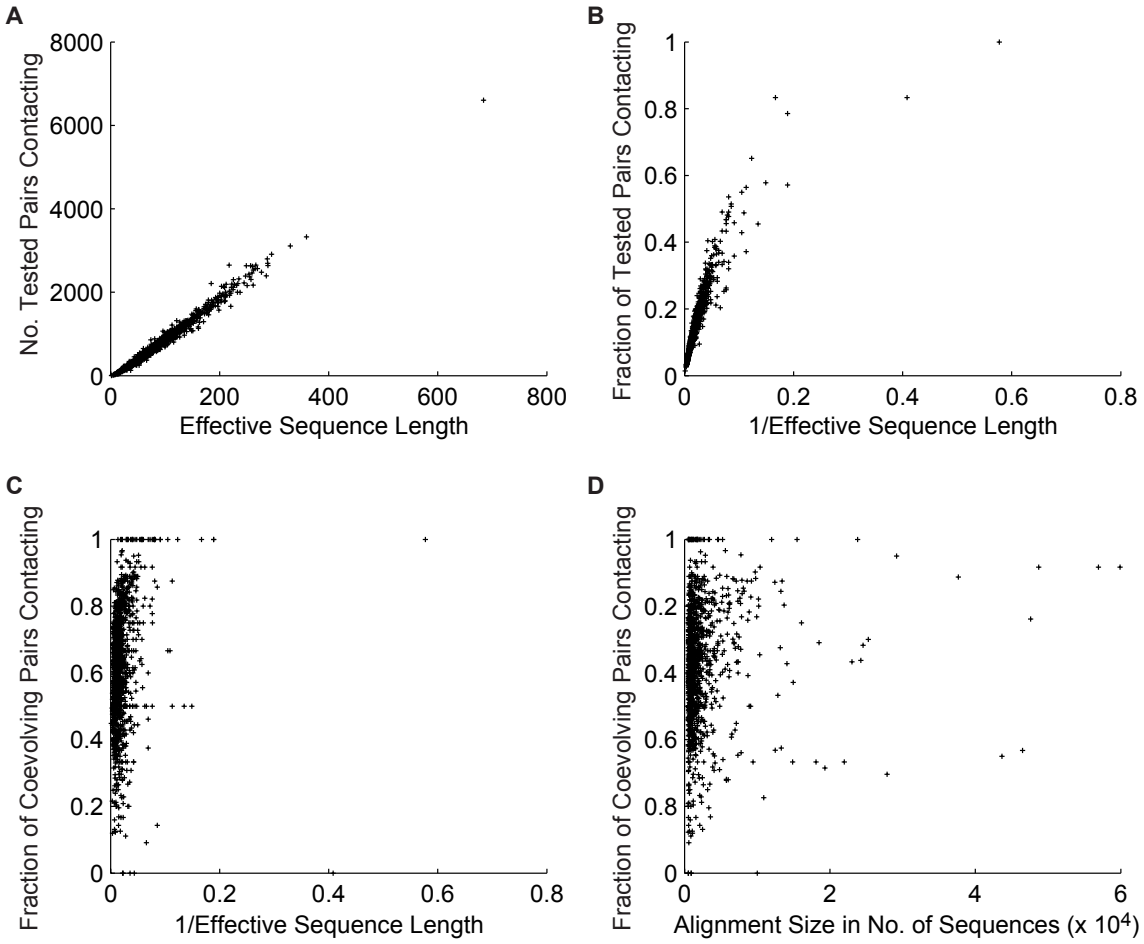

Figure B.2: Contact accuracy is weakly correlated with the reciprocal of protein length and alignment size. For each alignment for which a representative structure was available: (A) The number of tested residue pairs that were contacting each other was plotted against the effective protein sequence length; (B) The fraction of the tested residue pairs that were contacting each other was plotted against the reciprocal of effective sequence length; (C) The fraction of residue pairs identified as coevolving that were contacting each other was plotted against the reciprocal of effective sequence length; (D) The fraction of residue pairs identified as coevolving that were contacting each other was plotted against the alignment size.

