Bayesian Experimental Design in Bioinformatics: Application to Helical Peptide Studies

Joseph E. Lucas[2], Terrence G. Oas[3], Scott C. Schmidler[1]*

Abstract

The use of complex data-derived models is prevalent in bioinformatics and computational biology. When insufficient data is available to adequately determine such models, additional data must be collected. However, such data may be difficult, time-consuming, or expensive to obtain, and different observations often vary in the amount of information they provide towards improving the model. We outline the Bayesian decision-theoretic approach to designing experiments which collect maximally informative data. We develop this approach in the context of a statistical mechanical model for helical peptide folding, and provide Monte Carlo approaches for efficient calculation of required quantities. We demonstrate this approach by designing helical peptide studies aimed at maximally improving predictive performance of a state-of-the-art helix-coil model, selecting polypeptides for further experimental study under a variety of design constraints and utility measures. The identified peptides highlight areas where properties of helical peptide folding are insufficiently understood based on current experimental data in the protein science literature.

Keywords: Experimental design, Bayesian methods, helix-coil theory, protein folding, helical peptides

Running head: Bayesian Experimental Design

[1]Corresponding author: Email: schmidler@stat.duke.edu; Ph: (919) 684-8064; Fax: (919) 684-8594; Mailing address: Department of Statistical Science, 223 Old Chemistry Building, Box 90251, Duke University, Durham, NC 27708-0251

[2]Email: jel2@stat.duke.edu; Ph: (919) 684-8753; Fax: (919) 684-8594; Mailing address: Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708-0251

[3]Email: oas@biochem.duke.edu; Ph: (919) 684-4363; Fax: (919) 681-8862; Mailing address: Department of Biochemistry, 436 Nanaline Duke Building, Box 3711, Duke University Medical Center, Durham, NC 27710

1 Introduction

Probabilistic and statistical models are now a staple of bioinformatics and computational biology research. A wide variety of specialized models have been developed for problems including biological sequence analysis (Churchill, 1989; Krogh et al., 1994; Durbin et al., 1998; Liu and Lawrence, 1999) molecular structure analysis (Sippl, 1995; Wu et al., 1998; Schmidler, 2006), sequence-structure prediction (Eddy and Durbin, 1994; Schmidler et al., 2000, 2004; Lathrop et al., 1998), biological networks (Friedman, 2004), and tissue classification from mRNA expression (too many to mention), to name just a few. Such models are typically derived from a combination of biological knowledge, computational considerations such as existence of efficient algorithms, and statistical selection criteria such as fit to an existing data set or out-of-sample prediction accuracy. Parameters of these models are typically obtained by some manner of fitting to one or more experimental datasets.

When such models fail to perform adequately it may be a failure of the model structure itself, or simply the lack of sufficient data to adequately estimate all model parameters. In the latter case, we can consider improving the model by gathering new data, even running new experiments. However, not all data are created equal from the perspective of model improvement, and since data are often expensive, time-consuming, or risky (in the case of clinical data) to generate, we would like to maximize the information obtained while minimizing the associated costs. Principles of statistical experimental design may be brought to bear on this problem.

While a variety of statistical parameter estimation and statistical model selection criteria are well known and frequently used in the bioinformatics literature, statistical experimental design has received relatively little attention. A rare exception is in certain mRNA expression studies where (Kerr and Churchill, 2001) elucidate design principles; however this approach is somewhat different from the Bayesian model-based approach to experimental design de-

scribed here. Almost nothing appears in the bioinformatics literature on experimental design in the contexts of sequence analysis and structure analysis.

In this paper we review the fundamental principles of Bayesian experimental design for general statistical models. We then develop this approach for a predictive sequence-structure model derived from biophysical statistical mechanics, to which we have previously applied statistical parameter estimation and model selection techniques (Schmidler et al., 2007; Lucas, 2006). As described in Schmidler et al. (2007), this model has many similarities to hidden Markov models, which are widely utilized in many areas of bioinformatics and computational biology. We use this example to demonstrate the advantages and challenges of doing Bayesian design in bioinformatics-type models. We describe a Monte Carlo importance reweighting algorithm which can be used to dramatically decrease the computational expense of Bayesian experiment design calculations in more general models. Finally, we report the results of several design calculations using the helix-coil model example, where the polypeptide sequences selected for study highlight areas of helical peptide folding which are insufficiently understood based on currently available experimental data.

2 Bayesian experimental design

Statistical experimental design dates to the early days of formal statistical theory (Fisher, 1935). Non-Bayesian experimental design methodology is typically based on a series of related "alphabetical" optimality criteria (Box, 1982). For example, D-optimal designs involve maximization of the determinant of the Fisher information matrix. Sequential D-optimal designs (Wynn, 1970) are widely used used in applications; see e.g. Coffey et al. (2005); Berger (1994); Fujiwara et al. (2005) for recent examples.

The Bayesian approach to experimental design advocated in this paper is based on decision theory and the maximization of expected utility (Lindley, 1992). Given a set of possible actions $\mathcal{A}$, outcomes $\mathcal{O}$, and a *utility function* $\mathcal{U}(o) : \mathcal{O} \to \mathbb{R}$ assigning values to each outcome, decision theory dictates that we choose the action which *maximizes expected utility*:

$$A^* = \arg\max_{A \in \mathcal{A}} \int_{\mathcal{O}} \mathcal{U}(O \mid A) P(dO \mid A, I)$$

where I represents the current information available to the decision maker, before taking any action. Expected utility decision making has a long tradition and forms a foundational basis for statistics (Savage, 1954; Berger, 1985), as well as theories of rational behavior in economics (Raiffa, 1968) and artificial intelligence (Horvitz et al., 1988; Russell and Norvig, 2003).

Experimental design involves a two-stage decision process: first an experiment is selected among the many possibilities; the experiment is then performed and data collected; finally, the observed data are used to inform some additional decision. Thus in applying decision-theoretic reasoning to experimental design, the principle of maximizing expected utility is applied twice. In the first decision, the action set $\mathcal{A}_1$ is the set of possible experimental designs, or possible sets of measurements to be performed.[1] The outcomes $\mathcal{O}_1$ associated

[1] We limit our discussion to *sequential design*, where measurements are taken one at a time, and I is

with the first decision are the set of possible experimental observations. However, the goal of experimentation is to collect information, and the actions and outcomes of the second decision depend on to what use that information is to be put. Indeed, one of the strengths of the Bayesian approach to design is that it forces the experimenter to be explicit about the goal(s) of the experiment.

When applying design principles to statistical models, the action set $\mathcal{A}_2$ is often taken to be the set of possible choices for model parameter estimates $\hat{\theta} \in \Theta$, with O_2 being the true (unknown) parameter value $\theta \in \Theta$; alternatively, $\mathcal{A}_2$ may be a set of predictions for some unknown quantity or future event, with O_2 the true value. In each case, I_2 is represented by a posterior distribution $p(\theta \mid O_1)$ which incorporates the data collected as a result of the first (experiment choice) decision as well as any additional prior information or constraints. In such statistical estimation or prediction problems, $U(\theta)$ is commonly represented by a (negative, expected) *loss function* $L(\hat{\theta}, \theta)$ (see e.g Berger (1985)).

