Fighting against uncertainty: an essential issue in bioinformatics

Michiaki Hamada

Submitted: 22nd March 2013; Received (in revised form): 15th May 2013

Abstract

Many bioinformatics problems, such as sequence alignment, gene prediction, phylogenetic tree estimation and RNA secondary structure prediction, are often affected by the 'uncertainty' of a solution, that is, the probability of the solution is extremely small. This situation arises for estimation problems on high-dimensional discrete spaces in which the number of possible discrete solutions is immense. In the analysis of biological data or the development of prediction algorithms, this uncertainty should be handled carefully and appropriately. In this review, I will explain several methods to combat this uncertainty, presenting a number of examples in bioinformatics. The methods include (i) avoiding point estimation, (ii) maximum expected accuracy (MEA) estimations and (iii) several strategies to design a pipeline involving several prediction methods. I believe that the basic concepts and ideas described in this review will be generally useful for estimation problems in various areas of bioinformatics.

Keywords: Uncertainty of solutions; estimation problems; bioinformatics; sequence analysis

INTRODUCTION

Uncertainty of solutions

In estimation problems that appear in bioinformatics, such as sequence alignment [1], gene prediction [2], RNA secondary structure prediction [3, 4], RNA–RNA interaction prediction [5] and phylogenetic tree estimation [6], it is typical that the probability of any solution is extremely small even if that solution has the highest probability. For example, the probabilities of the minimum free energy (MFE) structures of a telomerase RNA and a ribosomal RNA are $8 \times 10^{-6}$ and $4 \times 10^{-23}$, respectively (This example shows that one needs to be careful of predicted secondary structures for long RNA sequences using simple minimum free minimization. In practical, a method recently proposed by [7] (or [8]) would partially mitigate this issue) (Figure 1). Here, the probabilities were computed by using the Boltzmann distribution (also called the Gibbs distribution) from statistical mechanics [9]:

$$p(\theta|x) = \frac{1}{Z(x)} \exp \left( \frac{-E(\theta, x)}{RT} \right)$$

where $\theta$ is a specific secondary structure (e.g. the MFE structure) of an RNA sequence $x$, $E(\theta, x)$ is the free energy of secondary structure $\theta$, $T$ is the temperature, $R$ is the ideal gas constant (8.31 J/mol K) and $Z(x)$ is the normalizing constant (partition function) that ensures $\sum_{\theta} p(\theta|x) = 1$. The situation is much worse in the problem of RNA deleterious mutation prediction [10–12], in which a specific RNA secondary structure is sought from among all possible secondary structures with up to $k$ mutations, and the partition function thus becomes much larger (see Figure 4 in Barash and Churkin [10]). The fact that the probability of a solution can be extremely small is known as the uncertainty of solutions, which leads to a critical issue in developing prediction algorithms or analyzing biological data in the field of bioinformatics [13–17]. In one Science article [15], for instance, the authors argued that the uncertainty of multiple sequence alignment greatly influences phylogenetic topology estimations: phylogenetic topologies estimated from multiple alignments predicted by five widely used aligners are different from one another. There is no doubt...
that this raises serious issues for reaching biological conclusions.

**Where does the uncertainty come from?**
The uncertainty described in the previous section arises for estimation problems on high-dimensional discrete spaces, in which the number of possible solutions is immense. For instance, the following results for the asymptotic number of solutions in particular problems are known:

- The number of possible secondary structures, \( S(n) \), of an RNA sequence with length \( n \) is estimated to be
  \[
  S(n) \sim \sqrt{\frac{15 + 7\sqrt{5}}{8\pi}} n^{-1.5} \left(\frac{3 + \sqrt{5}}{2}\right)^n
  \]
  where \( f(n) \sim g(n) \) means that \( \lim_{n \to \infty} f(n)/g(n) = 1 \).
  (The number of secondary structures of a specific RNA sequence will generally be smaller than this value because, in the equation, it is assumed that any pair of nucleotides can form a base pair.). See Theorem 13.2 in [18] for the proof.

