

Noname manuscript No.  
(will be inserted by the editor)

Gurinder S. Atwal · Justin B. Kinney

# Learning quantitative sequence-function relationships from high-throughput biological data

Received: date / Accepted: date

**Abstract** Understanding the transcriptional regulatory code, as well as other types of information encoded within biomolecular sequences, will require learning biophysical models of sequence-function relationships from high-throughput data. Controlling and characterizing the noise in such experiments, however, is notoriously difficult. The unpredictability of such noise creates problems for standard likelihood-based methods in statistical learning, which require that the quantitative form of experimental noise be known precisely. However, when this unpredictability is properly accounted for, important theoretical aspects of statistical learning which remain hidden in standard treatments are revealed. Specifically, one finds a close relationship between the standard inference method, based on likelihood, and an alternative inference method based on mutual information. Here we review and extend this relationship. We also describe its implications for learning sequence-function relationships from real biological data. Finally, we detail an idealized experiment in which these results can be demonstrated analytically.

## 1 Introduction

A major long-term goal in biology is to understand how biological function is encoded within the sequences of DNA, RNA, and protein. The canonical success story in this effort is the genetic code. Given an arbitrary sequence of messenger RNA, the genetic code allows us to predict with near certainty what peptide sequence will result.

But there are many other biological codes we would like to learn. How does the DNA sequence of a promoter or enhancer encode transcriptional regulatory programs? How does the sequence of pre-mRNA govern which exons are kept and which are removed from the final spliced mRNA? How does the peptide sequence of an antibody govern how strongly it binds to target antigens? For each of these questions, we aim to learn a sequence-function relationship: given the sequence of a biological heteropolymer, what will that molecule end up actually doing?

A major difference between the genetic code and these other codes is that while the former is qualitative in nature, the latter govern biological processes that are inherently quantitative. For example, the rate of transcription from a promoter can range continuously over orders of magnitude, and the value of this rate can be finely tuned by adjusting the DNA sequence of the promoter. Learn-

---

J. B. Kinney  
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory  
Cold Spring Harbor, NY 11724  
Tel.: +1-516-3675230  
E-mail: jkinney@cshl.edu

G. S. Atwal  
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory

ing these biological codes thus requires theoretical and computational methods to infer quantitative sequence-function relationships from data.

Experimental methods for measuring sequence-function relationships have improved dramatically over the last 15 years. The availability of microarray-based experiments [1,2], for example, made it possible to accurately measure the sequence-specificity of individual DNA-binding proteins, such as transcription factors. More recently, multiple techniques (e.g., [3–5]) for simultaneously measuring the activity of many different transcriptional regulatory sequences have been developed. Below we discuss one of these methods, Sort-Seq [5], in greater detail.

These experimental methods, which can measure the activities of many different biological sequences at once, are very unlike conventional experiments in physics. Such high-throughput biological experiments are typically very noisy and rarely provide direct readouts for the quantities that one cares about. Moreover, these measurements generally exhibit substantial day-to-day variability. However, as shown in this paper, it is still possible to precisely learn quantitative sequence-function relationships from such data, even though the noise characteristics of these data are difficult or impossible to characterize up front.

The ability to do this reflects subtle but important distinctions between two objective functions used for statistical inference: (i) likelihood, which requires *a priori* knowledge of the experimental noise function and (ii) mutual information, a fundamental quantity in information theory [6] that does not require a noise function. In contrast to the conventional wisdom that more experimental measurements will improve the model inference task, we show that there are instances in which the popular maximum likelihood approach will never learn the right model, even in the infinite data limit. Model inference based on mutual information does not suffer from this ailment, but it is unable to determine the values of a small subset of model parameters known as “diffeomorphic modes” [7]. This inability of mutual information to pin down parameters along diffeomorphic modes does not indicate a problem with mutual information, but rather reflects a fundamental distinction between how diffeomorphic and nondiffeomorphic model parameters are constrained by data.

In this paper we detail these two distinct inference approaches in the context of learning sequence-function relationships and elaborate their relationship to one another. We also work through an explicit analytical example of inference in an idealized Sort-Seq experiment. This example highlights the differences between likelihood- and mutual-information-based approaches to inference, as well as the emergence of diffeomorphic modes.

It should be noted that the inference of receptive fields in sensory neuroscience is another area of biology in which mutual information has proved useful as an objective function, and that work in this area has also provided important insights into basic aspects of machine learning [8–12]. This body of work, however, has largely avoided in-depth discussions of how mutual information relates to likelihood, which is a major focus of the present paper.

## 2 Sort-Seq and other experimental methods

We begin by describing Sort-Seq [5], one of the recently developed experimental techniques for measuring quantitative sequence-function relationships. This assay uses flow cytometry and high-throughput DNA sequencing to simultaneously measure the activity of a large number of different transcriptional regulatory sequences. It was first used for characterizing the biophysical basis of transcriptional regulation in *E. coli*. This basic approach has since been adapted for studying multiple aspects of gene regulation in bacteria *E. coli* [13], yeast [14], and human cells [15]. Sort-Seq-like methods have also been used to study protein sequence-function relationships in various systems [16,17].

First, a DNA plasmid is created in which a regulatory sequence of interest is placed upstream of a fluorescent reporter gene, such as green fluorescent protein (GFP). A large library of plasmids is then created by replacing this regulatory sequence with variant regulatory sequences containing randomly scattered substitution mutations. The resulting plasmids are then inserted into cells, which are allowed to grow under expression-inducing conditions. Cells then undergo fluorescence-activated cell sorting (FACS). This procedure sorts the cells into a small number of “batches” based on each cell’s measured fluorescence. FACS machines can sort thousands of cells a second, so it is not difficult to sort  $\sim 10^6$  cells into each batch.

Finally, the mutated regulatory sequences within each batch are amplified using PCR and then sequenced. The resulting data set consists of a large number of mutant regulatory sequences  $S$ , each

with a corresponding measurement  $M$  (the batch in which the sequence is found). In the original experiment [5],  $\sim 2.5 \times 10^5$  independent measurements were taken. Advances in DNA sequencing have since made it possible to accumulate much more data, and it is no longer difficult to assay the activities of millions of sequences in a single experiment.

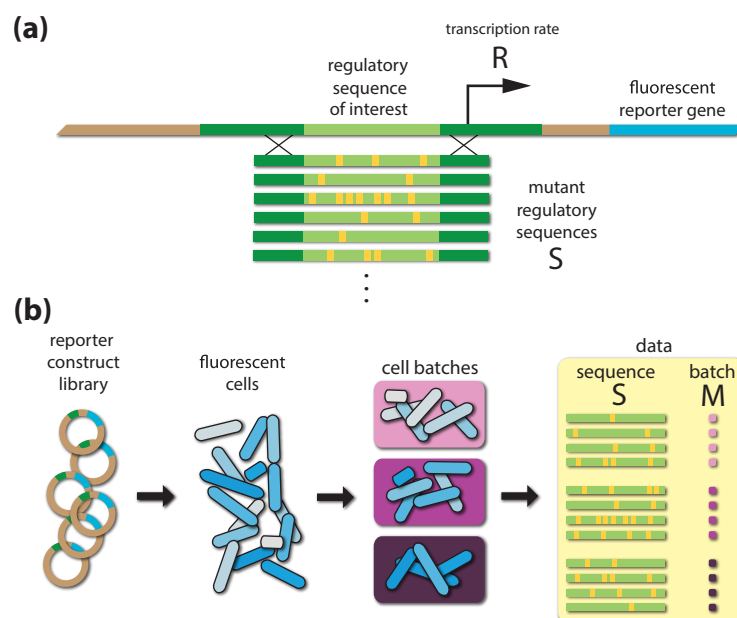


**Fig. 1** Overview of the Sort-Seq assay [5]. (a) A regulatory sequence of interest is inserted into a plasmid upstream of a fluorescent reporter gene. This regulatory sequence is then replaced with variant sequences, each having multiple scattered substitution mutations. (b) The resulting plasmid library is put into cells. Cells are cultured under inducing conditions, then sorted into batches according to fluorescence. The variant regulatory sequences in each batch of sorted cells are then sequenced *en masse*. The resulting data set is comprised of a large number of variant sequences  $S$ , each assigned to a batch  $M$ . Figure adapted from [5].

Fig. 2 depicts biophysical models of transcriptional regulation that were learned from Sort-Seq data in the study of [5]. These illustrate just a few of the quantitative sequence-function relationships that Sort-Seq and related experiments can elucidate. In the Sort-Seq experiments of [5], a 75 bp region of the *E. coli lac* promoter was interrogated. This region contains binding sites for two regulatory proteins, the RNA polymerase holoenzyme (RNAP) and the cAMP receptor protein (CRP) transcription factor (Fig. 2a). The resulting Sort-Seq data was used to learn “energy matrix” models for the sequence-dependent binding energies of CRP and RNAP (Fig. 2b,c). These data were also used to infer a biophysical model describing how these two proteins interact (Fig. 2d). By fitting these quantitative sequence-function relationships to data, it was possible to learn physically meaningful values for the *in vivo* molecular interaction energies that form the mechanistic basis for transcriptional regulation in this system. The specific way such models were fit to data is discussed in more detail in section 6.3 of this paper.

Before proceeding, we briefly describe other experimental methods that can produce data needing similar analysis. One alternative to sorting sequences is to directly quantify the abundance of transcripts produced from different regulatory sequences using RNA-Seq. Ways of using this approach to dissect regulatory sequences *in vitro* [4] and *in vivo* [18–20] have been described. Another method is to perform a selection assay that enriches for sequences according to some measured activity, then to sequence both the selected and non-selected batches of sequences [3,21,22].