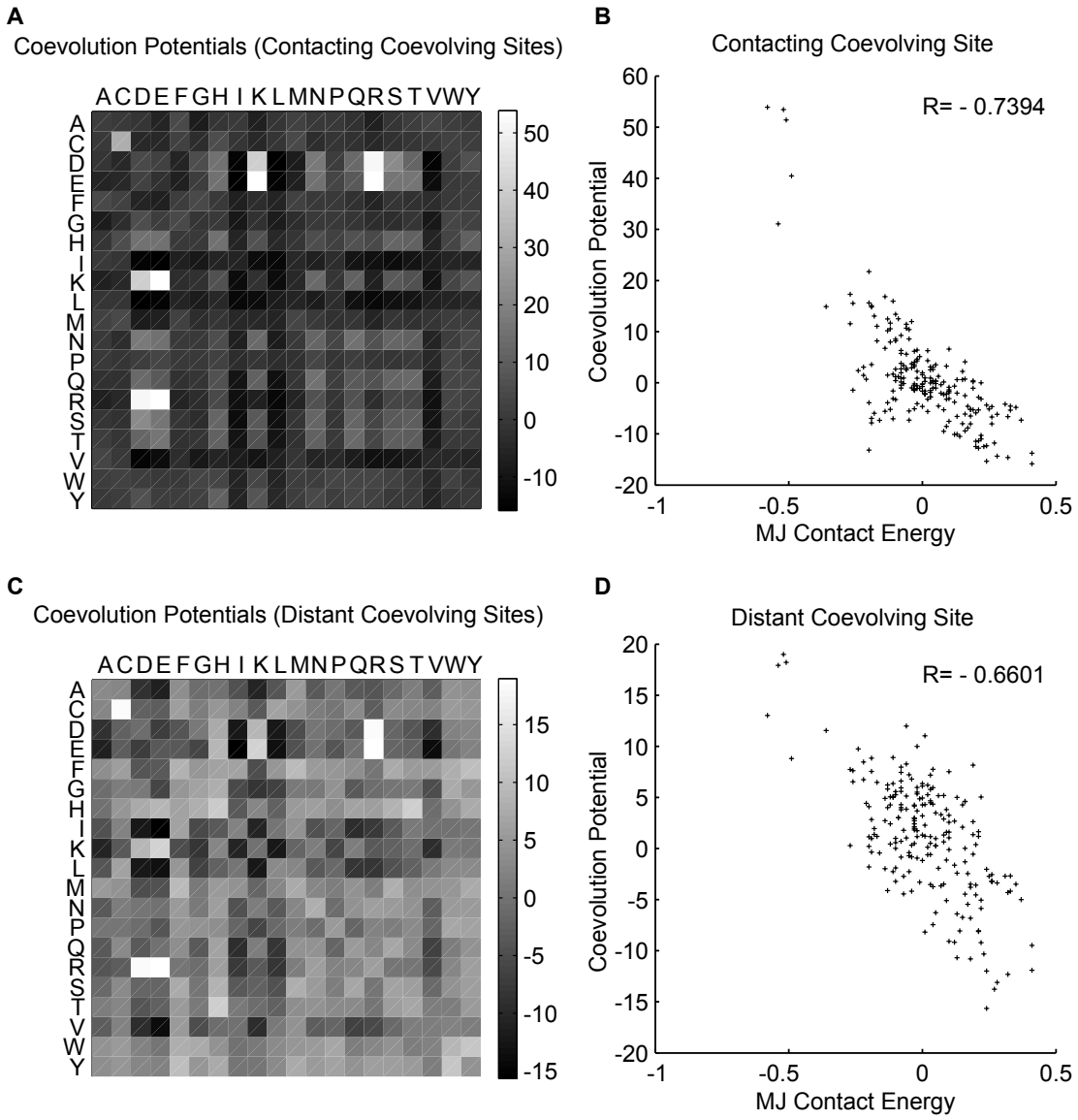

Figure B.3: Amino acid coevolution potentials for contacting and distant residue pairs. (A) Coevolution potentials calculated only for residues pairs no further than 6 $\AA$ apart (intra-molecular distance). (B) Coevolution potentials amongst contacting residue pairs are correlated with the MJ contact energies. (C) Coevolution potentials calculated only for residues pairs that are at least 6 $\AA$ apart. (D) Coevolution potentials amongst distant residue pairs are still correlated with the MJ contact energies.