- Let \( f(n, m) \) be the number of alignments of one sequence of \( n \) letters with another of \( m \) letters. Then \( f(n, n) \) is estimated to be
  \[
  f(n, n) \sim (1 + \sqrt{2})^{2n+1} n^{-1/2}.
  \]
  See Theorem 9.1 in [18] for the proof.

- The number of topologies of a phylogenetic tree (un-rooted binary tree) with \( n(>3) \) leaves (formally called ‘operational taxonomy units’) is
  \[
  \prod_{j=3}^{n} (2j - 5) = 1 \cdot 3 \cdot 5 \cdots (2n - 5). \]
  See Proposition 14.1 in [18] for the proof.

In the equations described earlier in the text, the number of solutions increases exponentially with \( n \), therefore, for example, the number of alignments of two sequences with length 1000 is
\[
 f(1000, 1000) \sim 10^{767.4}.
\]
which is immense because it has been estimated that the number of particles in the universe is \( \sim 10^{80} \). Because the sum of the probabilities of all solutions must be equal to 1, the probability of any particular solution tends to become extremely small because of the large number of potential solutions. In addition, there often exist many similar solutions: for a given RNA secondary structure \( s \), secondary structures produced by unpairing a base pair (in \( s \)) are similar to the original secondary structure.

**Purpose of this review**
In this review, I aim to introduce several methods to handle this uncertainty appropriately in the field of bioinformatics, presenting a number of actual studies. Most examples are adopted from either RNA informatics or sequence alignments. The methods include (i) avoiding point estimation (e.g. prediction of suboptimal solutions), (ii) maximum expected accuracy (MEA) estimations and (iii) several strategies to design a pipeline involving several prediction methods.

Throughout this article, the following notation and definitions are used:

- \( Y \) denotes a predictive (solution) space, including all the possible solutions that are discrete. For example, \( Y \) is the set of possible RNA secondary structures of an RNA sequence \( x \).
- \( p(\theta|D) \) denotes a posterior probability distribution on a predictive space \( Y \), given data \( D \). For example, \( p(\theta|x) \) is a probability distribution of secondary structures of an RNA sequence \( D = \{x\} \) given by Equation (1).
- Predicting one solution from the predictive space \( Y \) is called point estimation. The probability of the solution found by point estimation is sometimes extremely low as described in the previous section.

This article is organized as follows. In the ‘Avoiding Point Estimations’ section, I describe methods to
avoid point estimation by sampling several suboptimal solutions or visualizing the distribution of solutions. In the ‘What Should We Do When One Prediction is Required?’ section, I introduce methods to predict one solution, taking the uncertainty into account, because there are several situations in which only one solution is required. In the ‘Handling Uncertainty to Develop Complex Pipelines’ section, I give some strategies for developing complex pipelines or algorithms in which several prediction methods are involved. In the ‘Discussion’ section, these approaches are discussed, and the direction of future research in areas related to this study is considered.

**AVOIDING POINT ESTIMATIONS**

**Prediction of suboptimal solutions**

One possible method to handle the uncertainty described in the ‘Introduction’ section is to predict not only the one optimal solution (There might be several optimal solutions whose score is exactly the same as the optimal score. In this case, prediction of all optimal solutions also raises an issue) but also several suboptimal solutions. Zuker [19], in a study of RNA secondary structure prediction, was the first to introduce a method to predict suboptimal solutions, implemented in Mfold [20] (see Supplementary Section SA for the details).

**Stochastic sampling with dynamic programming**

Stochastic sampling (SS) in combination with a dynamic programming (DP) technique enables the efficient sampling of solutions from a distribution \( p(0|D) \) of solutions. For both RNA secondary structure predictions and pairwise alignments, stochastic sampling is realized by stochastically conducting a traceback procedure in a DP algorithm [1, 21]. For example, RNAsubopt [22] predicts the complete set of suboptimal secondary structures within a given energy range (say, 1 kcal/mol) from the minimum free energy.