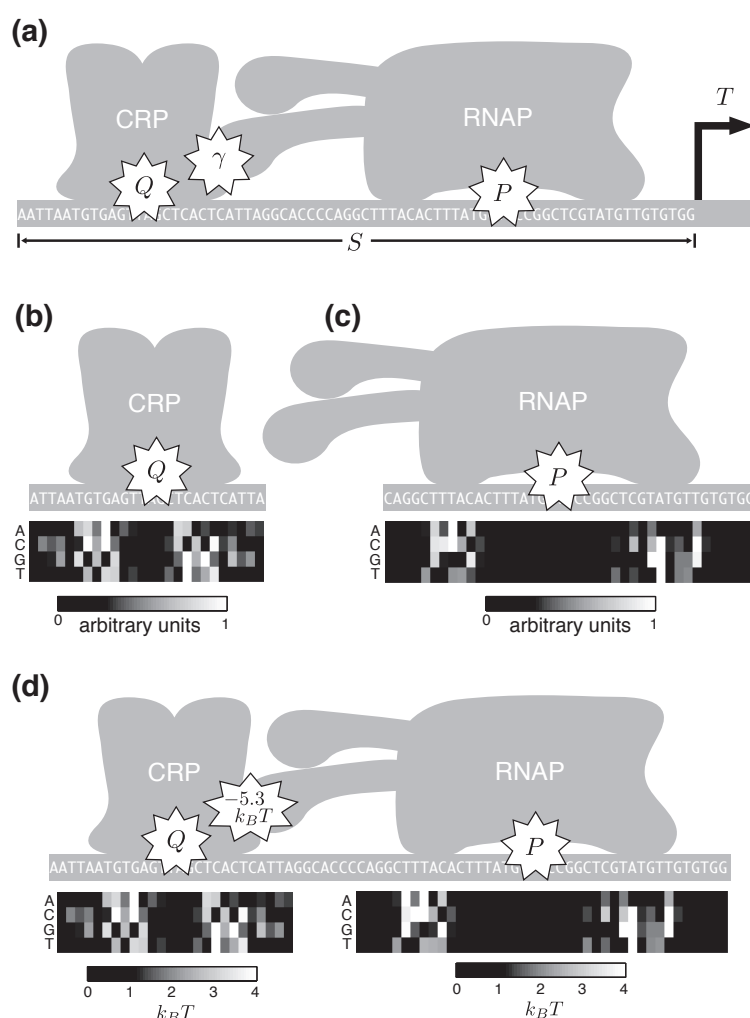


**Fig. 2** Sequence-function relationships fit to Sort-Seq data in [5]. (a) The authors inferred a biophysical model of transcriptional regulation by the *lac* promoter. This biophysical model included two “energy matrix” models, one describing the sequence-dependent binding energy of CRP ( $Q$ ), and one describing the binding energy of RNAP ( $P$ ). It also included a value for the interaction energy  $\gamma$  between these two proteins. (b) Fitting only the CRP energy matrix resulted in an energy matrix with unknown scale. (c) Similarly, fitting only the RNAP energy matrix resulted in an unknown energy scale. (d) Fitting the full biophysical model recovered both CRP and RNAP energy matrices in meaningful energy units, as well as a value for the protein-protein interaction energy  $\gamma$ . Figure adapted from [7].

### 3 Inference using likelihood

The inference of quantitative sequence-function relationships from Sort-Seq data can be phrased as follows. We have a data set consisting of a large number of sequences  $\{S_n\}_{n=1}^N$ , each sequence  $S$  having a corresponding measurement  $M$ . This process is stochastic: due to experimental noise, different measurements of the same sequence  $S$  can yield different values for  $M$ . Our experiment therefore has the following form

$$\underset{\text{sequence}}{S} \xrightarrow[\text{experiment}]{p(M|S)} \underset{\text{measurement}}{M}. \quad (1)$$

If we have an explicit parametric form for  $p(M|S)$ , we can learn the values of the parameters by maximizing the per-datum log likelihood,

$$L = \frac{1}{N} \sum_{n=1}^N \log p(M_n|S_n). \quad (2)$$

Here we have assumed that the measurements for each sequence are independent. In regression problems such as this, one introduces an additional layer of structure. In this case, we expect the measurement  $M$  of each sequence  $S$  to be a noisy readout of some underlying activity  $R$  that is a deterministic function of that sequence; we call this function our “model”, and denote it  $\theta(S)$ . This model is ultimately what we care about. The noisiness of the experiment is then characterized by a “noise function”  $\pi(M|R)$ . Our experiment is thus represented by

$$\begin{array}{ccccc} S & \xrightarrow[\text{sequence}]{\text{model } \theta(S)} & R & \xrightarrow[\text{activity}]{\text{noise function } \pi(M|R)} & M \\ & & & & \text{measurement} \end{array}, \quad (3)$$

where  $p(M|S) = \pi(M|\theta(S))$ . The corresponding likelihood is

$$L(\theta, \pi) = \frac{1}{N} \sum_{n=1}^N \log \pi(M_n|\theta(S_n)). \quad (4)$$

The model we adopt for our experiment therefore has two components:  $\theta$ , which we want to learn from data, and  $\pi$ , which we do not really care about.

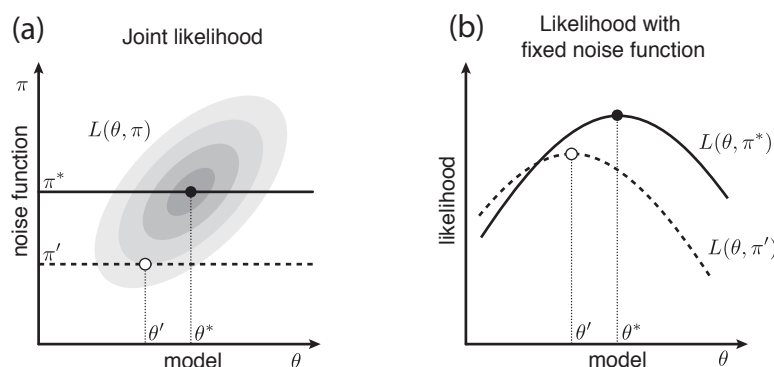


**Fig. 3** Schematic illustration of how likelihood  $L(\theta, \pi)$  depends on the model  $\theta$  and the noise function  $\pi$  in the  $N \rightarrow \infty$  limit. (a)  $L$  will typically have a correlated dependence on  $\theta$  and  $\pi$ . (b) If  $\pi$  is set equal to the correct noise function  $\pi^*$ , then  $L$  will be maximized by the correct model  $\theta^*$ . However, if  $\pi$  is set to an incorrect noise function  $\pi'$ ,  $L$  will typically attain a maximum at an incorrect model  $\theta'$ . Figure adapted from [7].

Standard statistical regression requires that the noise function  $\pi$  be specified up-front.  $\pi$  can be learned either by performing separate calibration experiments, or by assuming a functional form based on an educated guess. This can be problematic, however. Consider inference in the large data limit ( $N \rightarrow \infty$ ), which is illustrated in Fig. 3. Likelihood is determined by both the model  $\theta$  and the noise function  $\pi$  (Fig. 3a). If we know the correct noise function  $\pi^*$  exactly, then maximizing  $L(\theta, \pi^*)$  over  $\theta$  is guaranteed to recover the correct model  $\theta^*$ . However, if we assume an incorrect noise function  $\pi'$ , maximizing likelihood will typically recover an incorrect model  $\theta'$  (Fig. 3b).

#### 4 Inference using mutual information

Information theory provides an alternative inference approach. Suppose we hypothesize a specific model  $\theta$ , which gives predictions  $R$ . Furthermore, denote the true model  $\theta^*$  and the corresponding true activity

$R^*$ . The dependence between  $S$ ,  $M$ ,  $R^*$ , and  $R$  will then follow the graphical model,

$$R \xleftarrow{\theta} S \xrightarrow{\theta^*} R^* \xrightarrow{\pi} M. \quad (5)$$

From the simple fact that  $M$  depends on  $R$  through the value of  $R^*$  any dependence measure  $\mathcal{D}$  that satisfies the Data Processing Inequality (DPI) must satisfy

$$\mathcal{D}[R; M] \leq \mathcal{D}[R^*; M]. \quad (6)$$

Therefore, in the set of possible models  $\theta$ , the true model is guaranteed to globally maximize the objective function  $\mathcal{D}(\theta) \equiv \mathcal{D}[R; M]$ .

One particularly relevant dependence measure satisfying DPI is mutual information. Mutual information plays a fundamental role in information theory [6]. It also provides a self-equitable measure of statistical association [23]. In the case of Sort-Seq experiments, which have continuous  $R$  and discrete  $M$ , mutual information is given by

$$I(\theta) = I[R; M] = \sum_M \int dR p(R, M) \log \frac{p(R, M)}{p(R)p(M)}, \quad (7)$$

where  $p(M, R)$  is the distribution of activity predictions and measurements resulting from the model  $\theta$ . Note, in particular, that computing mutual information does not depend on the assumption of any noise function  $\pi$ . Rather, if one is able to estimate  $p(M, R)$  from a finite sample of data, mutual information can be used as an objective function for determining  $\theta$  without assuming any noise function  $\pi$ .