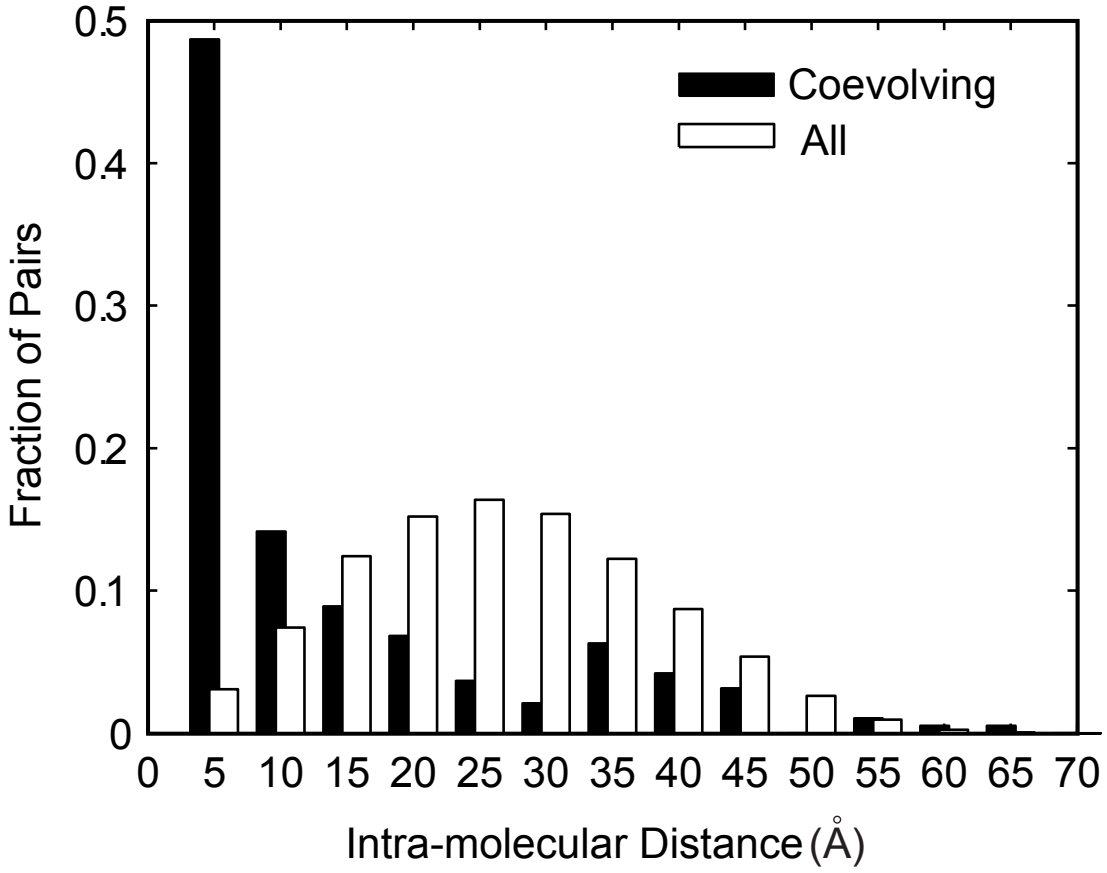

Figure B.4: Distribution of intra-molecular distances between coevolving residues of chorismate synthase The fraction of coevolving (black bars) or all (white bars) residue pairs that lie within the specified interval of physical distance from each other is depicted.