Applying expected utility maximization to this two-stage design problem yields the Bayesian optimal design:

$$A^* = \arg \max_{A_1 \in \mathcal{A}_1} \int_{\mathcal{O}_1} P(dO_1) \max_{A_2 \in \mathcal{A}_2} \int_{\mathcal{O}_2} \mathcal{U}(O_2 \mid A_2) P(dO_2 \mid A_2, I_2)$$

For a given statistical model, the key steps in Bayesian experimental design are thus to identify the set of possible experiments and their respective outcome possibilities, and to choose a utility or loss function. This approach is quite general and in fact gives rise to many non-Bayesian experimental design criteria (e.g. alphabetical optimality criteria such as D-optimality) as special cases under the choice of specific utility functions (Bernardo, 1979; Chaloner and Verdinelli, 1995). A comprehensive survey of Bayesian experimental design is

updated to reflect the new data obtained from an experiment before the next experiment is chosen. The Bayesian framework applies without modification to batch (multi-experiment) design, although computation of the optimal design is generally more difficult.

given in Chaloner and Verdinelli (1995); see also (Clyde, 2001; Bernardo and Smith, 1994).

3 The helix-coil model for peptides

The helix-coil model of state transitions in polymers has a long history (see Poland and Scheraga (1970); Qian and Schellman (1992) for reviews), and was pioneered for the study of α-helix formation in protein folding by Scholtz and Baldwin (1992), followed by significant additional work (Qian and Schellman, 1992; Doig et al., 1994; Shalongo and Stellwagen, 1995; Stapley et al., 1995; Andersen and Tong, 1997). Building on the work of Munoz and Serrano (1994), Schmidler et al. (2007) developed a predictive statistical mechanical model for the helix-coil transition in polypeptides. Let $R = (R_1, \ldots, R_l)$ denote the amino acid sequence of a peptide of length l, and $X = (x_1, \ldots, x_l)$ an associated vector of binary indicators with $x_i = 1$ if the i^{th} amino acid is in helical conformation and 0 otherwise.

The model is a Gibbs random field with short-range neighborhood interactions, with potential $U(X, R)$ given by

$$U(X,R) = \sum_{i=1}^{l} x_i \Delta G_{R_i} + \sum_{i=1}^{l-3} x_{i:i+3} \Delta\Delta G^3_{R_i R_{i+3}} + \sum_{i=1}^{l-4} x_{i:i+4} \Delta\Delta G^4_{R_i R_{i+4}} \quad (1)$$

where $x_{i:k} = \prod_{j=i}^{k} x_i$, individual side chain helical propensities are described by free energies

$$\Delta G_{R_i} = \Delta H x_{i-1:i+1} - T \Delta S_{R_i} x_i \quad (2)$$

and the $\Delta\Delta G$ terms in (1) represent energetic contributions of $i, i+3$ and $i, i+4$ side chain-side chain interactions. Although we have suppressed the dependence in the notation above

for clarity, in fact ΔH and ΔS_R are functions of both temperature

$$\Delta H(T) = \Delta H_0 + \Delta C_p(T - T_0)$$
$$\Delta S(T) = \Delta S_0 + \Delta C_p \log(T/T_0)$$

and pH

$$\Delta S_R \;=\; f_+ \Delta S_{R^+} + (1 - f_+)\Delta S_{R^-}$$

where $f_+ = (1 + 10^{pH-pK})^{-1}$ represents the fraction protonated for ionizable sidechains.

The resulting Boltzmann-Gibbs distribution over conformations is given by

$$P(X \in \mathcal{X} \mid R) = Z^{-1} \mathrm{e}^{-\frac{1}{k_B T} U(X,R)}$$

where Z is the normalizing constant or partition function involving a sum over all configurations $X \in \mathcal{X} = \mathbb{Z}_2^l$. The *helicity* of a peptide is then given by the expectation or ensemble average

$$\mathcal{H}(R) = \sum_{X \in \mathcal{X}} h(X) P(X \mid R)$$

where $h(X) = l^{-1} \sum_{i=1}^{l} x_i$, and it is this ensemble average quantity which can be compared with experimental measures of peptide helical content, such as those obtained by circular dichroism (CD).

Additional details of the model, including energetic terms for inclusion of N- and C-terminal blocking groups, capping effects, and positional parameters, and efficient algorithms for calculation of the partition function and helicity, are described in Schmidler et al. (2007).

3.1 Bayesian inference

Bayesian estimation of the parameters of the above helix-coil model (including ΔH; ΔS_R for all R; $\Delta\Delta G^3_{R_1R_2}$ and $\Delta\Delta G^4_{R_1R_2}$ for all R_1, R_2; ΔC_p; and other parameters) is described in Schmidler et al. (2007). Using θ to denote the vector of all of these parameters, we assign prior distributions $\pi_0(\theta)$ to all parameters, and given a set of experimental data denoted by D consisting of pairs of sequences and associated experimental (CD) helicities $\{R_i, \tilde{h}_i\}_{i=1}^n$, we obtain the posterior distribution

$$\begin{aligned}\pi(\theta \mid D) &= \frac{\sum_{\mathbf{x}} P(\mathbf{R}, \mathbf{x}, \mathbf{h} \mid \theta) P(\theta)}{\sum_{\mathbf{x}} P(\mathbf{R}, \mathbf{x}, \mathbf{h})} \\ &\propto \pi_0(\theta)(2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \mathrm{e}^{-\frac{1}{2\sigma^2}\left(\tilde{h}_i - \mathcal{H}(R_i,\theta)\right)^2} \end{aligned} \tag{3}$$

which corresponds to the additive noise model $\tilde{h}_R = \mathcal{H}(R, \theta) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. Informative priors based on the protein science literature and model selection shrinkage priors are described in detail in Schmidler et al. (2007) and Lucas (2006), but we do not focus on this here. Instead we consider the role of (3) in performing Bayesian experimental design calculations for the helix-coil model. Inference under the posterior distribution (3) is performed via a Markov chain Monte Carlo simulation algorithm described in Schmidler et al. (2007), yielding a finite set of samples $\theta^{(i)} \sim \pi(\theta \mid D)$, $i = 1, \ldots, m$ from which we may easily obtain most posterior quantities of interest by Monte Carlo approximation of expectation integrals:

$$\begin{aligned} E_\pi(f(\theta)) &= \int f(\theta)\pi(\theta \mid D) d\theta \\ &\approx \frac{1}{m} \sum_{i=1}^{m} f(\theta^{(i)}) \end{aligned} \tag{4}$$

for some function $f(\theta)$. (See Gilks et al. (1996) for an introduction to Markov chain Monte Carlo methods.)

The *posterior predictive distribution* for the helicity of a new observation R_* is given by the mixture distribution

$$q(h_* \mid R_*, D) = \int P(h_* \mid R_*, \theta)\pi(\theta \mid D)d\theta \tag{5}$$

which involves integration with respect to the posterior distribution of the parameters θ and accounts for the remaining uncertainty in these model parameters. Prediction of helicity based on this posterior predictive distribution is our second-stage decision problem: if we measure prediction utility for predictor $\hat{h}$ using the (negative) *squared loss*:

$$L(\hat{h}(R_*), h_*) = (\hat{h}(R_*) - h_*)^2$$

then the predictor which *maximizes expected utility* or *minimizes Bayesian/posterior expected loss* is given by the posterior predictive mean:

$$\begin{aligned} \hat{h}(R_*) = E_q(h_*) &= E_\pi(E(h_* \mid \theta)) \\ &= \int \mathcal{H}(R_*, \theta)\pi(\theta \mid D)d\theta = \mu_{\mathcal{H}(R_*)} \end{aligned} \tag{6}$$

The predictor (6) is approximated by

$$\hat{\mu}_{\mathcal{H}(R_*)} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{H}(R_*, \theta^{(i)})$$

from Monte Carlo samples $\theta^{(i)} \sim \pi(\theta \mid D)$ as in (4) above.

4 Bayesian design in the helix-coil model

We now demonstrate the Bayesian decision-theoretic approach to experimental design, by applying it to select experiments aimed at improving the helix-coil model described in Section 3.