It should be noted that there are multiple dependence measures  $\mathcal{D}$  that satisfy DPI. One might wonder whether maximizing multiple different dependence measures would improve on the optimization of mutual information alone. The answer is not so simple. In [7] it was shown that if the correct model  $\theta^*$  is within the space of models under consideration, then, in the large data limit, maximizing mutual information is equivalent to simultaneously maximizing every dependence measure that satisfies DPI. On the other hand, one rarely has any assurance that the correct model  $\theta^*$  is within the space of parameterized models one is considering. In this case, considering different DPI-satisfying measures might provide a test for whether  $\theta^*$  is noticeably outside the space of parameterized models. To our knowledge, this potential approach to the model selection problem has yet to be demonstrated.

## 5 Relationship between likelihood and mutual information

An alternative inference approach is to admit that we do not know the noise function  $\pi$  *a priori*, and to fit *both*  $\theta$  and  $\pi$  simultaneously by maximizing  $L(\theta, \pi)$  over this pair. It is easy to see why this makes sense: the division of the inference problem into first measuring  $\pi$ , then learning  $\theta$  using that inferred  $\pi$ , is somewhat artificial. The process that maps  $S$  to  $M$  is determined by both  $\theta$  and  $\pi$  and thus, from a probabilistic point of view, it makes sense to maximize likelihood over both of these quantities simultaneously.

We now show that, in the large  $N$  limit, maximizing likelihood over both  $\theta$  and  $\pi$  is equivalent to maximizing the mutual information between model predictions and measurements. Here we follow the argument given in [7]. In the large  $N$  limit, likelihood can be decomposed into three terms

$$L(\theta, \pi) = \sum_M \int dR p(R, M) \log \pi(M|R) = I(\theta) - D(\theta, \pi) - H[M]. \quad (8)$$

where

$$D(\theta, \pi) = \sum_M \int dR p(R, M) \log \frac{p(M|R)}{\pi(M|R)}, \quad (9)$$

is the Kullback-Leibler divergence between the assumed noise function  $\pi$  and the observed noise function  $p(M|R)$ , and  $H[M] = -\sum_M p(M) \log p(M)$  is the entropy of the measurements, which does not depend on  $\theta$ . To maximize  $L(\theta, \pi)$ , it therefore suffices to maximize  $I(\theta)$  over  $\theta$  alone, then to set the noise function  $\pi(M|R)$  equal to the empirical noise function  $p(M|R)$ , which causes  $D(\theta, \pi)$  to vanish.

Thus, when we are uncertain about the noise function  $\pi$ , we need not despair. We can, if we like, simply learn  $\pi$  at the same time that we learn  $\theta$ . We need not explicitly model  $\pi$  in order to do this; it suffices instead to maximize the mutual information  $I(\theta)$  over  $\theta$  alone.

The connection between mutual information and likelihood can further be seen in a quantity called the “noise-averaged” likelihood. This quantity was first described for the analysis of microarray data [24]; see also [7]. The central idea is to put an explicit prior on the space of possible noise functions, then compute likelihood after marginalizing over these noise functions. Explicitly, the per-datum log noise-averaged likelihood  $L_{na}(\theta)$  is related to  $L(\theta, \pi)$  via

$$e^{NL_{na}(\theta)} = \int d\pi p(\pi) e^{NL(\theta, \pi)}. \quad (10)$$

Under fairly general conditions, one finds that noise-averaged likelihood is related to mutual information via

$$L_{na}(\theta) = I(\theta) - \Delta(\theta) - H[M]. \quad (11)$$

Here,  $\Delta(\theta)$  is a quantity that vanishes in the  $N \rightarrow \infty$  limit, and thus becomes irrelevant for the inference problem on sufficiently large data sets.

## 6 Diffeomorphic modes

Mutual information has a mathematical property that is important to account for when using it as an objective function: the mutual information between any two variables is unchanged by an invertible transformation of either variable. So if a change in model parameters,  $\theta \rightarrow \theta'$ , results in changes in model predictions  $R \rightarrow R'$  that preserves the rank order of these predictions, then

$$I(\theta) = I[M; R] = I[M; R'] = I(\theta'), \quad (12)$$

and  $\theta$  and  $\theta'$  are judged to be equally valid.

By using mutual information as an objective function, we are therefore unable to constrain any parameters of  $\theta$  that, if changed, produce invertible transformations of model predictions. Such parameters are called “diffeomorphic parameters” or “diffeomorphic modes” [7]. The distinction between diffeomorphic parameters and nondiffeomorphic parameters is illustrated in Fig. 4.

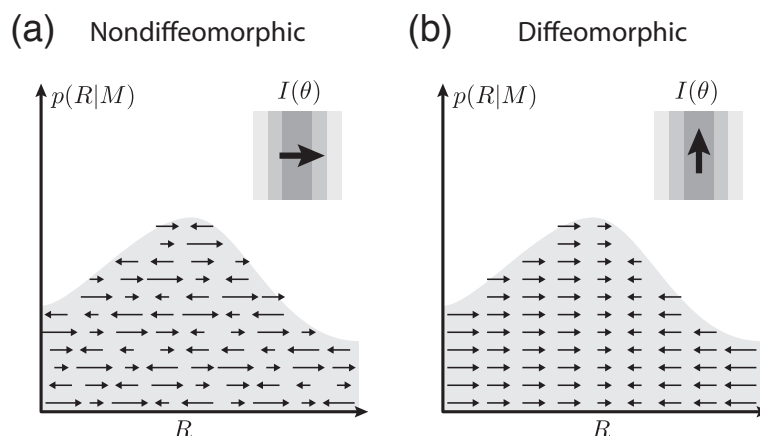


**Fig. 4** Illustration of diffeomorphic and nondiffeomorphic parameters. (a) Changing the value of a nondiffeomorphic parameter of  $\theta$  results in a sort of “diffusion” in which the  $R$  values assigned to different sequences change rank order. Such changes will typically alter the mutual information  $I(\theta) = I[R; M]$ . (b) Changing the value of a diffeomorphic parameter, however, results in a “flow” of  $R$  values that maintains their rank order. Such flows in  $R$ -space leave  $I(\theta)$  unchanged.

## 6.1 Criterion for diffeomorphic modes

Following [7], we now derive a criterion that can be used to identify all of the diffeomorphic modes of a model  $\theta$ .<sup>1</sup> Consider an infinitesimal change in model parameters  $\theta \rightarrow \theta + d\theta$ , where the components of  $d\theta$  are specified by

$$d\theta_i = \epsilon g_i \quad (13)$$

for some vector  $g_i$  in  $\theta$ -space. This change in  $\theta$  will produce a corresponding change in model predictions  $R \rightarrow R + dR$ , where

$$dR = \epsilon \sum_i g_i \frac{\partial R}{\partial \theta_i}. \quad (14)$$

In general, the derivative  $\partial R / \partial \theta_i$  can have arbitrary dependence on the underlying sequence  $S$ . However, this transformation will preserve the rank order of  $R$ -values if and only if  $dR$  is the same for all sequences having the same value of  $R$ . The change  $dR$  must therefore be a function of  $R$ , and have no other dependence on  $S$ . A diffeomorphic mode is a vector field  $g(\theta)$  that has this property at all points in parameter space. Specifically, a vector field  $g(\theta)$  is a diffeomorphic mode if and only if there is a function  $h(R, \theta)$  such that

$$\sum_i g_i(\theta) \frac{\partial R}{\partial \theta_i} = h(R, \theta). \quad (15)$$

## 6.2 Diffeomorphic modes of linear models

As a simple example, consider a situation in which each sequence  $S$  is a  $D$ -dimensional vector, and  $R$  is an affine function of  $S$ , i.e.

$$R = \theta_0 + \sum_{i=1}^D \theta_i S_i, \quad (16)$$

for model parameters  $\theta = \{\theta_0, \theta_1, \dots, \theta_D\}$ . The criterion (Eq. (15)) for a vector field  $g$  being a diffeomorphic mode then gives

$$g_0(\theta) + \sum_{i=1}^D g_i(\theta) S_i = h(R, \theta). \quad (17)$$

Because the left hand side is linear in  $S$ , and  $R$  is linear in  $S$ , the function  $h(R, \theta)$  must be linear in  $R$ . Thus,  $h$  must have the form

$$h(R, \theta) = a(\theta) + b(\theta)R. \quad (18)$$

So at each  $\theta$ , the vector field  $g(\theta)$  has at most two degrees of freedom, which correspond to the values of  $a(\theta)$  and  $b(\theta)$ . Specifically,

$$g_i(\theta) = \begin{cases} a(\theta) & i = 0 \\ b(\theta)\theta_i & i = 1, 2, \dots, D \end{cases} \quad (19)$$

The  $a$  component of  $g$  corresponds to adding a constant to  $R$ , while the  $b$  component corresponds to multiplying  $R$  by a constant.

Note that if we had instead chosen  $R = \sum_{i=1}^D \theta_i S_i$ , i.e. left out the constant component  $\theta_0$ , then there would be only one diffeomorphic mode, corresponding to multiplication of  $R$  by a constant. This fact will be used when we analyze the Gaussian selection model in Section 8.

<sup>1</sup> Here, as throughout this paper, we restrict our attention to situations in which  $R$  is a scalar. The case of vector-valued model predictions  $R$  is worked out in [7].