**MCMC and Gibbs sampling**

For a probability distribution whose probabilistic structure is more complicated, a Markov-chain Monte-Carlo (MCMC) algorithm [23], such as Gibbs sampling, is a widely applicable method to sample from the posterior distribution, although in most applications MCMC is used to obtain the optimized solution.

Meyer et al. [24] applied an MCMC method to optimize the joint probability distribution

\[
P(S, A, T|D) = \frac{1}{Z}P(D|S, A, T)P(S, A, T)
\]

where \( D \) stands for the data (i.e. unaligned RNA sequences), \( Z = P(D) \) is a normalizing constant (partition function), \( S \) is a consensus (or common) RNA secondary structure that might contain pseudoknots, \( A \) is a multiple sequence alignment and \( T \) is an evolutionary tree relating the sequences. It would be difficult not only to optimize this posterior probability using dynamic programming techniques but also to apply stochastic sampling with a dynamic programming described in the previous section because the probability structures are complex joint distributions of three states.

Metzler and Nebel [25] and Bon et al. [26] proposed a probabilistic model for RNA secondary structures with pseudoknots and presented an MCMC Method for sampling RNA structures according to their posterior distribution for a given sequence. In contrast to conventional RNA secondary structure prediction (which does not consider pseudoknots), RNA secondary structure prediction with pseudoknots entails a larger computational cost (see [27]), and the application of stochastic sampling is not realistic. Doose and Metzler [28] presented a method (PhyloQfold) that takes advantage of the evolutionary history of a group of aligned RNA sequences for sampling consensus secondary structures, including pseudoknots, according to their approximate posterior probability. Wei et al. [29] presented a new global structural alignment algorithm, RNAG, to predict consensus secondary structures for unaligned sequences, using a blocked Gibbs sampling algorithm.

For phylogenetic tree estimations, MrBayes [30] is a program for Bayesian phylogenetic analysis, which uses MCMC techniques to sample from the posterior probability distribution of phylogenetic trees for a given multiple sequence alignment. More recently, the program BigFoot [31] used an MCMC method for finding joint distributions of phylogenetic trees and multiple sequence alignments (MSA).

As shown by the aforementioned examples, MCMC and Gibbs sampling seem useful methods to sample suboptimal solutions when the structure of target probability distribution is complex.

**Non-stochastic approaches**

Typically, the approach described in the previous sections returns a huge number of secondary structures in which there are many structures that are
similar to each other (compare Figure 2). To overcome this, a method to predict representative suboptimal secondary structures, such as locally optimal secondary structures or alignments, has been proposed [19, 32, 33] (compare Supplementary Section SA). Moreover, for RNA secondary structure predictions, RNAsshapes [34] implements an algorithm to predict secondary structures for every abstract shape, which is realized by using an algebraic dynamic programming (ADP) technique [35]. This reduces the number of predictions (Figure 3). Note that a probabilistic version of RNAsshapes has also been proposed, which computes the accumulated probabilities of all structures that share a shape [36].

**Prediction of representative solutions after clustering**

In this approach, first stochastic sampling (Stochastic sampling with dynamic programming) is performed, then the suboptimal solutions are clustered, and finally a solution for each cluster is predicted. In this way, it is expected that a diverse variety of solutions (i.e. solutions are not similar to each other) will be obtained. For RNA secondary structure predictions, this approach is implemented in Sfold [37] and CentroidFold [38], Sfold provides a web interface for this approach (http://sfold.wadsworth.org/cgi-bin/index.pl). See Figure 4 for an example of the output of the Sfold Web Server for a typical tRNA sequence.

**Visualizing distributions of solutions**

Visualization of the distribution of solutions would include much richer information than a single solution. Yet, visualizing solutions is not trivial because the solution space is generally high-dimensional (compare with the ‘Where Does the Uncertainty Come from?’ section).