4.1 Helical peptide studies via circular dichroism

The vast majority of helical peptide studies are performed via circular dichroism (CD). Peptides are synthesized or isolated from natural protein digests and purified. The CD spectrum of a peptide residue is sensitive to its backbone conformation, which determines the optical activity of the chromaphore environment. The spectrum of a helical residue has been observed to contain a strongly negative peak at 222nm, and the relative intensity of the peak at this wavelength is commonly used to estimate the helical content of polypeptides (Scholtz and Baldwin, 1992). A large number of such studies have been performed on a wide variety of synthetic peptides and naturally-occurring protein fragments, due to their wide use as model experimental systems for studying protein folding.

The helix-coil model of Section 3 applies in principle to all peptides of various lengths. (In practice the model fails to account for long-range interactions observed in longer polypeptides, as well as non-monomeric peptides which may dimerize or aggregate.) Thus there are an exponential number of potential polypeptides which we could study experimentally to obtain additional data to improve the model. Due to the time, effort, and expense involved in such studies, we wish to guide future studies by choosing those experiments which are most likely to provide maximal additional information on top of currently available data. In particular, we wish to improve the predictive accuracy (and reduce predictive uncertainty) of the helix-coil model, enabling peptides of future interest to be predicted as accurately and precisely as possible. We may also be interested in clarifying the role of specific energetic

contributions or interactions by reducing uncertainty about the corresponding model parameters. We approach this goal of selecting maximally informative experiments by applying the principles of Bayesian decision-theoretic experimental design described in Section 2 to the statistical helix-coil model described in Section 3, under the parameter posterior distributions obtained using all currently available data (Schmidler et al., 2007).

4.2 Basic set-up

To apply the Bayesian design framework of Section 2 to helix-coil studies, we must specify the set of possible actions, outcomes, and utilities for the first-stage decision. For convenience we will restrict ourselves to sequential design, so the possible actions are simply the possible peptides R_* that we may consider synthesizing and studying experimentally via CD. Denote this design set by $\mathcal{P}$. Note that extension to batch design, where $\mathcal{P}$ is a set of subsets of possible peptides, is conceptually straightforward, but will demand a combinatorial increase in computational effort.

The set of possible outcomes is also clear: it is the set of possible experimental results $\tilde{h}_* \in [0, 1]$ obtained for the peptide chosen, along with the corresponding state of the model after incorporating the new results, represented by a posterior probability distribution on the helix-coil model parameters which we denote by $\pi(\theta \mid D \cup \tilde{h}_*)$.

Finally, we require a utility function on this outcome space. Since we are interested in improving out-of-sample predictive accuracy of the model, a natural choice is to define utility as the (negative) expected predictive loss. However, we must now be precise about which population of peptides we are interested in predicting. Let $\mathcal{R}$ denote the set of all possible polypeptides, which is infinite $|\mathcal{R}| = \bigcup_{l=1}^{\infty} 20^l$. Let $\omega(R)$ denote a distribution on $\mathcal{R}$. Defining the population $\omega(R)$ of interest is up to the experimenter; it may be as large as the set of all naturally occurring peptides, or as small as sequence variations on a specific

peptide of interest; it may be peptides of a certain length range, or only peptides which arise as fragments of naturally occurring proteins, or highly helical peptides only, or those whose helicity is stable under temperature or pH changes. Further, many peptides in $\mathcal{R}$ will form aggregates and $\omega(R)$ may be chosen to assign zero weight to predicting such peptides. The choice of $\omega(R)$, and in particular the range of peptides assigned non-zero weight, may impact the difficulty of the resulting optimization problem as described below.

As before, let h denote the true helicity of peptide R, and let $\hat{h}_D(R)$ denote our predicted helicity for R as given by (6), where the dependence on the data D through the posterior distribution over model parameters is now made explicit. We use the shorthand π_D to denote the posterior distribution $\pi(\theta \mid D)$ given in (3). Using squared loss $L(\hat{h}, h) = (h - \hat{h})^2$, we define our utility function U as the negative *expected prediction loss* over the predictive population $\omega(R)$ of interest:

$$\begin{aligned} U(D) &= E_\omega \left[E_{h_R|\pi_D} \left(L(\hat{h}_D(R), h_R) \right) \right] \\ &= \sum_{R \in \mathcal{R}} \omega(R) \int L(\hat{h}_D(R), h_R) q(h_R \mid R, D) dh_R \\ &= \sum_{R \in \mathcal{R}} \omega(R) \int \int L(\hat{h}_D(R), h_R) P(h_R \mid \theta) \pi(\theta \mid D) d\theta dh_R \\ &= \sum_{R \in \mathcal{R}} \omega(R) \mathrm{var}_{\pi_D}(h_R) = E_\omega \left[\mathrm{var}_{\pi_D}(h_R) \right] \end{aligned} \tag{7}$$

where $q(h_R \mid R, D)$ is the predictive distribution (5). Note that only the last line is specific to quadratic loss, and although we assume this below for ease of exposition, the approach holds for arbitrary loss functions. Since $U(D)$ is an expectation over $\mathcal{R}$, when $\mathcal{R}$ is too large to enumerate we may simply approximate $U(D)$ by Monte Carlo sampling from $\omega(R)$ in a manner similar to (4).

The *expected* utility to be obtained by studying R_* then becomes

$$\begin{aligned} EU(R_*) &= E_{h_{R_*}|D}\left(U(D \cup h_*)\right) \\ &= \int U(D \cup h_*) q(h_* \mid R_*, D) dh_* \\ &= E_\omega \int \left[\mathrm{var}_{\pi_{D_*}}(h_R) q(h_* \mid R_*, D) dh_*\right] \end{aligned} \tag{8}$$

where $\pi_{D_*} = \pi(\theta \mid D \cup h_*)$. ($EU(R)$ is sometimes called the *pre-posterior* expected utility, see e.g. Clyde (2001)). The design problem is therefore reduced to a computational problem of maximizing $EU(R)$ over all $R \in \mathcal{P}$.

In practice, improvement of predictive accuracy is not the only consideration when choosing when experiments to perform; relative time and cost may also play a role. For example, synthesis of longer peptides may be more difficult or expensive. Similarly, producing 10 point mutants of a single peptide is generally easier experimentally than synthesizing 10 unrelated peptides. Let $EC(R_i)$ denote the expected *cost* associated with the experiment to study peptide R_i. Then our expected utility becomes

$$EU(R) - EC(R)$$

Since costs may vary from lab to lab and because incorporating cost is only a trivial modification of the computation of expected utility, for examples in this paper we use $EC(R_i)$ constant and so suppress the notation in what follows.

4.3 Importance sampling

Calculation of (7) requires the evaluation of the expected prediction error $\Lambda_{R,D} = \mathrm{var}_{\pi_D}(h_R)$ for every peptide R in the prediction set $\mathcal{R}$, or perhaps for some Monte Carlo sample

$R_1, \ldots, R_m \sim \omega(R)$. Since $h_R \mid \theta \sim N(\mathcal{H}(R,\theta), \sigma^2)$ and recalling that $\sigma \in \theta$, we have

$$\begin{aligned}
\Lambda_{R,D} &= \int \left[\int (h_R - \hat{h}_R)^2 \phi \left(\frac{h_R - \mathcal{H}(R,\theta)}{\sigma} \right) dh_R \right] \pi(\theta \mid D) d\theta \\
&= \int \left[(\mathcal{H}(R,\theta) - \hat{h}_R)^2 + \sigma^2 \right] \pi(\theta \mid D) d\theta \\
&= \mathrm{var}_{\pi_D}(\mathcal{H}_R) + E_{\pi_D}\sigma^2 = \mathrm{var}_{\pi_D}(h_R)
\end{aligned}$$

Given samples $\theta^{(i)} \sim \pi(\theta \mid D)$ as described above, we have the Monte Carlo approximation

$$\begin{aligned}
\Lambda_{R,D} \approx \hat{\Lambda}_{R,D} &= \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\mu}_{\mathcal{H}(R)} - \mathcal{H}(R, \theta^{(i)}) \right)^2 + \sigma^{2(i)} \\
&= \hat{\sigma}^2_{\hat{h}_R} + \hat{\sigma}^2
\end{aligned}$$

We must then integrate this quantity, which depends on D through π_D, over the predictive distribution of the measurement h_* to be taken on the peptide R_*, in order to obtain the expected utility (8). Because $\mathrm{var}_{\pi_{D_*}}(h_R)$ depends on the unknown observation h_*, integrating numerically by quadrature or Monte Carlo requires generating samples $\theta^{(i)} \sim \pi(\theta \mid D_*)$ for each distinct value of h_* considered. However, sampling from the posterior $\pi(\theta \mid D)$ for some D involves a large MCMC simulation requiring multiple days (Schmidler et al., 2007). Thus repeating this for sufficiently many values of h_{R_*} to evaluate the integral (8), and doing so for a large number of possible R_*'s in order to maximize (7) over the design set $\mathcal{R}$, is prohibitively expensive.