### 6.3 Diffeomorphic modes of a biophysical model of transcriptional regulation

Diffeomorphic modes can become less trivial in more complicated situations. Consider the biophysical model of transcriptional regulation by the *E. coli lac* promoter (Fig. 2). This model was fit to Sort-Seq data in [5]. The form of this model is as follows. Let  $S$  denote a  $4 \times D$  matrix representing a DNA sequence of length  $D$  and having elements

$$S_{bl} = \begin{cases} 1 & \text{if base } b \text{ occurs at position } l \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

where  $b \in \{A, C, G, T\}$  and  $l = 1, 2, \dots, D$ . The binding energy  $Q$  of CRP to DNA was modeled in [5] as an “energy matrix”: each position in the DNA sequence was assumed to contribute additively to the overall energy. Specifically,

$$Q = \sum_{bl} \theta_Q^{bl} S_{bl} + \theta_Q^0 \quad (21)$$

where  $\theta_Q = \{\theta_Q^0, \theta_Q^{bl}\}$  are the parameters of this energy matrix. Similarly, the binding energy  $P$  of RNAP to DNA was modeled as an energy matrix

$$P = \sum_{bl} \theta_P^{bl} S_{bl} + \theta_P^0. \quad (22)$$

Both energies were taken to be in thermal units ( $k_B T$ ). The rate of transcription  $T$  resulting from these binding energies was taken to be proportional to the occupancy of RNAP at its binding site. This is given by

$$T = \frac{1}{1 + R^{-1}} \quad \text{where} \quad R = e^{-P} \frac{1 + e^{-Q-\gamma}}{1 + e^{-Q}}. \quad (23)$$

where  $\gamma$  is the interaction energy between CRP and RNAP (again in units of  $k_B T$ ).

Because the binding sites for CRP and RNAP do not overlap, one can learn the parameters  $\theta_Q$  and  $\theta_P$  from data separately by independently maximizing  $I[Q; M]$  and  $I[P; M]$ . The results of doing this are shown in Fig. 2b,c. Note, however, that the overall scale of each energy matrix remains undetermined, as does each chemical potential ( $\theta_P^0$  and  $\theta_Q^0$ ). The reason is that the energy scale and chemical potential are diffeomorphic modes of energy matrix models and therefore cannot be inferred by maximizing mutual information.

However, if  $Q$  and  $P$  are inferred together by maximizing  $I[T; M]$  instead, one is now able to learn both energy matrices with a physically meaningful energy scale. The chemical potential of CRP,  $\theta_Q^0$ , is also determined. The only parameter left unspecified is the chemical potential of RNA polymerase,  $\theta_P^0$ . The reason for this is that, while  $Q$  and  $P$  together have four diffeomorphic modes,  $T$  only has one. In the formula for  $T$ , the energies  $P$  and  $Q$  combine in a nonlinear way. This nonlinearity eliminates three of the four diffeomorphic modes. See [7] for the derivation of this result.

### 6.4 Conjugate noise modes

Diffeomorphic modes can be thought of as being “conjugate” to certain modes of the noise function. Consider a diffeomorphic transformation  $\theta \rightarrow \theta'$ . Because this transformation is diffeomorphic, the predictions  $R'$  of  $\theta'$  will be related to the predictions  $R$  of  $\theta$  by some invertible function  $f$ :

$$R' = f(R). \quad (24)$$

Now suppose that, at the same time we perform the transformation  $\theta \rightarrow \theta'$ , we also transform our noise function  $\pi \rightarrow \pi'$  where  $\pi'$  is defined by

$$\pi'(M|R) = \pi(M|f^{-1}(R)). \quad (25)$$

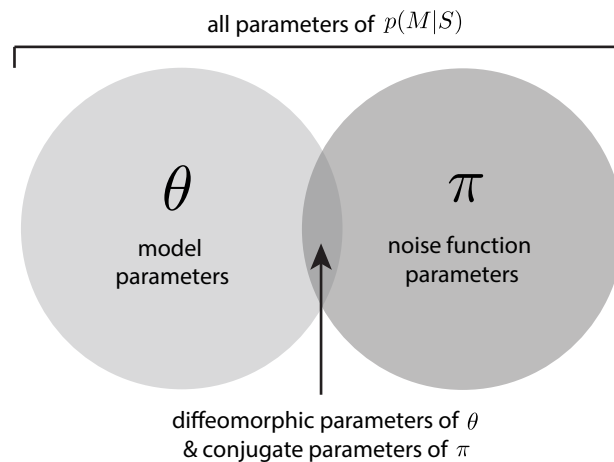


**Fig. 5** Venn diagram illustration of which parameters control which degrees of freedom of  $p(M|S)$ . Typically, altering the model  $\theta$  will change  $p(M|S)$  in a way that cannot be achieved by altering the noise function  $\pi$ . However, changes to the diffeomorphic parameters in  $\theta$  will change  $p(M|S)$  in a manner identical to changing the conjugate parameters of  $\pi$ . Thus, it is useful to think of diffeomorphic model parameters and their conjugate noise function parameters as lying within the intersection of  $\theta$  and  $\pi$ .

The resulting transformation  $(\theta, \pi) \rightarrow (\theta', \pi')$  leaves  $p(M|S)$  invariant, since

$$p(M|S) = \pi'(M|\theta'(S)) = \pi(M|\theta(S)) \quad (26)$$

for all sequences  $S$  and measurement values  $M$ . In what follows, we refer to the noise function transformation  $\pi \rightarrow \pi'$  as being conjugate to the model transformation  $\theta \rightarrow \theta'$ .

It is useful to think of the diffeomorphic parameters of  $\theta$  and the conjugate parameters of  $\pi$  in terms of the Venn diagram shown in Fig. 5. Typically, changes to most of the parameters of  $\theta$  will produce transformations of  $p(M|S)$  that cannot be achieved by any possible changes to  $\pi$ . The reverse is also true. However, changes to any diffeomorphic parameter of  $\theta$  will produce the same transformations of  $p(M|S)$  as a corresponding change to the conjugate parameter of  $\pi$ . Thus, if we think of  $\theta$  and  $\pi$  in terms of the transformations of  $p(M|S)$  that they generate, then the diffeomorphic parameters of  $\theta$  and the conjugate parameters of  $\pi$  are seen to lie within the intersection of  $\theta$  and  $\pi$ .

This finding has an intuitive interpretation: there is ambiguity in how we divide our experiment up into an activity model  $\theta$  and a noise function  $\pi$ . This ambiguity is parameterized by the diffeomorphic modes of  $\theta$  and their conjugate parameters of  $\pi$ . In what follows, we will see that being cognizant of this ambiguity is critical for correctly inferring activity models from data.

## 7 Inference using likelihood, noise-averaged likelihood, and mutual information

The diffeomorphic modes of parametric models respond to data in different ways than do nondiffeomorphic modes. This is illustrated in Fig. 6. Consider the range of parameter values consistent with the posterior distribution

$$p(\theta|\text{data}) \sim e^{NL(\theta, \pi)}, \quad (27)$$

computed assuming various noise functions  $\pi$ . If one assumes the correct noise function  $\pi^*$ , the inferred values for all parameters of  $\theta$  will be consistent with the true parameter values of  $\theta^*$ . Moreover, the uncertainty in all of these parameters will decrease as  $\delta\theta \sim N^{-1/2}$ .

If we instead assume an incorrect noise function, the uncertainty in all parameters will still vary as  $\delta\theta \sim N^{-1/2}$ . However, the use of an incorrect noise function will lead to bias in the inferred  $\theta$ . Which biases occur depend on which aspects of the noise function are wrong.

If we use a noise function  $\pi'$  that differs from  $\pi^*$  in a parameter conjugate to a diffeomorphic parameter of  $\theta$ , then that diffeomorphic parameter will be biased away from  $\theta^*$  in order to counteract the inaccuracy of  $\pi'$ . So likelihood-based inference will give a wrong answer, but perhaps this state of

affairs isn't too severe: the inferred model  $\theta^o$  will deviate from  $\theta^*$  in the values of parameters that we know we cannot accurately infer anyway.

However, if we use a noise function  $\pi''$  that differs from  $\pi^*$  in one or more parameters that are not conjugate to any of the diffeomorphic modes of  $\theta$ , then the inferred model  $\theta^o$  will, in general, deviate from  $\theta^*$  in the values of both diffeomorphic and nondiffeomorphic parameters. This is genuinely problematic. It means that any incorrect assumptions about the non-conjugate aspects of  $\pi$  can be expected to result in incorrect values for any of the parameters of  $\theta$ . We therefore see that performing likelihood-based inference with an assumed noise function can, in the large  $N$  limit, carry a substantial risk of error propagating to the model  $\theta$  that one wishes to learn.

The proper Bayesian approach to dealing with uncertainty in  $\pi$  is to formalize one's uncertainty by adopting an explicit prior  $p(\pi)$ , then computing the noise-averaged likelihood described above. This results in a situation schematized in Fig. 6d. Doing this will place tight constraints ( $\delta\theta \sim N^{-1/2}$ ) along all of the nondiffeomorphic modes of  $\theta$ . However, along diffeomorphic modes one finds weak constraints that do not decrease to zero in the  $N \rightarrow \infty$  limit. What constraints there are arise only from the choice of the noise function prior  $p(\pi)$ .

We therefore see that diffeomorphic model parameters and nondiffeomorphic model parameters respond to data in fundamentally different ways. Uncertainty in the values of nondiffeomorphic parameters decrease as  $N^{-1/2}$  in the large  $N$  limit, and this is true regardless of whether one has accurate information about the noise function  $\pi$ . Uncertainty in the values of diffeomorphic parameters, however, is constrained entirely by prior knowledge of the noise function  $\pi$  when  $N$  is large. Unless the correct noise function is known with absolute certainty, this difference in behavior will, in the  $N \rightarrow \infty$  limit, cause likelihood based inference to underestimate the uncertainty in all diffeomorphic parameters. This fact is a fundamental aspect of all statistical regression problems.