**Visualizing distributions of solutions with sampling**

In many cases, a solution lies in a high-dimensional discrete space and the number of possible solutions is immense; therefore, it is difficult to visualize solutions directly. To address this, multidimensional distance scaling (MDS) [39] or principal component analysis is used in combination with the sampling of solutions by stochastic sampling or MCMC.

For RNA secondary structure predictions, Sfold produces this kind of visualization of a distribution of RNA secondary structures in terms of base pair distance (The base pair distance is equal to the number of base pairs that differ between two secondary structures [37]) (Figure 4).

> AL138651.1/64525-64597 [500]
GGGGAUGUAGCUCAUAUGGUAGAGCGCUCGCUUUGCAUGCGAGAGGCACAGGGUUCGAUUCCCUGCAUCUCCA -3030
(((((((((....))))))).).(((((.......)))))) -25.60
(((((((((....))))))).).(((((.......)))))) -25.50
(((((((((....))))))).).(((((.......)))))) -25.50
(((((((((....)))))))..(((((.......)))))) -25.50
(((((((((....)))))))..(((((.......)))))) -25.50
(((((((((....)))))))..(((((.......)))))) -25.50
(((((((((....)))))))..(((((.......)))))) -25.50
(((((((((....)))))))..(((((.......)))))) -25.50

**Figure 2:** Suboptimal secondary structures of a tRNA sequence predicted by RNAsubopt [22], enumerating all possible suboptimal RNA secondary structures within a given energy range from the minimum free energy (MFE).
For phylogenetic tree estimations, Amenta and Klingner [40] and Hillis et al. [41] proposed a method to visualize phylogenetic trees by using MDS based on a distance between two phylogenetic trees called the Robinson–Foulds (RF) distance [42] (the sum of the internal edges in disagreement between the two trees). More recently, Huang and Li [43] have developed some software MASTtreedist that uses MDS with the maximum agreement subtree (MAST) distance, which is the number of leaves in common for the maximum subtree between the two trees [44]. This leads to better clustering results than using the RF distance.
Visualizing distributions with one or two reference solutions

Instead of visualizing the distribution of solutions directly (the ‘Stochastic Sampling with Dynamic Programming’ section), it has been proposed to visualize the distribution with one or two reference solutions. Let us first consider a reference case, in which one reference solution \( s_0 \) and a distance, \( d(\cdot, \cdot) \), between two solutions are given. For each integer value \( k \), \( p(k) \) is equal to the sum of probabilities for solutions whose distance from \( s_0 \) is equal to \( k \):

\[
p(k) = \sum_{s' \in \{d(s_0, s) = k\}} p(s'|D)
\]

where \( p(\cdot|D) \) is a probability distribution over solutions. Then, the distribution \( p(k) (k = 1, 2, 3, \cdots) \) can be easily visualized. One advantage of this visualization is that, for several problems such as alignments and RNA secondary structure predictions, the distribution can be computed exactly by using dynamic programming techniques (Newberg and Lawrence [45]). (The approach described in Stochastic sampling with dynamic programming requires candidate solutions to be visualized. These candidate solutions are usually given by sampling techniques because it is infeasible to visualize the complete distribution). The programs RNABor [46, 47] and RNABorMEA [48] implement this approach. Figure 5 shows an example of RNABor results for a 101 nucleotide SAM riboswitch. It should be noted that points with a large base pair distance from the reference structure may include a lot of structures that might be different from each other.

This approach can be extended to two (or more) solutions. If two reference solutions \( s_0 \) and \( s_1 \) are given, the distribution \( p(k, l) (k, l = 1, 2, \cdots) \) is defined by

\[
p(k, l) = \sum_{s' \in \{d(s_0, s) = k \land d(s_1, s) = l\}} p(s'|D).
\]

In this case, the distribution becomes more specific and includes richer information than that with one reference case. The entire distribution \( p(k, l) \) with two reference solutions can be efficiently computed by using dynamic programming for RNA secondary structures and alignments [50].