Alternatively, if π_D and π_{D_*} are "sufficiently close" in the sense that the observation h_* does not dramatically shift the posterior mass into the tails of π_D, then we may achieve a dramatic savings in computational time by *importance reweighting*: given samples $\theta^{(i)} \sim \tilde{\pi}(\theta \mid D)$, $i = 1, \ldots, m$ drawn from some distribution $\tilde{\pi}(\theta) \neq \pi(\theta)$, we may approximate

expectations under $\pi(\theta)$ by replacing (4) with

$$E_\pi(f(\theta)) \approx \frac{\sum_{i=1}^m w_i f(\theta^{(i)})}{\sum_{i=1}^m w_i}$$

where $w_i = \pi(\theta^{(i)})/\tilde{\pi}(\theta^{(i)})$ are the *importance weights*. Applying this to the samples $\theta^{(i)}$ drawn from π_D, we get:

$$\Lambda_{R,D_*} \approx \hat{\Lambda}^w_{R,D_*} = \sum_{i=1}^m \frac{w_i}{w} \left[\left(\hat{\mu}^w_{\mathcal{H}(R)} - \mathcal{H}(R, \theta^{(i)}) \right)^2 + \sigma^{2(i)} \right]$$

where

$$\hat{\mu}^w_{\mathcal{H}(R)} = \sum_{i=1}^m \frac{w_i}{w} \mathcal{H}(R, \theta^{(i)}) \qquad \text{and} \qquad w = \sum_{i=1}^m w_i$$

and

$$\begin{aligned} w_i = \frac{\pi_{D_*}(\theta^{(i)})}{\pi_D(\theta^{(i)})} &= \mathcal{L}(R_*, h_*; \theta^{(i)}) \\ &= \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \mathrm{e}^{-\frac{1}{2\sigma^{2(i)}}(h_* - \mathcal{H}(R_*, \theta^{(i)}))^2} \end{aligned}$$

Thus (9) can be calculated very quickly without the need to sample directly from π_{D*}. Note that the assumption that $\pi_D \approx \pi_{D_*}$ is somewhat suspect here, since we are searching for peptides R_* which will have the greatest effect on the posterior; however the examples in Section 5 show that this assumption can work well in practice.

To calculate (8), we must integrate (7) over the predictive distribution (5). This may easily be done by Monte Carlo integration as well, by drawing $h_*^{(i)}$ from $q(\cdot \mid R_*, D)$ for each $\theta^{(i)}$. However, because it is a one-dimensional integral, direct quadrature is more efficient. Given the $\theta^{(i)}$'s, we have the mixture density approximation

$$q(h_* \mid R_*, D) \approx \frac{1}{m} \sum_{i=1}^m \phi\left(\left(h_* - \mathcal{H}(R, \theta^{(i)}) \right) / \sigma^{(i)} \right)$$

leading to the numerical integral

$$\begin{aligned} EU(R_*) &= E_\omega \int \Lambda_{R_*,D_*} q(h_* \mid R_*, D) dh_* \\ &\approx E_\omega \int \frac{\hat{\Lambda}^w_{R_*,D_*}}{m} \sum_{i=1}^{m} \phi\left(\frac{h_* - \mathcal{H}(R, \theta^{(i)})}{\sigma^{(i)}}\right) dh_* \end{aligned}$$

The expected change in prediction error due to studying peptide R_* is then

$$\begin{aligned} &\sum_{R \in \mathcal{R}} \omega(R) \left(\Lambda_{R,D} - \Lambda_{R,D_*}\right) \\ &\approx \sum_{R \in \mathcal{R}} \omega(R) \left(\hat{\Lambda}_{R,D} - \int \sum_{i=1}^{m} \phi\left(\frac{h_* - \mathcal{H}(R, \theta^{(i)}}{\sigma^{(i)}}\right) \hat{\Lambda}^w_{R,D_*} dh_*\right) \end{aligned}$$

If the predictive set $\{R \in \mathcal{R} : \omega(R) > 0\}$ contains k peptides and the design set is of size $|\mathcal{P}| = p$, then $\sum_i w_i \mathcal{H}(R, \theta^{(i)})$ must be calculated kp times for each of the t quadrature points in the numerical integration over h_*, taking $O(kpmt)$ time overall.

5 Examples

We demonstrate the Bayesian design methodology described above for helical peptides on several examples. For convenience, in this section we let $\mathcal{R}$ denote the prediction set of peptides, which may in practice be a Monte Carlo sample from $\omega(R)$. Recall that $\mathcal{P}$ denotes the design set of peptides under consideration for possible experimental study.

5.1 Example: Individual Peptides

In the first example, we obtained a random subsample $\mathcal{R}_S$ of size 220 from the database of peptides described in Schmidler et al. (2007). For each peptide $R \in \mathcal{R}_S$, we performed a design calculation using $\mathcal{R} = \{R\}$ as the prediction set, and $\mathcal{P} = \mathcal{R}_S$ as the design set. That is, we consider the effect of studying any single peptide in $\mathcal{R}_S$ on the predictive accuracy for peptide R.

The expected utilities resulting from these 220 design calculations are shown as a single matrix in Figure 1, with peptides sorted alphabetically by sequence. The strong diagonal reflects the intuitively obvious fact that the greatest improvement in predictive accuracy for a particular peptide is typically obtained by studying the peptide itself. The block-diagonal patterns reflect the sequence similarity implied by the alphabetical ordering, and the presence of some peptides in the random sample which may have identical sequence but distinct temperature or pH, or be point mutations of one another.

[Figure 1 about here.]

The large off-diagonal values are more interesting. For example, one of these (circled in red) corresponds to the peptide 'PANLKALEAQKQKEQR', where studying 'Y(AEAAKA)^{8}F' yields an expected prediction improvement for 'PANLKALEAQKQKEQR' of .10, compared to .11 for studying the peptide itself. Predicted helicities for these two peptides both depend on several model parameters: the temperature dependence, the individual amino acid

parameters for A,E, and K, and the $i, i+3$ interaction parameters K-E, E-K, K-R, and L-L. Of these, K-E and E-K occur seven and eight times respectively in the longer polypeptide. Figure 2 shows histograms of the posterior distributions for all side chain interaction model parameters which appear in 'PANLKALEAQKQKEQR'. Both L-L and K-R have maximum a-posteriori estimates of zero and significant mass both above and below zero. Thus, 'Y(AEAAKA)^{8}F' contains multiple repetitions of both of the non-zero interactions present in 'PANLKALEAQKQKEQR'.

[Figure 2 about here.]