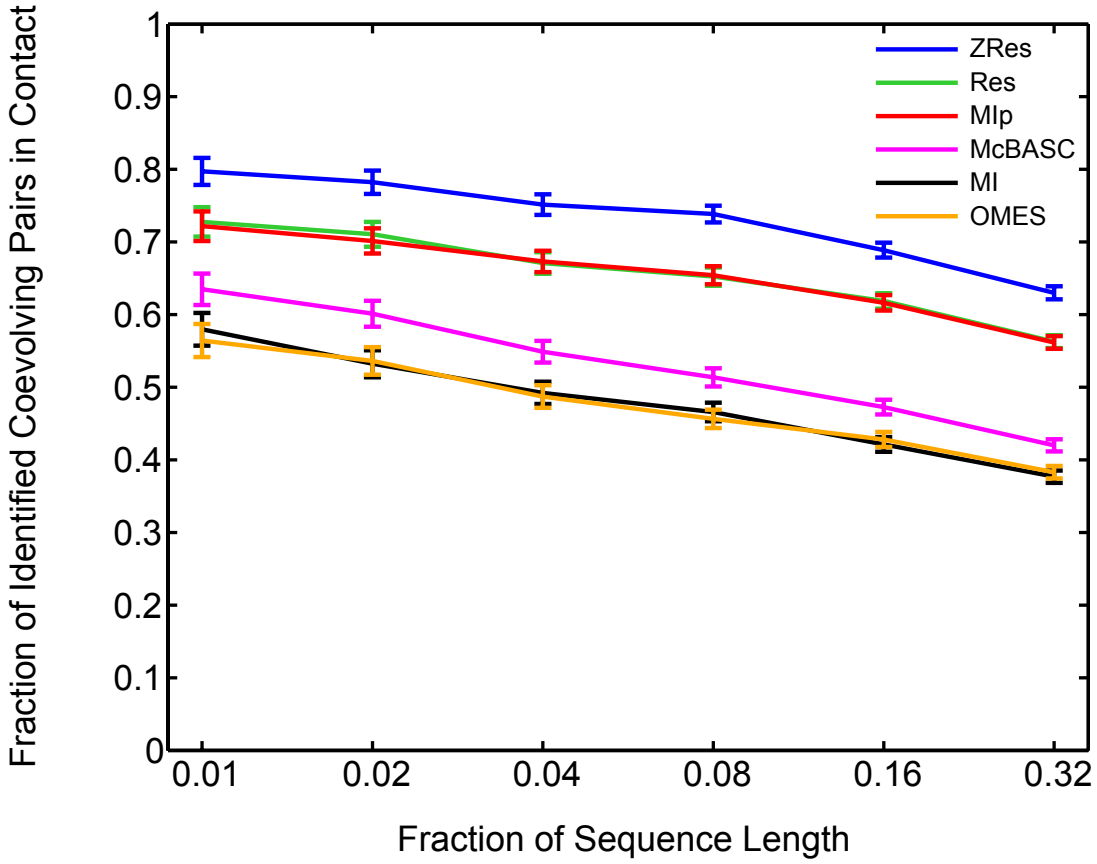

Figure B.5: The same dataset used in Figure 10.1A was reanalyzed to normalize for variation in protein sequence length. Again all residue pairs were ranked from highest to lowest *ZRes* score. For each protein, a number of top ranking residue pairs proportional to the length of the protein sequence (characterized as fractions of the protein sequence length) were considered. The fraction of these high-ranking residue pairs that lied within 6 *Å* of each other was then calculated. The average of this contact accuracy across all alignments was then plotted (blue). The process was repeated with the *Res* (green), OMES (brown), McBASC (magenta), MIp (red), and *MI* (black) measures.

Bibliography

[1] Curiosity. In NorbertM. Seel, editor, *Encyclopedia of the Sciences of Learning*, pages 894–894. Springer US, 2012.

[2] P. Abbeel and A.Y. Ng. Exploration and apprenticeship learning in reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 1–8. ACM, 2005.

[3] Naoki Abe, Naval Verma, Chid Apte, and Robert Schroko. Cross channel optimized marketing by reinforcement learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 767–772. ACM, 2004.

[4] J. Archer and L. Birke. *Exploration in animals and humans*. Van Nostrand Reinhold (UK) Co. Ltd., 1983.

[5] J. Asmuth, L. Li, M.L. Littman, A. Nouri, and D. Wingate. A bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 19–26. AUAI Press, 2009.

[6] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich. Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B-Condensed Matter and Complex Systems*, 63(3):329–339, 2008.

[7] E. N. Baker and R. E. Hubbard. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.*, 44(2):97–179, 1984.

[8] P. F. Baldi and L. Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666, Jun 2010.

[9] Izhar Bar-Gad, Genela Morris, and Hagai Bergman. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71(6):439 – 473, 2003.

[10] J. Baron. Rationality and intelligence. 1985.

[11] A.G. Barto and S.P. Singh. On the computational economics of reinforcement learning. In *Connectionist Models: Proceedings of the 1990 Summer School. Morgan Kaufmann*. Citeseer, 1990.

[12] Jonathan Baxter, Andrew Tridgell, and Lex Weaver. Knightcap: A chess programm that learns by combining td (lambda) with game-tree search. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 28–36. Morgan Kaufmann Publishers Inc., 1998.