Recently, a more efficient method for visualization of both one and two reference cases has been proposed. This method uses a discrete Fourier transform (DFT) and parallel computing (Mori et al., manuscript in preparation).

Visualizing marginal probabilities

Marginal probabilities are the sums of probabilities of solutions that satisfy a specific condition and are formally represented by

\[
p_C = \sum_{0 \in \mathcal{C}} p(0|x)
\]

where \( \mathcal{C} \) is a subset of the predictive space \( X \) satisfying a condition of interest. Marginal probabilities with respect to an appropriate condition are often much larger (typically, near to 1) than the probability of a solution itself (which is, typically, <10^{-5}; Figure 1), and visualization of marginalized probabilities includes rich information, as shown in the following examples.

Example 1 [Base pairing probabilities (BPPs)]

A base pairing probability (BPP), \( p_{ij} \), of an RNA sequence \( x \) is the marginal probability that \( x_i \) and \( x_j \) (the \( i \)-th and \( j \)-th nucleotides of \( x \)) form a base-pair:

\[
p_{ij} = \sum_{0 \in \mathcal{C}(i,j)} p(0|x)
\]

where \( \mathcal{C}(i,j) \) is a set of secondary structures whose \( i \)-th position forms a base pair with the \( j \)-th position, and \( p(0|x) \) is a probability distribution of secondary structures of \( x \), for example, given by the McCaskill model [9] [compare Equation (1)].

The set of all base pairing probabilities for all pairs of nucleotides in a given RNA sequence is called a ‘base pairing probability matrix’ (BPPM) and is represented as a triangular matrix \( P = [p_{ij}]_{1 \leq i < j \leq |x|} \) where \( p_{ij} \in [0, 1] \) is defined in Equation (4). The BPPM of typical models for RNA secondary structures can be computed in a polynomial order time by using a dynamic programming algorithm (e.g. see [9]).

In Figure 6, the base pairing probability matrix for an RNA aptamer is shown. Notice that although the base pairing probability matrix does not specify a secondary structure, it includes richer information about structures than can be obtained from a single RNA secondary structure. See the ‘Analyses Using Marginal Probabilities’ section for the details.

Recently, other visualizations of base pairing probabilities have been proposed (RNABow [52]; http://rna.williams.edu/rnabows/single.html). See Supplementary Figure S2 for an example of RNABow output.
Example 2 [Aligned pairing probabilities (APPs)]

An aligned pairing probability (APP), \( p_{ik} \), of two sequences \( x \) and \( x' \) gives the probability that \( x_i \) and \( x'_k \) align:

\[
p_{ik} = \sum_{y \in C(i, k)} p(y|x, x')
\]

where \( C(i, k) \) is the set of pairwise alignments whose \( i \)th base (nucleotide or amino acid) in \( x \) aligns with the \( k \)th base in \( x' \), and \( p(y|x, x') \) is a probability distribution for pairwise alignments between \( x \) and \( x' \), such as the Miyazawa model [53] or pair hidden Markov models (pHMMs) [1].

The entire set of probabilities for every possible pair of base alignments between two sequences \( x \) and \( x' \) is called an ‘aligned pairing probability matrix’ (APPM). It is represented by a matrix \( P = \{p_{ik}\}_{1 \leq i \leq |x|, 1 \leq k \leq |x'|} \) where \( p_{ik} \in [0, 1] \) is defined in Equation (5). The APPM of typical models for pairwise alignments can be computed in polynomial order time by using a dynamic programming algorithm (see, e.g. Durbin et al. [1]). Like a BPPM for an RNA sequence, an APPM can be easily visualized.