In fact, there is a temperature curve in the prediction set associated with 'Y(AEAAKA)^{8}F'. Returning to the full database of 1085 peptides and querying this temperature curve against the prediction set containing only 'PANLKALEAQKQKEQR' we see that the amount of information about model parameters that is available in a particular polypeptide shows a striking temperature dependence (see Figure 3). The helicity of this peptide along the temperature curve ranges from 10% at the highest temperature to 92% at the lowest temperature, with the largest amount of information about the parameters available at around 50% helicity (at a temperature of 311). The Bayesian design approach tells us not only which peptides are most informative to study, but other aspects of the experimental set up such as temperature as well.

[Figure 3 about here.]

5.2 Example: Point Mutations

Naturally-occurring peptides often have low intrinsic helicity. Because helix formation can be a bottleneck in protein folding, the relative differences in these small helicities can have a significant impact on protein folding mechanisms (Burton et al., 1998). Thus it is of particular interest to improve the predictive accuracy of the model for peptides in the low helicity

range. To address this, we select a prediction set $\mathcal{R}$ from our database containing only those peptides with measured helicity $< .3$ and temperature 273-280K (nearly all peptides have low helicity at high temperatures). This yields $|\mathcal{R}| = 335$ peptides in our prediction set. As our design set, we consider host-guest studies on the neuropeptide Y analog 'APAELKA-AXAAFKRHGPY' (Petukhov et al., 1996) consisting of point mutations at the X position, measured at pH of either 4, 7, or 10. This gives a design set of size $|\mathcal{P}| = 60$.

Figure 4 shows the expected utility to be obtained by studying each peptide in the design set, for each peptide in the prediction set.

[Figure 4 about here.]

One peptide which stands out is at position 298 in Figure 4. Peptide 298 of the prediction set is 'YGKFRFEQQKKEKEARKK', which contains a number of parameters in common with the design host. This includes ΔS_r parameters for E,G,Y,K,R,A, and F, as well as potential $i, i+3$ and $i, i+4$ parameters involving 'X'.

It is also interesting that in general there appears to be more information to be gained by performing this experiment at higher pH: of the 20 point mutants, fifteen yield their maximum expected utility at pH 10. This may be attributed to the potential gain of information regarding ΔS_K and $\Delta\Delta S_K$, ΔS_Y and $\Delta\Delta S_Y$, and the interaction parameters $\Delta\Delta G_{KX}^{i,i+3}$ and $\Delta\Delta G_{XK}^{i,i+4}$. In particular, peptide 124 of the prediction set ('YGGKAVAAKAVAAKAVAAK') shows large expected improvement when the design peptide is studied at pH 10, but significantly less at pH 4 or 7. This peptide contains several Lysines (pK_a just over 10), and a Tyrosine (pK_a just under 10), so the parameters $\Delta\Delta S_Y$, $\Delta\Delta S_K$, ΔpK_K and ΔpK_Y will have a more significant effect on the helicity of this polypeptide at higher pH (due to the relative fractions of protonated versus unprotonated side chains), and three of these occur in the design peptides. Of the five amino acid substitutions that are not best studied at high pH, 3 of them (P, G, and M) show no discernable difference at different pH's.

The two mutants offering the lowest gain in expected utility are 'X=G' and 'X=P' (peptides 1 and 2 of the design set in Figure 4). Although we are targeting prediction of low helicity peptides, there appears to be little to learn from studying these well-known helix-breakers, presumably because these sidechains do not participate in interactions, and their individual ΔS parameters are already sufficiently well-determined to have large values.

Maximizing expected utility over the entire prediction set leads to choice of the design peptides 'X=I' and 'X=L' (peptides 5 and 6 in Figure 4). Both are hydrophobic sidechains and have potential $i, i+3$ interactions with Phe and $i-4, i$ interactions with Leu. The remainder of the top five desings are listed in table 1.

[Table 1 about here.]

Table 2 shows the number of peptides in the prediction set containing hydrophobic interactions in these positions relative to 'F' and 'L'; all five amino acids in the hydrophobic interaction parameter group are shown.

[Table 2 about here.]

Interactions involving Leu and Ile are seen to be the most abundant, suggesting that reducing uncertainty in these parameters will contribute to improving prediction of the largest number of prediction set peptides. This also serves to demonstrate the impact of choice of prediction set on the resulting optimal design.

5.3 Helicity prediction variability

One might expect that the optimal choice of peptides to study would simply be those with high uncertainty in their predicted helicity under the model (predictive variance), or those containing amino acids or interactions for which the model parameters have high uncertainty

(posterior variance). In that case, such criteria could be used for heuristic design of experiments. However, that turns out not to be the case: the full Bayesian design framework utilizes additional important information in choosing peptides.

To demonstrate this, we estimated the fraction of each peptide's predictive variance which can be attributed to a particular parameter, by fixing all other parameters at their posterior means and calculating the resulting predictive variance integrating the free parameter over its posterior marginal distribution. Figure 5 shows that this does not correlate well with the design calculation of Section 5.2.

[Figure 5 about here.]

For example, there appears to be significant uncertainty associated with parameter ΔS_K, which might seem to suggest that substituting Lys into the host peptide would provide a high level of information. However, such a design criteria fails to account for the details of the host peptide: the peptide 'APAELKAAKAAFKRHGPY' contains both $i, i+3$ and $i, i+4$ 'K-K'; as a result, the Lys guest peptide is significantly less helical (14% versus 21% when Alanine is substituted in its place) and therefore provides little information about any parameters which contribute to helicity. Choosing Lys because of the uncertainty associated with ΔS_K parameter is therefore a poor choice. Similarly, the design calculation tells us that Pro and Gly (well known helix breakers) are the two worst choices for the guest peptide.

The parameter contributing the greatest variability to predicted helicities in the database is ΔH. This is not surprising since ΔH contributes to all peptides regardless of sequence. However, it also has high posterior correlation with ΔS_A, since Ala appears in nearly all peptides as well. In the presence of such correlation, our procedure of fixing all parameters except one will tend to significantly overestimate the contribution to prediction variability; it is for exactly this reason that we see a high estimate for ΔS_A in Figure 5. The ΔS_T parameter is affected by a similar problem, as it has a posterior correlation of .72 with

$\Delta\Delta S_T$.

5.4 Example: Protein folding kinetics of λ-repressor

The λ-repressor protein is critical in the lifecycle of Enterobacteria phage λ, a virus that infects E. Coli. It has been studied extensively by one of us (TGO) as a model system for the protein folding kinetics. The native conformation of λ-repressor contains five α-helices, and previous work (Burton et al., 1998) has shown that the helicity of these peptides in the unfolded state has a significant impact on the overall folding rate of the protein. As such, it is of particular interest to be able to accurately predict the (relatively low) helicity of these peptides and their variants in solution.

Although studying variants of the λ-repressor helices themselves may be expected to provide the most information, it requires synthesizing each of the five α-helical peptides and mutations thereof. An alternative is to consider a single host peptide, and perform host-guest studies which will provide the most improvement in the model's predictive accuracy on *all five* λ-repressor helical peptides simultaneously.

We generated a random sample of 2225 single, double, and triple point mutants of the five λ-repressor helix sequences to use as our prediction set. The sample was generated as follows: For each sample, select one of the five helical regions of λ-repressor uniformly at random. Then select with equal probability a definition of that helix as defined in either Burton et al. (1998) or Marqusee and Sauer (1994) (these differ on the endpoints of three of the five helicies). We then choose to use single, double, or triple point mutations with probabilities 60%, 25%, and 15% respectively. Finally, we choose the appropriate number of positions and make substitutions at random from among the 20 amino acids. After generating 3000 such polypeptides and removing duplicates, we are left with 2225 mutants of the α-helicies in the λ-repressor protein, which we define to be our prediction set. This approximates our

goal of choosing the experiments which will maximally improve prediction performance on variants of the λ-repressor peptides.