Unlike likelihood, which is readily computed, it is unclear how one should compute noise-averaged likelihood, or even what sort of priors  $p(\pi)$  might make sense. The similarity of noise-averaged likelihood and mutual information, however, suggests that one might instead perform inference using

$$p(\theta|\text{data}) \sim e^{NI(\theta)} \quad (28)$$

where  $I(\theta) = I[R; M]$  is an estimate of the mutual information between model predictions  $R$  and measurements  $M$ . This was the strategy used in [5] to fit the biophysical model shown in Fig. 2. If mutual information can be accurately computed to  $O(1/N)$ , then such inference will produce the constraints depicted in Fig. 6e: tight constraints on nondiffeomorphic parameters and no constraints whatsoever on diffeomorphic parameters. Doing this eliminates three problems: the need to posit an explicit noise function prior  $p(\pi)$ , the need to integrate over all noise functions, and potential concerns about the effects of the prior  $p(\pi)$  on the inferred values of diffeomorphic model parameters. In this way, using mutual information in place of per-datum log noise-averaged likelihood more accurately reflects the knowledge we gain about  $\theta$  from the data.

The downside of using Eq. (28) is that one needs to be able to estimate mutual information to an accuracy of  $\sim N^{-1}$ . At present it remains unclear how to do this reliably. Simple approaches, such as the kernel smoothing method used in [5], do appear to work well in practice on real data. However, the accuracy of this sampling approach has yet to be systematically evaluated.

## 8 Worked example: Gaussian selection

The above principles can be illustrated in the following analytically tractable model of a Sort-Seq experiment, which we call the ‘‘Gaussian selection model’’. In this model, our experiment starts with a large library of ‘‘DNA’’ sequences  $S$ , each of which is a  $D$ -dimensional vector drawn from a Gaussian probability distribution

$$p_{\text{lib}}(S) = (2\pi)^{-D/2} \exp\left(-\frac{|S - \mu|^2}{2}\right). \quad (29)$$

Here,  $\mu$  is a  $D$ -dimensional vector defining the average sequence in the library. From this library we extract sequences into two batches, labeled  $M = 0$  and  $M = 1$ . We fill the  $M = 0$  batch with

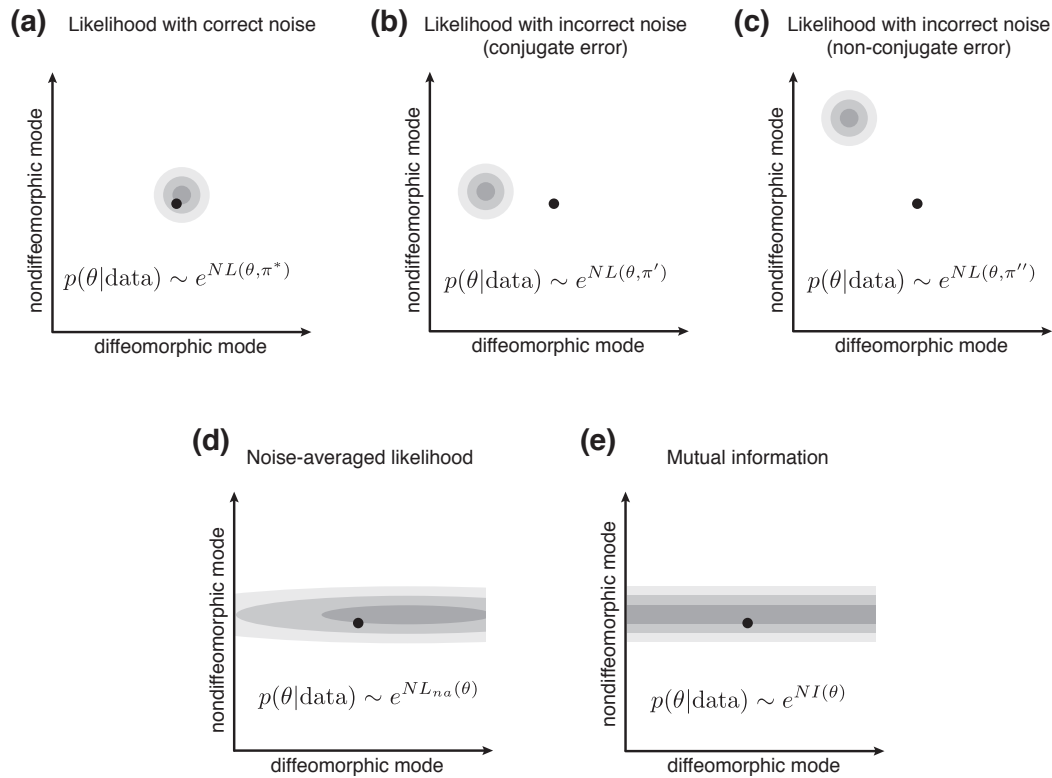


**Fig. 6** Schematic illustration of the posterior distribution  $p(\theta|\text{data})$  (gray shaded area), computed using different objective functions, in relation to the correct model  $\theta^*$  (dot). (a) Using likelihood with the correct noise function  $\pi^*$  leads to inference consistent with  $\theta^*$ . (b) Using likelihood with a noise function  $\pi'$  that differs from  $\pi^*$  only in conjugate parameters typically recovers nondiffeomorphic parameters that are consistent with  $\theta^*$  along with diffeomorphic parameters that are inconsistent with  $\theta^*$ . (c) Using likelihood with a noise function  $\pi''$  that differs from  $\pi^*$  in non-conjugate parameters typically leads to inference of  $\theta$  that is inconsistent with  $\theta^*$  in both diffeomorphic and nondiffeomorphic parameters. (d) Using noise-averaged likelihood results in strong constraints on nondiffeomorphic parameters and weak constraints on diffeomorphic parameters. The constraints on the diffeomorphic parameters are determined by the choice of the noise function prior  $p(\pi)$ . (e) Using mutual information is in principle similar to using noise-averaged likelihood, except that it provides no constraints whatsoever on diffeomorphic parameters. The problem of estimating mutual information to sufficient precision, however, remains. Figure adapted from [7].

sequences sampled indiscriminately from the library, and fill the  $M = 1$  batch with sequences sampled with relative probability

$$\frac{p(M = 1|S)}{p(M = 0|S)} = \exp(a^* + b^* R^*) \quad (30)$$

where the activity  $R^*$  is defined as the dot product of the sequence  $S$  with some “correct” model  $\theta^*$  (which is also a  $D$ -dimensional vector), i.e.

$$R^* = S^T \theta^*. \quad (31)$$

We will use  $N_M$  to denote the number of sequences in each batch  $M$ , along with  $N = N_0 + N_1$ . All of our calculations will take place in the limit where  $N_1$  is large but for which  $N_0$  is far larger. We will use  $\epsilon$  to denote the ratio

$$\epsilon = \frac{p(M = 1)}{p(M = 0)} = \frac{N_1}{N_0}, \quad (32)$$

All calculations will be carried out only to first order in  $\epsilon$ . This model Sort-Seq experiment is illustrated in Fig. 7.

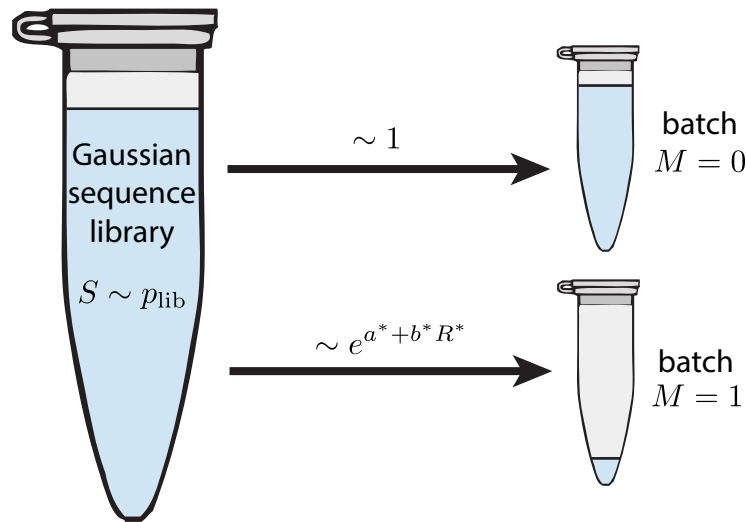


**Fig. 7** Illustration of the Gaussian selection model. We start with a library of sequences  $S$  distributed according to  $p_{\text{lib}}(S)$ .  $N_0$  sequences are sampled indiscriminately into batch 0.  $N_1$  sequences are sampled into batch 1 with relative weights  $\exp(a^* + b^* R^*)$ , where  $R^* = S^T \theta^*$ . All calculations are performed in the  $N_0 \gg N_1$  limit, as suggested by the liquid volumes shown in the two tubes.

Our goal is this: given the sequences in the two batches, recover the parameters  $\theta^*$  defining the sequence-function relationship for  $R^*$ . To do this, we adopt the following model for the sequence-dependent activity:

$$R = S^T \theta, \quad (33)$$

where  $\theta$  is the  $D$ -dimensional vector of parameters we wish to infer, and  $R$  is the predicted activity. From the arguments above and in [7], it is readily seen that the magnitude of  $\theta$ , i.e.  $|\theta|$ , is the only diffeomorphic mode of the model: changing this parameter rescales  $R$ , which preserves rank order.