[13] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[14] DE Berlyne. Novelty and curiosity as determinants of exploratory behaviour1. *British Journal of Psychology. General Section*, 41(1-2):68–80, 1950.

[15] D.E. Berlyne. Conflict, arousal, and curiosity. 1960.

[16] D.E. Berlyne. Curiosity and exploration. *Science*, 153(3731):25, 1966.

[17] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nat. Struct. Biol.*, 10(12):980, Dec 2003.

[18] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[19] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pages 23–37. Springer, 2009.

[20] L. Burger and E. van Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a bayesian method. *Mol. Syst. Biol.*, 4:165, 2008.

[21] G.M. Burghardt. *The Genesis of Animal Play: Testing the Limits*. Bradford Bks. Mit Press, 2005.

[22] Kathryn A. Buss, David R. Cooper, Cheryl Ingram-Smith, James G. Ferry, David Avram Sanders, and Miriam S. Hasson. Urkinase: Structure of acetate kinase, a member of the askha superfamily of phosphotransferases. *Journal of Bacteriology*, 183(2):680–686, 2001.

[23] E.L. Charnov et al. Optimal foraging, the marginal value theorem. *Theoretical population biology*, 9(2):129–136, 1976.

[24] Michael X Cohen and Charan Ranganath. Reinforcement learning signals predict future decisions. *The Journal of neuroscience*, 27(2):371–378, 2007.

[25] T.M. Cover and J.A. Thomas. *Elements of information theory*, volume 6. Wiley Online Library, 1991.

[26] Marco Crisma, Fernando Formaggio, Alessandro Moretto, and Claudio Toniolo. Peptide helices based on -amino acids. *Peptide Science*, 84(1):3–12, 2006.

[27] James P. Crutchfield and Karl Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, Jul 1989.

[28] JP Crutchfield and DP Feldman. Regularities unseen, randomness observed: levels of entropy convergence. *Chaos (Woodbury, NY)*, 13(1):25, 2003.

[29] Peter Dayan and Yael Niv. Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185 – 196, 2008. ¡ce:title¿Cognitive neuroscience¡/ce:title¿.

[30] Edward L Deci and Richard M Ryan. The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4):227–268, 2000.

[31] Kenji Doya. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10(6):732 – 739, 2000.

[32] S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008.

[33] Piero Fariselli, Osvaldo Olmea, Alfonso Valencia, and Rita Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14(11):835–843, 2001.

[34] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. The pfam protein families database. *Nucleic Acids Research*, 36(suppl 1):D281–D288, 2008.

[35] WalterM. Fitch and Etan Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5):579–593, 1970.

[36] Anthony A. Fodor and Richard W. Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2):211–221, 2004.

[37] K. Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

[38] H. Gimbert. Pure stationary optimal strategies in markov decision processes. *STACS 2007*, pages 200–211, 2007.

[39] B. G. Giraud, Alan Lapedes, and Lon Chang Liu. Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E*, 58:6312–6322, Nov 1998.

[40] J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

[41] M. Glanzer. Curiosity, exploratory drive, and stimulus satiation. *Psychological Bulletin*, 55(5):302, 1958.

[42] Gregory B. Gloor, Louise C. Martin, Lindi M. Wahl, and Stanley D. Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, 2005. PMID: 15882054.

[43] Jan Glscher, Nathaniel Daw, Peter Dayan, and John P. O'Doherty. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585 – 595, 2010.

[44] Ulrike Gobel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994.

[45] G. Gordon, D.M. Kaplan, B. Lankow, D.Y.J. Little, J. Sherwin, B.A. Suter, and L. Thaler. Toward an integrated approach to perception and action: Conference report and future directions. *Frontiers in Systems Neuroscience*, 5, 2011.

[46] R. Gouveia-Oliveira and A. G. Pedersen. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms Mol Biol*, 2:12, 2007.

[47] R. W. Guillery. Anatomical pathways that link perception and action. *Prog. Brain Res.*, 149:235–256, 2005.

[48] R. W. Guillery and S. M. Sherman. Branched thalamic afferents: what are the messages that they relay to the cortex? *Brain Res Rev*, 66:205–219, Jan 2011.

[49] Suzanne Hidi and Judith M Harackiewicz. Motivating the academically unmotivated: A critical issue for the 21st century. *Review of educational research*, 70(2):151–179, 2000.