As our design set, we consider all 8000 host/guest peptides from the host sequence 'AEAAAxyAAzAAAKA'. We then apply the Bayesian design approach of Section 4 to select the sequences in this design set to study experimentally in order to maximally improve the model accuracy on the prediction set. For comparison, the design calculations were also done using the 2225 λ-repressor mutants (prediction set) as the design set.

[Figure 6 about here.]

The results of the design calculation are shown in Figure 6 and Table 3. The top five selected peptides (Table 3) are all single point mutations of λ-repressor helix number 1.

[Table 3 about here.]

A heatmap showing expected gain for each pair under study is shown in Figure 6. Not surprisingly, there are notable blocks of high expected gain corresponding to point mutations within specific helix regions.

[Figure 7 about here.]

A comparison of expected gain from each of the study groups is shown in Figure 7. The largest expected gain comes from studying point mutations in the first helix region, with slightly better results from studying the longer stretch of amino acids as defined in Burton et al. (1998); note that helix 1 is also the most helical. As can be seen from Figure 6, the large gains from studying this region derive mostly from improved predictive accuracy of the region itself. Changes in the relative importance of each of the regions, as well as differences in cost would affect the choice to study a host/guest sequence versus point mutations of the λ-repressor helicies. Experimental design provides a quantitative framework for making this decision.

There is an overall increase in expected gain from studying the polypeptides at 310 kelvin (body temperature) versus 298 kelvin (room temperature), with the strongest effect seen in the host/guest peptide group. All peptides in the prediction and design sets were studied between pH 6.0 and pH 8.0, and there is no clear effect, for this experiment, of pH on the calculated expected gain.

There is a striking consistency in the top host/guest style polypeptides: all include the $L - I$ $i, i + 3$ interaction. Indeed, 14 of the top 20 contain this interaction. Five of the remaining six contain $E - R$ in $i, i + 4$ configuration, and the last contains $E - R$ in $i, i + 3$ configuration. Note that both helicies 1 and 4, the two most helical, contain $L - I$ in $i, i + 3$ position and that helix 1 contains $E - R$ in $i, i + 4$ position. Thus the formal design criteria is choosing guest peptides which best contribute to improving predictive accuracy *simultaneously* on the *entire* prediction set. In fact we find that if we create a host/guest style peptide which contains both interactions: 'AEAAARALAAIAAAAKA' we get an expected gain of .0323, a significant improvement over any of the 8000 peptides in the original host/guest experiment.

6 Conclusion

The Bayesian approach to experimental design provides a natural framework for quantifying the uncertainty and relevant costs and benefits from an experimental study. Although Bayesian design has not seen significant use in bioinformatics and computational biology to date, the potential is clear. We have outlined the general framework here, and developed a concrete example in the context of helical peptide studies. As demonstrated, this approach has the ability to optimally direct investment of experimental resources to improve predictive or other aspects of computational models, as well as to provide insight into uncertainty in current models, and limitations in currently available experimental data. The potential scope of such applications in bioinformatics is very large, and we hope that the Bayesian design framework may come to see much broader usage.

Acknowledgments

SCS and JL were partially supported by NSF grant DMS-0204690 (SCS). TGO was supported by NIH grant GM045322. Computing resources were provided by NSF infrastructure grant DMS-0112340 (SCS).

References

Andersen, N. H. and Tong, H. (1997). Empirical parameterization of a model for predicting peptide helix/coil equilibrium populations. *Protein Science*, 6:1920–1936.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition.

Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics*, 19(1):43–56.

Bernardo, J. (1979). Expected information as expected utility. *Annals of Statistics*, 7:686–690.

Bernardo, J. M. and Smith, A. F. (1994). *Bayesian Theory*. John Wiley & Son.

Box, G. E. P. (1982). Choice of response surface design and alphabetic optimality. *Utilitas Mathematica*, 21B:11–55.

Burton, R. E., Myers, J. K., and Oas, T. G. (1998). Protein folding dynamics: Quantitative comparison between theory and experiment. *Biochemistry*, 37(16):5337–5343.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1):79–94.

Clyde, M. (2001). Experimental design: A Bayesian perspective. In Smelser, editor, *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier Science, New york.

Coffey, T., Gennings, C., Simmons, J. E., and Herr, D. W. (2005). D-optimal experimental designs to test for departure from additivity in a fixed-ration mixture ray. *Toxicological Sciences*, 88(2):467–476.

Doig, A. J., Chakrabartty, A., Klingler, T. M., and Baldwin, R. L. (1994). Determination of free energies of N-capping in α-helices by modification of the Lifson-Roig helix-coil theory to include N- and C-capping. *Biochemistry*, 33:3396–3403.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eddy, S. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088.

Fisher, R. A. (1935). *Design of Experiments*. Oliver and Boyd, Edinburgh.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Fujiwara, M., Nagy, Z. K., Chew, J. W., and Braatz, R. D. (2005). First-principles and direct design approaches for the control of pharmaceutical cyrstallization. *Journal of Process Control*, 15:493–504.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Horvitz, E. J., Breese, J. S., and Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, 2(3):247–302.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531.

Lathrop, R. H., Rogers, R. G., Smith, T. F., and White, J. V. (1998). A Bayes-optimal probability theory that unifies protein sequence-structure recognition and alignment. *Bulletin of Mathematical Biology*, 60:1039–1071.

Lindley, D. V. (1992). *Bayesian Statistics, A Review*. SIAM, Philadelphia.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52.

Lucas, J. E. (2006). *Sparsity Modeling for High Dimensional Systems: Applications in Genomics and Structural Biology*. PhD thesis, Duke University.

Marqusee, S. and Sauer, R. T. (1994). Contributions of a hydrogen bond/salt bridge network to the stability of secondary and tertiary structure in λ repressor. *Protein Science*, 3:2217–2225.

Munoz, V. and Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nature Structural Biology*, 1(6):399–409.

Petukhov, M., Yumoto, N., Murase, S., Onmura, R., and Yoshikawa, S. (1996). Factors that affect the stabilization of α-helices in short peptides by a capping box. *Biochemistry*, 35:387–397.

Poland, D. and Scheraga, H. A. (1970). *Theory of Helix-Coil Transitions in Biopolymers: Statistical Mechanical Theory of Order-Disorder Transitions in Biological Macromolecules*. Academic Press.

Qian, H. and Schellman, J. A. (1992). Helix-coil theories: A comparative study for finite-length polypeptides. *Journal of Physical Chemistry*, 96:3987–3994.

Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Addison-Wesley.

Russell, S. J. and Norvig, P. (2003). *Artifical Intelligence: A Modern Approach*. Prentice Hall, 2nd edition.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York.

Schmidler, S. C. (2006). Bayesian flexible shape matching with applications to structural bioinformatics. (submitted to *Journal of the American Statistical Association*).

Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, 7(1):233–248.

Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2004). Bayesian modeling of non-local interactions in protein sequences. (*submitted*).

Schmidler, S. C., Lucas, J., and Oas, T. G. (2007). Statistical estimation in statistical mechanical models: Helix-coil theory and peptide helicity prediction. *Journal of Computational Biology*, 14(10):(to appear).

Scholtz, J. M. and Baldwin, R. L. (1992). The mechanism of α-helix formation by peptides. *Annual Review of Biophysics and Biomolecular Structure*, 21:95–118.

Shalongo, W. and Stellwagen, E. (1995). Incorporation of pairwise interactions into the Lifson-Roig model for helix prediction. *Protein Science*, 4:1161–1166.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, 5:229–235.

Stapley, B. J., Rohl, C. A., and Doig, A. J. (1995). Addition of side chain interactions to modified Lifson-Roig helix-coil theory: Application to energetics of Phenylalanine-Methionine interactions. *Protein Science*, 4:2383–2391.