Example 3 (Leaf splitting probabilities)

A leaf splitting probability, \( p_{X, Y} \), for a leaf set \( S \) gives the marginal probability that there exists a partition \( (X, Y) \) of \( S \) (i.e., \( X \cup Y = S \) and \( X \cap Y = \emptyset \)) formed by cutting an internal edge in a phylogenetic tree:

\[
p_{X, Y} = \sum_{\theta \in C(X, Y)} p(\theta|S)
\]

where \( C(X, Y) \) is the set of phylogenetic tree topologies that can be split into \( X \) and \( Y \) by cutting one edge in \( T \), and \( p(\theta|S) \) is a probability distribution of phylogenetic tree topologies, such as the one given in [30].

Unlike base paring probabilities or aligned pairing probabilities, it is difficult to visualize the complete set of splitting probabilities because it cannot be
represented as a ‘matrix’. The following is a non-trivial visualization of phylogenetic trees, with leaf splitting probabilities.

**Example 4 [Centroid Wheel Tree (CWT) [54]]**

For phylogenetic tree estimations, a Centroid Wheel Tree (CWT) [54] provides a novel visualization of a phylogenetic tree topology with marginal probabilities (leaf splitting probabilities; Example 3). See Figure 7 for an example. See [54] for the detailed definition and algorithms for computing a CWT. A CWT can be computed through a website (http://cwt.cb.k.u-tokyo.ac.jp/).

**WHAT SHOULD WE DO WHEN ONE PREDICTION IS REQUIRED?**

Although the approaches described in the previous section are good ways to handle uncertainty, it is often necessary to predict only one solution.

**Maximum likelihood estimator and its drawback**

A straightforward approach to make a prediction for a given probability distribution, \( p(y|D) \), on the predictive space \( Y \) is to find the solution \( \hat{y} \) with the maximum probability:

\[
\hat{y} = \arg \max_{y \in Y} p(y|D). \tag{7}
\]

This approach is known as maximum likelihood (ML) estimation. Many existing tools or algorithms use ML estimators; however, recent studies have indicated that ML estimators are not always superior estimators, e.g. [55, 56]. The following is a typical example in which the ML estimator is inappropriate.

**Example 5 (Carvalho and Lawrence [55])**

The predictive space is the \( n \)-dimensional binary space \( Y = \{0, 1\}^n \), in which similar vectors in \( Y \) represent similar predictions.
We introduce a probability distribution on $Y = \{0, 1\}^n$ as follows:

$$p(\theta | D) = \begin{cases} p_1 := \frac{1}{n+3} & \text{if } \theta \in \mathcal{S} := \{(x_1, \ldots, x_n) | x_k \in \{0, 1\}, \sum_{k=1}^n x_k \leq 1\} \\
p_2 := \frac{2}{n+3} & \text{if } \theta = \theta^1 := (1, 1, \ldots, 1) \\
0 & \text{otherwise} \end{cases}$$

(clearly, $\sum_{\theta \in Y} p(\theta | D) = 1$). Then, the maximum likelihood estimator gives the solution $\theta^1 = (1, 1, \ldots, 1)$. However, it can be seen that

$$\sum_{\theta : H(\theta', 0) \leq 1} p(\theta | D) = \frac{n+1}{n+3}$$

where $\theta^0$ is the zero vector in $\{0, 1\}^n$ (i.e. all elements in the vector are equal to 0) and $H(\cdot, \cdot)$ is the Hamming distance. This means that when $n = 997$, 

---

**Figure 7:** An example of a Centroid Wheel Tree (CWT) computed using the website: http://cwt.cb.k.u-tokyo.ac.jp/. In this representation, leaf splitting probabilities are shown next to each edge.
the sum of the probabilities solutions around (similar to) $\theta^0$, for which the Hamming distance from $\theta^0$ is $\leq 1$, is 0.998, whereas the probability of the ML estimation (i.e. $\hat{\theta}$) is only 0.002 (Figure 8), indicating that $\theta^0$ is a much better estimation than the ML estimation $\hat{\theta}$. Note that the ML estimation $\hat{\theta}$ is a long way from the solution $\theta^0$ because $H(\theta^0, \hat{\theta}) = n$.