### 8.1 Batch-specific distributions

From this model, we can readily calculate the sequence distribution  $p(S|M)$  in each batch, as well as the model prediction distribution  $p(R|M)$  for any hypothesized model  $\theta$ . From the selection procedure described above and illustrated in Fig. 7. Since sequences are sampled for batch 0 are indiscriminately from  $p_{\text{lib}}$ ,

$$p(S|M=0) = p_{\text{lib}}(S) = (2\pi)^{-D/2} \exp\left(-\frac{|S - \mu|^2}{2}\right) \quad (34)$$

Using Bayes's theorem, we can then use this to compute the distribution of sequences in batch 1 as follows.

$$p(S|M=1) = \frac{p(M=0) p(M=1|S)}{p(M=1) p(M=0|S)} p(S|M=0) \quad (35)$$

$$= \epsilon^{-1} e^{a^* + b^* R^*} (2\pi)^{-D/2} \exp\left(-\frac{|S - \mu|^2}{2}\right). \quad (36)$$

Completing the square in the exponent, we obtain

$$p(S|M=1) = (2\pi)^{-D/2} \exp\left(-\frac{|S - \mu - b^* \theta^*|^2}{2}\right). \quad (37)$$

Moreover, from the normalization requirement on  $p(S|M=1)$ , we find that  $\epsilon$  is related to  $a^*$ ,  $b^*$ , and  $\theta^*$  via

$$\epsilon = \exp \left( a^* + b^* \mu^T \theta^* + \frac{b^{*2} |\theta^*|^2}{2} \right). \quad (38)$$

We compute the distribution of model predictions for each batch as follows. For each  $M$ , this distribution is defined as

$$p(R|M) = \int dS \delta(R - \theta^T S) p(S|M). \quad (39)$$

This can be analytically calculated for both of the batches owing to the Gaussian form of each. We find that

$$p(R|M=0) = \frac{1}{\sqrt{2\pi}|\theta|} \exp \left( -\frac{(R - \mu^T \theta)^2}{2|\theta|^2} \right), \quad (40)$$

$$p(R|M=1) = \frac{1}{\sqrt{2\pi}|\theta|} \exp \left( -\frac{(R - [\mu + b^* \theta^*]^T \theta)^2}{2|\theta|^2} \right). \quad (41)$$

See Appendix for details.

## 8.2 Likelihood

To compute likelihood, we must posit a noise function  $\pi(M|R)$ . Based on our prior knowledge of the selection procedure, we choose  $\pi(M|R)$  so that

$$\frac{\pi(M=1|R)}{\pi(M=0|R)} = \exp(a + bR), \quad (42)$$

where  $a$  and  $b$  are scalar parameters that we may or may not know *a priori*. This, combined with the normalization requirement  $\sum_M \pi(M|R) = 1$ , gives

$$\pi(M=1|R) = \frac{e^{a+bR}}{1 + e^{a+bR}}, \quad \pi(M=0|R) = \frac{1}{1 + e^{a+bR}}. \quad (43)$$

This noise function  $\pi$  is correct when  $a = a^*$  and  $b = b^*$ . The parameter  $b$  is conjugate to the diffeomorphic mode  $|\theta|$ , while the parameter  $a$  is not conjugate to any diffeomorphic mode.

Using this noise function, the per-datum log likelihood  $L$  becomes a function of  $\theta$ ,  $a$ , and  $b$ . We compute this quantity as follows:

$$L(\theta, a, b) = \frac{1}{N} \log \left[ \prod_{n=1}^N \pi(M_n|R_n) \right] \quad (44)$$

$$= \frac{1}{N} \sum_{n|M_n=1} \log [e^{a+bR_n}] - \frac{1}{N} \sum_{\text{all } n} \log [1 + e^{a+bR_n}] \quad (45)$$

$$\approx \epsilon \langle a + bR \rangle_{M=1} - \langle \log[1 + e^{a+bR}] \rangle_{\text{all}} \quad (46)$$

$$\approx \epsilon \langle a + bR \rangle_{M=1} - \langle e^{a+bR} \rangle_{\text{all}} \quad (47)$$

$$\approx \epsilon \langle a + bR \rangle_{M=1} - \langle e^{a+bR} \rangle_{M=0}. \quad (48)$$

Here,  $n$  indexes the sequence-measurement pairs  $(M_n, S_n)$  that comprise our data.  $R_n = S_n^T \theta$  indicates the model prediction for sequence  $S_n$ , and subscripts on the angle brackets indicate averages taken over either  $p(S|M=1)$ ,  $p(S|M=0)$ , or  $p(S)$ . In going from Eq. (46) to Eq. (47), we have used the assumption that  $e^{a+bR} \ll 1$ , which we expect to hold in the  $\epsilon \rightarrow 0$  limit. In the last step, we approximate the expectation over all sequences by the expectation over sequences only within batch

0. Applying the expressions for  $p(R|M)$  derived above, we find that the per-datum log likelihood is, in the large  $N$  limit, given by

$$L(\theta, a, b) = \epsilon[a + b\theta^T \mu + bb^* \theta^T \theta^*] - \exp\left(a + b\theta^T \mu + \frac{b^2 |\theta|^2}{2}\right). \quad (49)$$

We now consider the consequences of various approaches for using  $L(\theta, a, b)$  to estimate  $\theta^*$ . In each case, the inferred optimum will be denoted by a superscript ‘o.’ Standard regression requires that we assume a specific value for  $a$  and for  $b$ , then optimize  $L(\theta, a, b)$  over  $\theta$  alone. Setting

$$0 = \frac{\partial L}{\partial \theta_i} \Big|_{\theta^o, a, b} \quad (50)$$

for each component  $i$ . By this criteria we find that the optimal model  $\theta^o$  is given by a linear combination of  $\theta^*$  and  $\mu$ :

$$\theta^o = \frac{cb^*}{b} \theta^* + \frac{c-1}{b} \mu, \quad (51)$$

where  $c$  is a scalar defined by the transcendental equation

$$c = \exp\left[(a^* - a) + \frac{(1-c^2)}{2} \left(|b^* \theta^* + \mu|^2 + \frac{|\mu|^2}{1+c}\right)\right]. \quad (52)$$

See Appendix for details. Note that  $c$  is determined only by the value of  $a$  and not by the value of  $b$ . Moreover,  $c = 1$  if and only if  $a = a^*$ .

If our assumed noise function is correct, i.e.,  $a = a^*$  and  $b = b^*$ , then

$$\theta^o = \theta^*. \quad (53)$$

Thus, maximizing likelihood will identify the correct model parameters. This exemplifies the general behavior illustrated in Fig. 6a.

If  $a = a^*$  but  $b \neq b^*$ , our assumed noise function will differ from the correct one only in a manner conjugate to the diffeomorphic mode  $|\theta|$ . In this case we find that

$$\theta^o = \frac{b^*}{b} \theta^*, \quad (54)$$

and thus  $\theta^o$  is proportional but not equal to  $\theta^*$ . This comports with our claim above that the diffeomorphic mode of the inferred model, i.e.  $|\theta^o|$ , will be biased so as to compensate for the error in the conjugate parameter  $b$  of the noise function. This finding follows the behavior described in Fig. 6b.

If  $a \neq a^*$ , however,  $c \neq 1$ . As a result,  $\theta^o$  is a nontrivial linear combination of  $\theta^*$  and  $\mu$  and will thus point in a different direction than  $\theta^*$ . This is true regardless of the value of  $b$ . This behavior is illustrated in Fig. 6c: errors in non-conjugate parameters of the noise function will typically lead to errors in nondiffeomorphic parameters of the model.

We now consider the error bars that likelihood places on model parameters. Setting  $\theta = \theta^o + \delta\theta$  and expanding  $L(\theta, a, b)$ , we find that

$$NL(\theta, a, b) = NL(\theta^o, a, b) - \frac{N_1 b^2}{2} \sum_{i,j} \Lambda_{ij} \delta\theta_i \delta\theta_j, \quad (55)$$

where  $\Lambda_{ij} = \delta_{ij} + (\mu_i + b\theta_i^o)(\mu_j + b\theta_j^o)$ . Note that all eigenvalues of  $\Lambda$  are greater or equal to 1. Adopting the posterior distribution

$$p(\theta|\text{data}) \sim e^{NL(\theta, a, b)} \quad (56)$$

therefore gives a covariance matrix on  $\theta$  of

$$\langle \delta\theta_i \delta\theta_j \rangle = \frac{\Lambda_{ij}^{-1}}{N_1 b^2}. \quad (57)$$

Thus,  $\delta\theta \sim N_1^{-1/2}$  in all directions of  $\theta$ -space. This is illustrated in Figs. 6a-c. The dependence of these fluctuations on  $N_1$  does not depend on whether the noise function  $\pi$  is correct. Therefore, when the noise function is incorrect, the finite bias introduced into  $\theta^o$  will cause  $\theta^*$  to fall outside the inferred error bars for sufficiently large  $N$ .

### 8.3 Mutual information

In the  $\epsilon \rightarrow 0$  limit, Eq. (7) simplifies to

$$I(\theta) = \epsilon \int dR p(R|M=1) \log \frac{p(R|M=1)}{p(R|M=0)}. \quad (58)$$

The right hand side can be evaluated exactly using Eq. (40) and Eq. (41):

$$I(\theta) = \frac{\epsilon b^{*2}}{2} \frac{(\theta^T \theta^*)^2}{|\theta|^2}. \quad (59)$$

See Appendix for details. Note that the expression on the right is invariant under rescaling of  $\theta$ . This reflects the fact that  $|\theta|$  is a diffeomorphic mode of the model defined in Eq. (33).