Wu, T. D., Schmidler, S. C., Hastie, T., and Brutlag, D. L. (1998). Regression analysis of multiple protein structures. *Journal of Computational Biology*, 5(3):597–607.

Wynn, H. (1970). The sequential generation of D-optimum experimental designs. *Annals of Mathematical Statistics*, 41:1655–1664.

List of Figures

1 Expected improvement in predictive accuracy for 220 randomly sampled peptides when any single peptide in the set is chosen for experimental study. . . 34

2 Posterior distributions for the $i, i+3$ interaction terms (at relevant pH) occurring in the peptide of example 1. Shown are Lys-Glu, Glu-Lys, Lys-Arg, and Leu-Leu $\Delta\Delta G$ parameters. 35

3 The expected gain in prediction accuracy for the polypeptide 'PANLKALEAQKQKEQR' associated with the study of polypeptide 'Y(AEAAKA)^{8}F' at all of the temperatures available in the full data set. The points marked with 'x' are the ones that were randomly chosen to be a part of the prediction and design sets. 36

4 Expected utility for each peptide in the prediction set to be obtained by studying each peptide in the design set at pH 4 (a) or pH 10 (b). The results for pH 7 (not shown) are nearly identical to those for pH 4. 37

5 The variance of mean helicity in the prediction set attributable to each of the individual amino acid ΔS parameters. Amino acids are ordered (lowest to highest) according to the expected gain obtained from the design calculation in Section 5.2; the two criteria show little correlation. 38

6 A heatmap showing expected gain for predicting helicity of peptides along the x-axis from studying the top five peptides in each group (y-axis). There are notable blocks due to the nature of the construction of the prediction set (point mutations of fixed domains of the λ-repressor protein). 39

7 The expected gain by group for each of the helix regions and for the host/guest peptide. 40

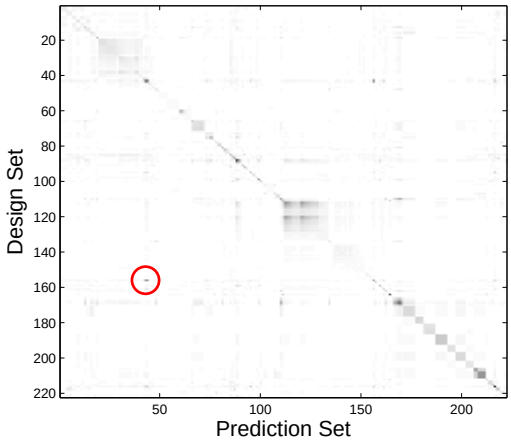


Figure 1: Expected improvement in predictive accuracy for 220 randomly sampled peptides when any single peptide in the set is chosen for experimental study.

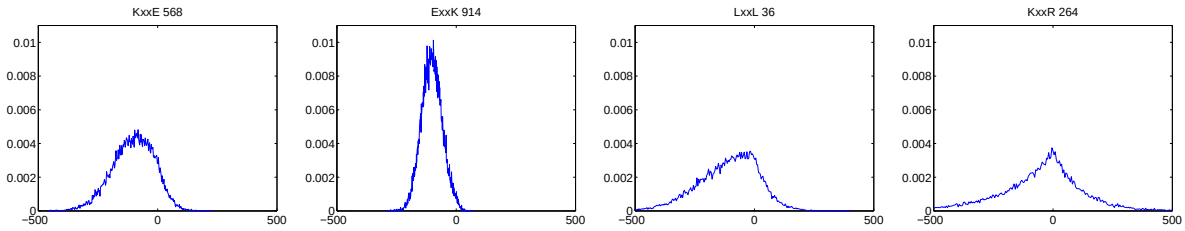


Figure 2: Posterior distributions for the $i, i+3$ interaction terms (at relevant pH) occurring in the peptide of example 1. Shown are Lys-Glu, Glu-Lys, Lys-Arg, and Leu-Leu $\Delta\Delta G$ parameters.

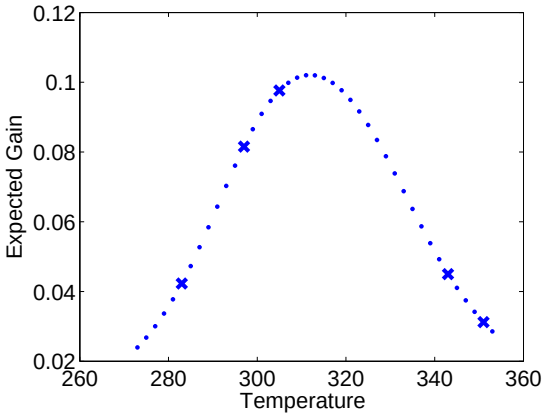
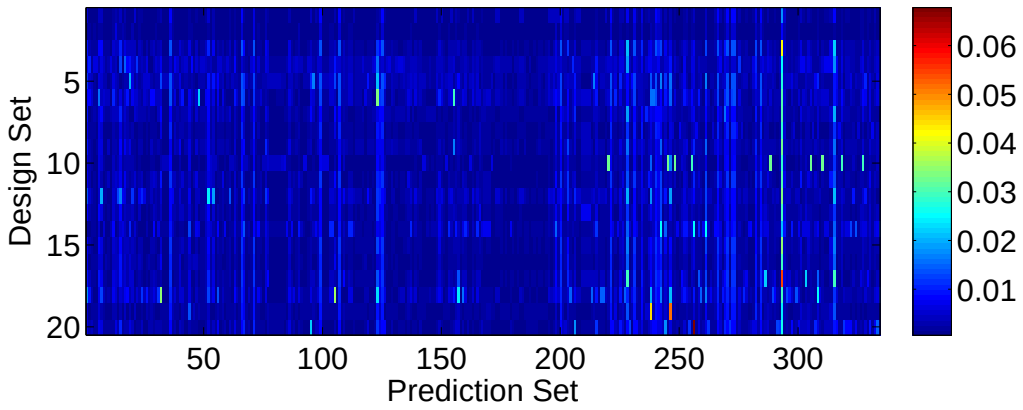
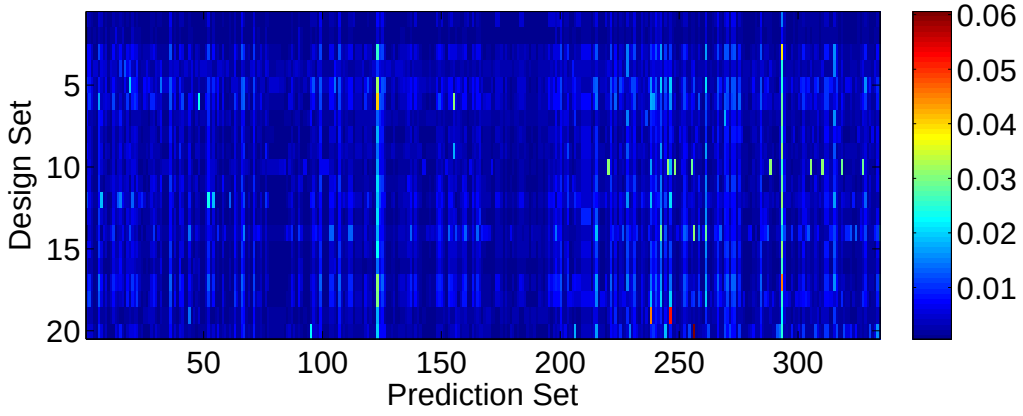


Figure 3: The expected gain in prediction accuracy for the polypeptide 'PANLKALEAQKQKEQR' associated with the study of polypeptide 'Y(AEAAKA)^{8}F' at all of the temperatures available in the full data set. The points marked with 'x' are the ones that were randomly chosen to be a part of the prediction and design sets.