[50] C. Hutt. Specific and diversive exploration1. *Advances in child development and behavior*, 5:119, 1970.

[51] C. Hutt and R. Bhavnani. Predictions from play. *Nature*, 1972.

[52] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Advances in neural information processing systems*, 18:547, 2006.

[53] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, May 2009.

[54] L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Arxiv preprint cs/9605103*, 1996.

[55] R. Kaplan and S. Kaplan. Cognition and environment: functioning in an uncertain world. *Ann Arbor*, 1983.

[56] Stephen Kaplan. Cognitive maps, human needs and the designed environment 5.4. *Environmental Design Research: Selected papers*, 1:275, 1973.

[57] T.B. Kashdan, M.W. Gallagher, P.J. Silvia, B.P. Winterstein, W.E. Breen, D. Terhar, and M.F. Steger. The curiosity and exploration inventory-ii: Development, factor structure, and psychometrics. *Journal of research in personality*, 43(6):987–998, 2009.

[58] Todd B Kashdan and John E Roberts. Trait and state curiosity in the genesis of intimacy: Differentiation from related constructs. *Journal of Social and Clinical Psychology*, 23(6):792–816, 2004.

[59] Todd B Kashdan, Paul Rose, and Frank D Fincham. Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. *Journal of Personality Assessment*, 82(3):291–305, 2004.

[60] Itamar Kass and Amnon Horovitz. Mapping pathways of allosteric communication in groel by analysis of correlated mutations. *Proteins: Structure, Function, and Bioinformatics*, 48(4):611–617, 2002.

[61] M. Kawato and K. Samejima. Efficient reinforcement learning: computational theories, neuroscience and robotics. *Current opinion in neurobiology*, 17(2):205–212, 2007.

[62] Kyeong Kyu Kim, Hyun Kyu Song, Dong Hae Shin, Kwang Yeon Hwang, Senyon Choe, Ook Joon Yoo, and Se Won Suh. Crystal structure of carboxylesterase from¡ i¿ pseudomonas fluorescens¡/i¿, an α/β hydrolase with broad substrate specificity. *Structure*, 5(12):1571–1584, 1997.

[63] J. Klayman and Y.W. Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, 94(2):211, 1987.

[64] Vasily Klucharev, Kaisa Hytnen, Mark Rijpkema, Ale Smidts, and Guilln Fernndez. Reinforcemeklucharev2009140nt learning signal predicts social conformity. *Neuron*, 61(1):140 – 151, 2009.

[65] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, volume 3, pages 2619–2624. IEEE, 2004.

[66] B T Korber, R M Farber, D H Wolpert, and A S Lapedes. Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences*, 90(15):7176–7180, 1993.

[67] Shulamith Kreitler and Hans Kreitler. Motivational and cognitive determants of exploration. In *Curiosity and exploration*, pages 259–284. Springer, 1994.

[68] Daniel Y. Little and Lu Chen. Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PLoS ONE*, 4(3):e4762, 03 2009.

[69] Daniel Ying-Jeh Little and Friedrich Tobias Sommer. Learning and exploration in action-perception loops. *Frontiers in Neural Circuits*, 7(37), 2013.

[70] G. Loewenstein. The psychology of curiosity: A review and reinterpretation. *psychological Bulletin*, 116(1):75, 1994.

[71] G Lowenstein. The psychology of curiosity. *Psychological Bulletin*, 116(1):75–98, 1994.

[72] R.H. MacArthur and E.R. Pianka. On optimal use of a patchy environment. *American Naturalist*, pages 603–609, 1966.

[73] D.J.C. MacKay and L. Peto. A hierarchical dirichlet language model. *Natural language engineering*, 1(3):1–19, 1995.

[74] Oded Maimon and Shahar Cohen. A review of reinforcement learning methods. *Data Mining and Knowledge Discovery Handbook*, pages 401–417, 2010.

[75] C.D. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. 2008.

[76] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, 2005.

[77] Ernst Mayr. Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, 134(3489):1501–1506, 1961.

[78] Robert R McCrae. Social consequences of experiential openness. *Psychological Bulletin*, 120(3):323, 1996.

[79] Christopher S. Miller and David Eisenberg. Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, 24(14):1575–1582, 2008.

[80] Sanzo Miyazawa and Robert L. Jernigan. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins: Structure, Function, and Bioinformatics*, 34(1):49–68, 1999.