To find the model  $\theta^o$  that maximizes mutual information, we set

$$0 = \frac{\partial I}{\partial \theta_i} \Big|_{\theta^o} = \frac{\epsilon b^{*2} \theta^{oT} \theta^*}{|\theta^o|^2} \left[ \theta_i^* - \theta_i^o \frac{\theta^{oT} \theta^*}{|\theta^o|^2} \right] \quad (60)$$

The optimal model  $\theta^o$  must therefore be parallel to  $\theta^*$ , i.e.

$$\theta^o \propto \theta^*. \quad (61)$$

Expanding about  $\theta = \theta^o + \delta\theta$  as above, we find that

$$NI(\theta) = NI(\theta^o) - \frac{N_1 b^{*2}}{2} (\delta\theta_{\perp})^2 \quad (62)$$

where  $\delta\theta_{\perp}$  is the component of  $\delta\theta$  perpendicular to  $\theta^*$ . Therefore, if we use the posterior distribution

$$p(\theta|\text{data}) \sim e^{NI(\theta)} \quad (63)$$

to infer  $\theta$ , we find uncertainties in directions perpendicular to  $\theta^*$  of magnitude  $N_1^{-1/2}$ . These error bars are only slightly larger than those obtained using likelihood, and have the same dependence on  $N$ . However, we find no constraint whatsoever on the component of  $\delta\theta$  parallel to  $\theta^*$ . These results are illustrated by Fig. 6e

### 8.4 Noise-averaged likelihood

We can also compute the per-datum log noise-averaged likelihood,  $L_{na}(\theta)$ , in the case of a uniform prior on  $a$  and  $b$ , i.e.  $p(\pi) = p(a, b) = \mathcal{C}$  where  $\mathcal{C}$  is an infinitesimal constant. We find that

$$\exp[NL_{na}(\theta)] = \int d\pi p(\pi) \exp[NL(\theta, \pi)] \quad (64)$$

$$= \mathcal{C} \int_{-\infty}^{\infty} da \int_{-\infty}^{\infty} db \exp \left( N_1 [a + b\theta^T \mu + bb^* \theta^T \theta^*] - N \exp \left[ a + b\theta^T \mu + \frac{b^2 |\theta|^2}{2} \right] \right) \quad (65)$$

$$= \mathcal{C} \Gamma(N_1) \sqrt{\frac{2\pi}{N|\theta|^2}} \exp \left( \frac{N_1 b^{*2}}{2} \frac{(\theta^T \theta^*)^2}{|\theta|^2} \right). \quad (66)$$

See the Appendix for details. Thus,

$$L_{na}(\theta) = I(\theta) - \frac{1}{N} \log |\theta| + \text{const}, \quad (67)$$

where the constant (which absorbs  $\mathcal{C}$  entirely) does not depend on  $\theta$ . Therefore, if we perform Bayesian inference using noise-averaged likelihood, i.e. using

$$p(\theta|\text{data}) \sim e^{NL_{na}(\theta)}, \quad (68)$$

we will find in the large  $N$  limit that  $\delta\theta_{\perp}$  is constrained in the same way as if we had used mutual information, i.e.

$$\theta^o \propto \theta^*. \quad (69)$$

However, the noise function prior we have assumed further results in weak constraints on  $|\theta|$  that do not tighten as  $N$  increases.<sup>2</sup> This is illustrated schematically in Fig. 6d.

## 9 Discussion

Understanding the features of sequence-function relationships in molecular biology is going to require the ability to glean quantitative models from high-throughput experiments with poorly characterized noise. Standard regression methods require that the user assume a specific noise function up front, then find the explanatory model that has maximum likelihood. However, if the noise function one assumes is incorrect, this will in general cause errors in the sequence-function relationships that one infers.

In this paper we have shown that maximizing mutual information instead of likelihood allows one to infer sequence-function relationships from large data sets without any assumptions about the quantitative form of experimental noise. On the other hand, maximizing mutual information may fail to pin down certain model parameters. These are called diffeomorphic parameters, and are constrained by data in a fundamentally different way than are nondiffeomorphic parameters. Such nondiffeomorphic parameters can be constrained to arbitrary precision (given enough data) without any knowledge of the experimental noise function. By contrast, diffeomorphic parameters are constrained only by knowledge of the noise function and the precision of these constraints does not increase with the amount of data. The relevant aspects of inference using likelihood, mutual information, and noise-averaged likelihood, were illustrated here in an analytically tractable “Gaussian selection” model.

The development of methods for learning sequence-function relationships from Sort-Seq data and data produced by similar experiments still requires much attention. We are only in the infancy of solving this machine learning problem. One practical issue is how to accurately estimate the mutual information used for inferring model parameters, and wholly satisfactory methods for doing this have yet to be described.

## 10 Appendix

### 10.1 Derivation of Eqs. 40 and 41

Here we describe how to compute  $p(R|M)$  where  $R = \theta^T S$ . We first consider the case of  $M = 0$ .

$$p(R|M=0) = \int dS p(S|M=0) \delta(R - S^T \theta) \quad (70)$$

$$= \int dS p(S|M=0) \delta([R - \mu^T \theta] - [S - \mu]^T \theta) \quad (71)$$

$$= \int dS p(S|M=0) \delta(R' - S'^T \theta) \quad (72)$$

where  $R' = R - \mu^T \theta$  and  $S' = S - \mu$ . Now, split  $S'$  up into the components parallel and perpendicular to  $\theta$ :

$$S' = S'_{\perp} + S'_{\parallel} \hat{\theta}, \quad (73)$$

<sup>2</sup> In the case at hand,  $|\theta^o|$  is pushed all the way to zero. This is an artifact of the simple flat prior  $p(a, b)$ . If we instead adopt a weak Gaussian prior on  $b$ , we can still carry out the computation of  $L_{na}$  analytically, and in this case we find that  $|\theta^o|$  is finite.

where  $S'_\perp$  is a vector of dimension  $D - 1$ ,  $S'_\parallel$  is a scalar, and  $\hat{\theta} = \theta/|\theta|$ . Continuing the integration,

$$p(R|M=0) = \int dS'_\perp \int_{-\infty}^{\infty} dS'_\parallel \delta(R' - S'_\parallel |\theta|) (2\pi)^{-D/2} \exp\left(-\frac{S'^2_\perp}{2} - \frac{S'^2_\parallel}{2}\right) \quad (74)$$

$$= \int_{-\infty}^{\infty} dS'_\parallel \delta(R' - S'_\parallel |\theta|) (2\pi)^{-1/2} \exp\left(-\frac{S'^2_\parallel}{2}\right) \quad (75)$$

$$= \int_{-\infty}^{\infty} dS'_\parallel \delta\left(\frac{R'}{|\theta|} - S'_\parallel\right) |\theta|^{-1} (2\pi)^{-1/2} \exp\left(-\frac{S'^2_\parallel}{2}\right) \quad (76)$$

$$= |\theta|^{-1} (2\pi)^{-1/2} \exp\left(-\frac{R'^2}{2|\theta|^2}\right). \quad (77)$$

Thus we find

$$p_0(R) = \frac{1}{\sqrt{2\pi}|\theta|} \exp\left(-\frac{(R - \mu^T \theta)^2}{2|\theta|^2}\right). \quad (78)$$

To compute  $p(R|M=1)$ , we just replace  $\mu \rightarrow \mu + b^* \theta^*$ , giving

$$p(R|M) = \frac{1}{\sqrt{2\pi}|\theta|} \exp\left(-\frac{(R - [\mu + b^* \theta^*]^T \theta)^2}{2|\theta|^2}\right). \quad (79)$$

## 10.2 Derivation of Eqs. 51 and 52

Here we show how to derive the optimal  $\theta$  for  $L(\theta, a, b)$ , with  $a$  and  $b$  fixed. Setting the variation of  $L$  to zero,

$$0 = \left. \frac{\partial L}{\partial \theta_i} \right|_{\theta^o} = \epsilon b(\mu_i + b^* \theta_i^*) - b(\mu_i + b \theta_i^o) \exp\left(a + b \mu^T \theta^o + \frac{b^2 |\theta^o|^2}{2}\right). \quad (80)$$

This gives

$$\mu_i + b \theta_i^o = \epsilon(\mu_i + b^* \theta_i^*) \exp\left(-a - b \mu^T \theta^o - \frac{b^2 |\theta^o|^2}{2}\right) \quad (81)$$

$$= c(\mu_i + b^* \theta_i^*) \quad (82)$$

where  $c$  is a constant satisfying

$$c = \epsilon \exp\left(-a - b \mu^T \theta^o - \frac{b^2 |\theta^o|^2}{2}\right) = \exp\left([a^* - a] + \mu^T [b^* \theta^* - b \theta^o]^T + \frac{b^{*2} |\theta^*|^2 - b^2 |\theta^o|^2}{2}\right) \quad (83)$$

Note that the right hand side of the above equation depends implicitly on  $c$  through the value of  $\theta$ . Each component  $\theta_i$  must therefore satisfy

$$b \theta_i^o = c b^* \theta_i^* + (c - 1) \mu_i, \quad (84)$$

which gives Eq. (51). Using this formula for  $b \theta^o$ , we find that

$$\mu^T [b^* \theta^* - b \theta^o] + \frac{b^{*2} |\theta^*|^2 - b^2 |\theta^o|^2}{2} = \frac{1 - c^2}{2} |b^* \theta^* + \mu|^2. \quad (85)$$

Plugging this into Eq. (83) gives Eq. (52).