(a)



(b)

Figure 4: Expected utility for each peptide in the prediction set to be obtained by studying each peptide in the design set at pH 4 (a) or pH 10 (b). The results for pH 7 (not shown) are nearly identical to those for pH 4.

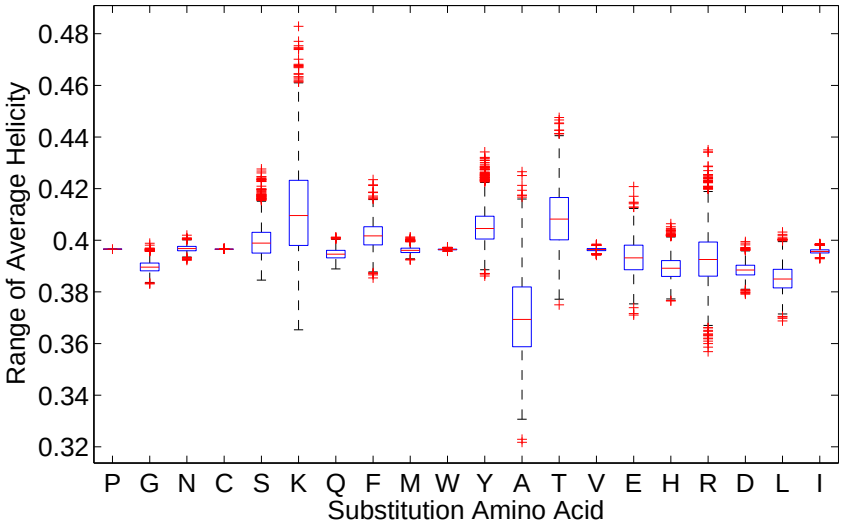


Figure 5: The variance of mean helicity in the prediction set attributable to each of the individual amino acid ΔS parameters. Amino acids are ordered (lowest to highest) according to the expected gain obtained from the design calculation in Section 5.2; the two criteria show little correlation.

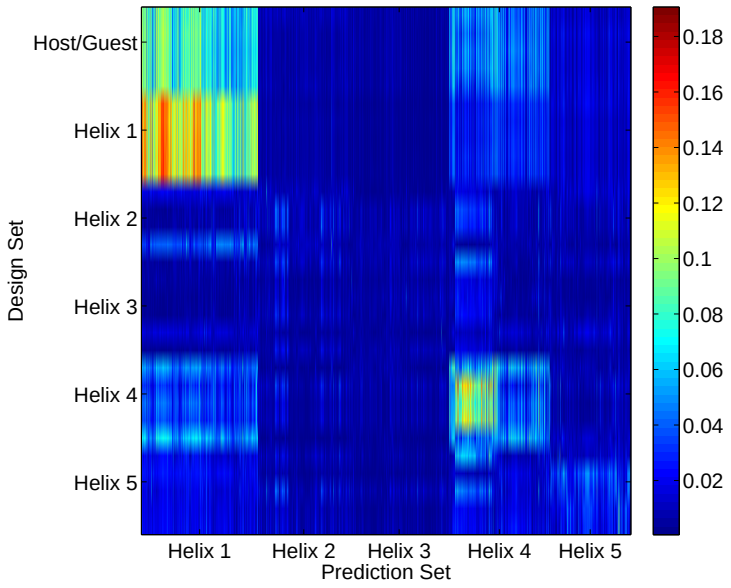


Figure 6: A heatmap showing expected gain for predicting helicity of peptides along the x-axis from studying the top five peptides in each group (y-axis). There are notable blocks due to the nature of the construction of the prediction set (point mutations of fixed domains of the λ-repressor protein).

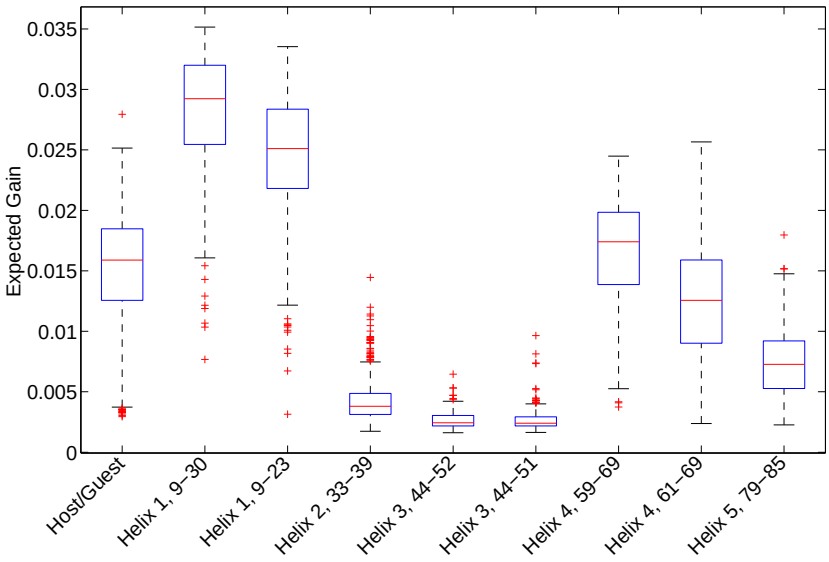


Figure 7: The expected gain by group for each of the helix regions and for the host/guest peptide.

List of Tables

1 The top five peptides recommended for study, ordered by the expected gain in accuracy of helicity prediction. 42

2 Counts of hydrophobic $i, i+3$ and $i, i+4$ sidechain interactions occurring in prediction set. Interactions involving Leu and Ile occur substantially more often than the other hydrophobic amino acids. 43

3 The top five polypeptides from the Helix 1 and the Host/Guest groups. For the helix 1 polypeptides, point mutations are highlighted in red. For the host/guest peptides, the guest positions are highlighted in red. All of the top five host/guest peptides contain the L-I $i, i+3$ interaction. The first helix in the λ-repressor protein also contains this interaction. w.t.=wildtype 44

Rank	1	2	3	4	5
Substitution	L	I	R	D	E
pH	10	10	10	4	10
Expected gain (%)	.0055	.0053	.0051	.0051	.0049

Table 1: The top five peptides recommended for study, ordered by the expected gain in accuracy of helicity prediction.

Amino Acid	Interaction Pair	Count	Interaction Pair	Count	Total Count
L	L...L	7	L..F	5	12
I	L...I	8	I..F	3	11
V	L...V	5	V..F	1	6
F	L...F	2	F..F	2	4
M	L...M	0	M..F	0	0

Table 2: Counts of hydrophobic $i, i+3$ and $i, i+4$ sidechain interactions occurring in prediction set. Interactions involving Leu and Ile occur substantially more often than the other hydrophobic amino acids.

Polypeptide	Expected gain	
Helix 1 (w.t.)	QEQLEDARRLKAIYEKKKNELG	.0341
Helix 1	QEILEDARRLKAIYEKKKNELG	.0352
	QEQLEDARRLKAIYEKKKKELG	.0345
	QEQLEDARRLKAIYEKKKNELA	.0348
	QEQLEDARRLKAIYEKKKNELL	.0350
	QEQLEDARRLKAIYMKKKAELG	.0351
Host/Guest	AEAAAADLAAIAAAAKA	.0312
	AEAAAAELAAIAAAAKA	.0311
	AEAAAANLAAIAAAAKA	.0309
	AEAAAAQLAAIAAAAKA	.0307
	AEAAAAYLAAIAAAAKA	.0304

Table 3: The top five polypeptides from the Helix 1 and the Host/Guest groups. For the helix 1 polypeptides, point mutations are highlighted in red. For the host/guest peptides, the guest positions are highlighted in red. All of the top five host/guest peptides contain the L-I $i, i+3$ interaction. The first helix in the λ-repressor protein also contains this interaction. w.t.=wildtype