It is worth noting that several bioinformatics problems (including RNA secondary structure prediction, RNA–RNA interaction prediction, sequence alignment and phylogenetic tree estimation) can be formulated as a point estimation problem in (a subset of) binary spaces $Y = \{0, 1\}^n$ with large $n$ [56]. Hence, a similar situation to the one in the above (artificial) example could occur in real problems in bioinformatics (for example, in Figure 4, the clover leaf structure of tRNA does not lie in the first rank cluster).

**Maximum expected gain/accuracy (MEG/MEA) estimators**

Because ML estimators are not always good estimators, as shown in Example 5, we will introduce alternatives to ML estimators in this section, on the basis of given accuracy/evaluation measures for the prediction problem.

**Definition**

Given a probability distribution $p(\theta|D)$ on a predictive space $Y$, the maximum expected gain (MEG) estimator is defined by

$$\hat{y} = \arg \max_{y \in Y} \sum_{\theta \in Y} G(\theta, y)p(\theta|D)$$

(8)

where $G(\theta, y)$ is called the gain function, which returns the similarity between two solutions in $Y$. When the gain function is designed according to an accuracy or evaluation measure for the target problem, in which $y$ and $\theta$ are considered as a prediction and reference, respectively, the estimator is often called a maximum expected accuracy (MEA) estimator [57–59] [When this estimator is called a ‘maximum expected accuracy’ (MEA) estimator, $G(\theta, y)$ in Equation (8) is equal to an accuracy measure (or is designed according to an accuracy measure) for a reference $\theta$ and a prediction $y$. This also implies that $p(\theta|D)$ is considered to be a probability distribution of references, which is misleading because $p(\theta|D)$ does not usually represent such a distribution. In RNA secondary structure prediction, for example, the McCaskill model provides not a probability distribution of reference secondary structures but rather a full ensemble of possible secondary structures [9]]. MEA estimators predict the solution by maximizing the expected accuracy when the solutions are distributed according to $p(\theta|D)$. These estimates are, therefore, appropriate for the given accuracy measure [MEA estimators are closely related to posterior decoding algorithms (compare Algorithms using marginal probabilities: posterior decoding).

**ML estimators from the viewpoint of MEA**

Let us consider the ML estimator from the viewpoint of MEA. The ML estimator of Equation (7) can be represented as

$$\hat{y} = \arg \max_{y \in Y} p(y|D)$$

$$= \arg \max_{y \in Y} \sum_{\theta \in Y} \delta(\theta, y)p(\theta|D)$$

where $\delta(\theta, y)$ is the Kronecker delta function that returns 1 only if $\theta = y$. This means the ML estimator is a kind of MEG/MEA estimator in which the gain (accuracy) function is the delta function. However, the delta function is rarely used as an accuracy measure for most problems. In RNA secondary structure predictions, for instance, using the delta function as an accuracy measure would mean that a predicted secondary structure must be exactly same as the reference structure; this seems too strict. Instead, the evaluation is usually based on a comparison of base pairs in predicted and reference secondary structures (Supplementary Section SB).

Note that the delta function $\delta(\theta, y)$ is much stricter than the accuracy measures in Supplementary Section SB. The MFE structure is equivalent to the ML estimator with the probability distribution of Equation (1); therefore, the MFE structure is not optimal to the accuracy measures aforementioned. How do we design an MEA estimator for which it is appropriate to use those accuracy measures?