### 10.3 Derivation of Eqs. 58 and 59

We derive Eq. (58) as follows. To ease notation a bit, we define  $p_M(R) = p(R|M)$ .

$$I[R; M] = \sum_{M=0,1} \int dR p(M, R) \log \frac{p_M(R)}{p(R)} \quad (86)$$

$$= p(M=0) \int dR p_0(R) \log \frac{p_0(R)}{p(R)} + p(M=1) \int dR p_1(R) \log \frac{p_1(R)}{p(R)} \quad (87)$$

$$= p(M=1) \int dR p_1(R) \log \frac{p_1(R)}{p_0(R)} + \int dR p(R) \log \frac{p_0(R)}{p(R)} \quad (88)$$

the first term in Eq. (88) is the right hand side of Eq. (58). We will now show that the second term is of order  $\epsilon^2$ , and can therefore be ignored. Rearranging

$$p(R) = (1 - \epsilon)p_0(R) + \epsilon p_1(R) \quad (89)$$

gives

$$p_0(R) = \frac{p(R) - \epsilon p_1(R)}{1 - \epsilon}. \quad (90)$$

Plugging this into the second term of Eq. (88) gives

$$\int dR p(R) \log \frac{p_0(R)}{p(R)} = \int dR \left[ \log \frac{1}{1 - \epsilon} \left( 1 - \epsilon \frac{p_1(R)}{p(R)} \right) \right] \quad (91)$$

$$= \int dR p(R) \log \left[ 1 + \epsilon \left( 1 - \frac{p_1(R)}{p(R)} \right) + O(\epsilon^2) \right] \quad (92)$$

$$= \epsilon \int dR p(R) \left( 1 - \frac{p_1(R)}{p(R)} \right) + O(\epsilon^2) \quad (93)$$

$$= O(\epsilon^2). \quad (94)$$

Eq. (59) is derived as follows:

$$I(\theta) = \epsilon \left\langle \log \frac{p(R|M=1)}{p(R|M=0)} \right\rangle_{M=1} \quad (95)$$

$$= \epsilon \left\langle \frac{(R - \mu^T \theta)^2}{2|\theta|^2} - \frac{([R - \mu^T \theta] - b^* \theta^T \theta^*)^2}{2|\theta|^2} \right\rangle_{M=1} \quad (96)$$

$$= \frac{\epsilon}{2|\theta|^2} \langle 2[R - \mu^T \theta] b^* \theta^T \theta^* - (b^* \theta^T \theta^*)^2 \rangle_{M=1} \quad (97)$$

$$= \frac{\epsilon}{2|\theta|^2} (2[\langle R \rangle_{M=1} - \mu^T \theta] b^* \theta^T \theta^* - (b^* \theta^T \theta^*)^2) \quad (98)$$

$$= \frac{\epsilon}{2|\theta|^2} (2[b^* \theta^T \theta^*] b^* \theta^T \theta^* - (b^* \theta^T \theta^*)^2) \quad (99)$$

$$= \frac{\epsilon b^{*2}}{2} \frac{(\theta^T \theta^*)^2}{|\theta|^2}. \quad (100)$$

### 10.4 Derivation of Eq. 66

Here we show how to evaluate the equation, Eq. (65), for the noise-averaged likelihood  $e^{NL_{na}(\theta)}$ . First, interchange the order of integration and define  $a' = a + b\theta^T \mu$ . This gives,

$$e^{NL_{na}(\theta)} = \mathcal{C} \int_{-\infty}^{\infty} db \int_{-\infty}^{\infty} da \exp \left[ N_1 [a' + bb^* \theta^T \theta^*] - N \exp \left( a' + \frac{b^2 |\theta|^2}{2} \right) \right]. \quad (101)$$

Next, define  $M = N \exp\left(\frac{b^2|\theta|^2}{2}\right)$ ,  $u = Me^{a'}$ , and so  $e^{a'} = u/M$ ,  $e^{a'} da' = du/M$ . This gives

$$e^{NL_{na}(\theta)} = \mathcal{C} \int_{-\infty}^{\infty} db e^{N_1 bb^* \theta^T \theta^*} \int_{-\infty}^{\infty} (e^{a'} da) (e^{a'})^{N_1-1} \exp[-Me^{a'}] \quad (102)$$

$$= \mathcal{C} \int_{-\infty}^{\infty} db e^{N_1 bb^* \theta^T \theta^*} M^{-N_1} \int_0^{\infty} du u^{N_1-1} \exp[-u] \quad (103)$$

$$= \mathcal{C} \Gamma(N_1) \int_{-\infty}^{\infty} db e^{N_1 bb^* \theta^T \theta^*} M^{-N_1} \quad (104)$$

$$= \mathcal{C} \Gamma(N_1) \int_{-\infty}^{\infty} db \exp\left[N_1 bb^* \theta^T \theta^* - N_1 \frac{b^2|\theta|^2}{2}\right] \quad (105)$$

$$= \mathcal{C} \Gamma(N_1) \int_{-\infty}^{\infty} db \exp\left[\frac{N_1|\theta|^2}{2} \left(2bb^* \frac{\theta^T \theta^*}{|\theta|^2} - b^2\right)\right] \quad (106)$$

$$= \mathcal{C} \Gamma(N_1) \int_{-\infty}^{\infty} db \exp\left[\frac{N_1|\theta|^2}{2} \left(\frac{b^{*2}(\theta^T \theta^*)^2}{|\theta|^4} - \left[b - \frac{b^* \theta^T \theta^*}{|\theta|^2}\right]^2\right)\right] \quad (107)$$

$$= \mathcal{C} \Gamma(N_1) \exp\left(\frac{N_1 b^{*2}}{2} \frac{(\theta^T \theta^*)^2}{|\theta|^2}\right) \int_{-\infty}^{\infty} db \exp\left(\frac{N_1|\theta|^2}{2} \left[b - \frac{b^* \theta^T \theta^*}{|\theta|^2}\right]^2\right) \quad (108)$$

$$= \mathcal{C} \Gamma(N_1) \sqrt{\frac{2\pi}{N_1|\theta|^2}} \exp\left(\frac{N_1 b^{*2}}{2} \frac{(\theta^T \theta^*)^2}{|\theta|^2}\right), \quad (109)$$

which is Eq. (66).

**Acknowledgements** We would like to thank L. Peliti, O. Revoire, and T. Mora for organizing this special issue. This work was supported by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

## References

1. B. Ren, F. Robert, J. Wyrick, O. Aparicio, E. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. Volkert, C. Wilson, S. Bell, R. Young, *Science* (New York, NY) **290**(5500), 2306 (2000)
2. S. Mukherjee, M. Berger, G. Jona, X. Wang, D. Muzzey, M. Snyder, R. Young, M. Bulyk, *Nat Genet* **36**(12), 1331 (2004)
3. Y. Zhao, D. Granas, G.D. Stormo, *PLoS Computational Biology* **5**(12), e1000590 (2009)
4. R.P. Patwardhan, C. Lee, O. Litvin, D.L. Young, D. Pe'er, J. Shendure, *Nature Biotechnology* **27**(12), 1173 (2009)
5. J.B. Kinney, A. Murugan, C.G. Callan, E.C. Cox, *Proc. Natl. Acad. Sci. USA* **107**(20), 9158 (2010)
6. T. Cover, J. Thomas, *Elements of information theory* (1st ed.) (1991)
7. J.B. Kinney, G.S. Atwal, *Neural Comput.* **26**(4), 637 (2014)
8. L. Paninski, *Network* **14**(3), 437 (2003)
9. T. Sharpee, N. Rust, W. Bialek, *Neural Computation* **16**(2), 223 (2004)
10. T. Sharpee, H. Sugihara, A. Kurgansky, S. Rebrink, M. Stryker, K. Miller, *Nature* **439**(7079), 936 (2006)
11. M. Kouh, T.O. Sharpee, *Network: Computation in Neural Systems* **20**(2), 49 (2009)
12. K. Rajan, A. Marre, G. Tkacik, *NIPS* (2011)
13. S. Kosuri, D.B. Goodman, G. Cambray, V.K. Mutalik, Y. Gao, A.P. Arkin, D. Endy, G.M. Church, *Proc Natl Acad Sci U S A* **110**(34), 14024 (2013)
14. E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, E. Segal, *Nature Biotechnology* **30**(6), 521 (2012)
15. P. Oikonomou, H. Goodarzi, S. Tavazoie, *Cell reports* **7**(1), 281 (2014)
16. S. Dutta, A.E. Keating, *Journal of Molecular Biology* (2014)
17. A.I. Podgornaia, M.T. Laub, *Science* (New York, NY) **347**(6222), 673 (2015)
18. R.P. Patwardhan, J.B. Hiatt, D.M. Witten, M.J. Kim, R.P. Smith, D. May, C. Lee, J.M. Andrie, S.I. Lee, G.M. Cooper, N. Ahituv, L.A. Pennacchio, J. Shendure, *Nature Biotechnology* **30**(3), 265 (2012)
19. A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C.G. Callan, J.B. Kinney, M. Kellis, E.S. Lander, T.S. Mikkelsen, *Nature Biotechnology* **30**(3), 271 (2012)
20. J.C. Kwasniewski, I. Mogno, C.A. Myers, J.C. Corbo, B.A. Cohen, *Proceedings of the National Academy of Sciences of the United States of America* **109**(47), 19498 (2012)
21. D.M. Fowler, C.L. Araya, S.J. Fleishman, E.H. Kellogg, J.J. Stephany, D. Baker, S. Fields, *Nature Methods* **7**(9), 741 (2010)

- 
22. M. Slattey, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, Cell (2011)
  23. J.B. Kinney, G.S. Atwal, Proc. Natl. Acad. Sci. USA (2014)
  24. J.B. Kinney, G. Tkacik, C.G. Callan, Proc. Natl. Acad. Sci. USA **104**(2), 501 (2007)