[81] KC Montgomery. Exploratory behavior as a function of" similarity" of stimulus situation. *Journal of Comparative and Physiological Psychology*, 46(2):129, 1953.

[82] A.K. Myers and N.E. Miller. Failure to find a learned drive based on hunger; evidence for learning motivated by" exploration.". *Journal of Comparative and Physiological Psychology*, 47(6):428, 1954.

[83] J.D. Nelson. Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112(4):979, 2005.

[84] Andrew Ng, Adam Coates, Mark Diel, Varun Ganapathi, Jamie Schulte, Ben Tse, Eric Berger, and Eric Liang. Autonomous inverted helicopter flight via reinforcement learning. *Experimental Robotics IX*, pages 363–372, 2006.

[85] Henry W. Nissen. A study of exploratory behavior in the white rat by means of the obstruction method. *The Pedagogical Seminary and Journal of Genetic Psychology*, 37(3):361–376, 1930.

[86] A. Noë. *Action in perception*. the MIT Press, 2004.

[87] M. Oaksford and N. Chater. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4):608, 1994.

[88] Osvaldo Olmea, Burkhard Rost, and Alfonso Valencia. Effective use of sequence correlation and conservation in fold recognition. *Journal of Molecular Biology*, 293(5):1221 – 1239, 1999.

[89] J.K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–972, 2001.

[90] J.K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(05):939–973, 2002.

[91] Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, and Alfonso Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271(4):511 – 523, 1997.

[92] Anthony D Pellegrini, Danielle Dupuis, and Peter K Smith. Play in evolution and development. *Developmental Review*, 27(2):261–276, 2007.

[93] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. Ucsf chimeraa visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[94] E. Pfaffelhuber. Learning and information theory. *International Journal of Neuroscience*, 3(2):83–88, 1972.

[95] David Philipona, J Kevin O'Regan, and J-P Nadal. Is there something out there? inferring space from sensorimotor dependencies. *Neural Computation*, 15(9):2029–2049, 2003.

[96] J. Piaget. *The Origins of Intelligence in Children*. Norton, New York, 1952.

[97] W. Pisula. Costs and benefits of curiosity: The adaptive value of exploratory behavior. *Polish Psychological Bulletin*, 2003.

[98] W. Pisula. Play and exploration in animalsa comparative analysis. *Polish Psychological Bulletin*, 39(2):104–107, 2008.

[99] W. Pisula. *Curiosity and information seeking in animal and human behavior*. Brown Walker Pr, 2009.

[100] Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(suppl 1):D129–D133, 2004.

[101] Scott Proper and Prasad Tadepalli. Scaling model-based average-reward reinforcement learning for product delivery. *Machine Learning: ECML 2006*, pages 735–742, 2006.

[102] Adrian Raine, Chandra Reynolds, Peter H Venables, and Sarnoff A Mednick. Stimulation seeking and intelligence: a prospective longitudinal study. *Journal of Personality and Social Psychology*, 82(4):663, 2002.

[103] M J Renner. Learning during exploration: the role of behavioral topography during exploration in determining subsequent adaptive behavior. *Int J Comp Psychol*, 2(1):4356, 1988.

[104] M.J. Renner. Neglected aspects of exploratory and investigatory behavior. *Psychobiology*, 1990.

[105] F. Rezā. *An Introduction to information theory*. Dover Books on Mathematics Series. Dover Publications, Incorporated, 1961.

[106] P. Rochat. Object manipulation and exploration in 2-to 5-month-old infants. *Developmental Psychology*, 25(6):871, 1989.

[107] Jerome I Rotgans and Henk G Schmidt. Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction*, 21(1):58–67, 2011.

[108] Kazuyuki Samejima, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. Representation of action-specific reward values in the striatum. *Science*, 310(5752):1337–1340, 2005.

[109] M. Sato, K. Abe, and H. Takeda. Learning control of finite markov chains with an explicit trade-off between estimation and control. *Systems, Man and Cybernetics, IEEE Transactions on*, 18(5):677–684, 1988.

[110] Jonathan Schaeffer, Markian Hlynka, and Vili Jussila. Temporal difference learning applied to a high-performance game-playing program. In *Proceedings of the 17th international joint conference on Artificial intelligence-Volume 1*, pages 529–534. Morgan Kaufmann Publishers Inc., 2001.

[111] Ulrich Schiefele, Andreas Krapp, and Adolf Winteler. Interest as a predictor of academic achievement: A meta-analysis of research. *The Role of interest in Learning and Development*, page 183, 1992.

[112] George Shackelford and Kevin Karplus. Contact prediction using mutual information and neural nets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):159–164, 2007.

[113] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[114] B. Si, J.M. Herrmann, and K. Pawelzik. Gain-based exploration: From multi-armed bandits to partially observable environments. 2007.

[115] P.J. Silvia. Interest and interests: The psychology of constructive capriciousness. *Review of General Psychology*, 5(3):270, 2001.

[116] P.J. Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89, 2005.

[117] P.J. Silvia. *Exploring the psychology of interest*. Oxford University Press, USA, 2006.

[118] P.J. Silvia. Interestthe curious emotion. *Current Directions in Psychological Science*, 17(1):57, 2008.

[119] S. Singh, R.L. Lewis, A.G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *Autonomous Mental Development, IEEE Transactions on*, 2(2):70–82, 2010.

[120] Nicholas J. Skelton, Michael F. T. Koehler, Kerry Zobel, Wai Lee Wong, Sherry Yeh, M. Theresa Pisabarro, Jian Ping Yin, Laurence A. Lasky, and Sachdev S. Sidhu. Origins of pdz domain ligand specificity: Structure determination and mutagenisis of the erbin pdz domain. *Journal of Biological Chemistry*, 278(9):7645–7654, 2003.

[121] S. Still. Information-theoretic approach to interactive learning. *EPL (Europhysics Letters)*, 85:28005, 2009.

[122] Peter Stone and Richard S Sutton. Scaling reinforcement learning toward robocup soccer. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 537–544. Morgan Kaufmann Publishers Inc., 2001.

[123] J. Storck, S. Hochreiter, and J. Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *ICANN'95*. Citeseer, 1995.

[124] Richard S Sutton and Andrew G Barto. *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, chapter 12. 1988.

[125] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT Press, 1998.

[126] R.S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems 8*, pages 1038–1044, 1996.

[127] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[128] S.B. Thrun. Efficient exploration in reinforcement learning. *Technicalreport, School of-ComputerScience, Carnegie-MellonUniversity*, 1992.

[129] Elisabeth R.M. Tillier and Thomas W.H. Lui. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, 19(6):750–755, 2003.

[130] N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 1999.

[131] N. Tishby and D. Polani. Information theory of decisions and actions. *Perception-Action Cycle*, pages 601–636, 2011.

[132] Simon A. A. Travers and Mario A. Fares. Functional coevolutionary networks of the hsp70hophsp90 system revealed through computational analyses. *Molecular Biology and Evolution*, 24(4):1032–1044, 2007.

[133] M. Tribus. *Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications*. University series in basic engineering. Van Nostrand, 1961.

[134] M. Vergassola, E. Villermaux, and B.I. Shraiman. infotaxis as a strategy for searching without gradients. *Nature*, 445(7126):406–409, 2007.

[135] ZhengyuanO. Wang and DavidD. Pollock. Coevolutionary patterns in cytochrome c oxidase subunit i depend on structural and functional context. *Journal of Molecular Evolution*, 65(5):485–495, 2007.

[136] Zhongmin Wang and Florante A. Quiocho. Complexes of adenosine deaminase with two potent inhibitors: x-ray structures in four independent molecules at ph of maximum activity,. *Biochemistry*, 37(23):8314–8324, 1998.

[137] H. Wässle, BB Boycott, et al. Functional architecture of the mammalian retina. *Physiological reviews*, 71(2):447, 1991.

[138] Everett Waters and L.Alan Sroufe. Social competence as a developmental construct. *Developmental Review*, 3(1):79 – 97, 1983.

[139] Kurt R. Wollenberg and William R. Atchley. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences*, 97(7):3288–3291, 2000.

[140] Chen-Hsiang Yeang and David Haussler. Detecting coevolution in and among protein domains. *PLoS Comput Biol*, 3(11):e211, 11 2007.

[141] Kevin Y. Yip, Prianka Patel, Philip M. Kim, Donald M. Engelman, Drew McDermott, and Mark Gerstein. An integrated system for studying residue coevolution in proteins. *Bioinformatics*, 24(2):290–292, 2008.