**Example: RNA secondary structure prediction (I)**

In RNA secondary structure predictions, more true predictions (TP/TN) and fewer false predictions (FP/FN) of base pairs should occur in a predicted secondary structure (compare Supplementary Section SB). Based on the principle of the MEA estimator, the following MEG estimators (called
generalized centroid estimators) have been introduced [58]:

$$\hat{y} = \arg \max_{y \in S(x)} \sum_{\theta \in S(x)} G(\theta, y)p(\theta|x)$$

where

$$G(\theta, y) = \alpha_1 \cdot TP + \alpha_2 \cdot TN - \alpha_3 \cdot FP - \alpha_4 \cdot FN$$

with the $\alpha_k$ representing arbitrary positive constants, and $p(\theta|x)$ is a probability distribution on a set $S(x)$ of possible secondary structures of RNA sequence $x$; TP, TN, FP and FN are defined in Supplementary Section SB. Hamada et al. [58] showed that the number of parameters can be reduced to only one because the MEA estimator with this gain function is equivalent to the MEG estimator with the gain function

$$G_c(\theta) = \gamma \cdot TP + TN,$$     \hspace{1cm} (9)

where $\gamma > 0$. When $\gamma = 1$, the estimator is equivalent to the centroid estimator [55]; hence, the estimator is called the generalized centroid estimator or $\gamma$-centroid estimator. By using the parameter $\gamma$, users can adjust the balance between sensitivity and PPV (compare Supplementary Equations S1 and S2).

CONTRAfold [57] uses a different MEA estimator, but Hamada et al. [58] theoretically showed that the gain function used in CONTRAfold includes a term that is biased with respect to the accuracy measure of RNA secondary structure prediction. Computational experiments reported in [58] support this theoretical result because they obtained consistent results for three different probabilistic models for RNA secondary structures. This example indicates the importance of designing the gain function appropriately.

**Example: RNA secondary structure prediction (II)**

In the previous section, a linear combination of TP, TN, FP and FN is used as a gain function. Here, we try to optimize the accuracy given by the measure MCC (Supplementary Equation S3 in Supplementary Section SB) directly in the MEA estimators. Thus, the MEA estimator

$$\hat{y} = \arg \max_{y \in S(x)} \sum_{\theta \in S(x)} \text{MCC}(\theta, y)p(\theta|x)$$ \hspace{1cm} (10)

is introduced. Unfortunately, it is difficult to compute this estimator efficiently because neither the expectation of MCC nor the arg max operation in Equation (10) can be computed efficiently. In [60], the authors addressed this problem by proposing approximate estimators (called maximum pseudo-expected accuracy estimators) as follows:

$$\hat{y} = \arg \max_{y \in S(x)} \sum_{\theta \in S(x)} \text{MCC}^0(y),$$    \hspace{1cm} (11)

where

$$\text{MCC}^0(y) = \frac{\hat{\text{TP}} \cdot \hat{\text{TN}} - \hat{\text{FP}} \cdot \hat{\text{FN}}}{\sqrt{(\hat{\text{TP}} + \hat{\text{FP}})(\hat{\text{TP}} + \hat{\text{FN}})(\hat{\text{TN}} + \hat{\text{FP}})(\hat{\text{TN}} + \hat{\text{FN}})}}$$

and $\hat{\text{TP}}$, $\hat{\text{TN}}$, $\hat{\text{FP}}$ and $\hat{\text{FN}}$ are, respectively, the expected values of TP, TN, FP and FN [under a given probability distribution $p(\theta|x)$]. In contrast to the expected value of MCC, $\text{MCC}^0(y)$ can be
computed efficiently from the base-pairing probability matrix. Stochastic sampling is used to approximate the ‘argmax’ operation in Equation (11).

It should be emphasized that the aforementioned estimators can be adapted to use more general accuracy measures. See Supplementary Section SC for details.

**Example: sequence alignments**
The MEA estimators introduced in the previous sections (for RNA secondary structure predictions) can be directly applied to pairwise sequence alignments, where TP, TN, FP and FN are computed with respect to bases (nucleotides or amino acids) that are aligned (and un-aligned) between two sequences; therefore, the γ-centroid estimator is expected to produce more correctly aligned bases in the predicted alignment. Actually, computational experiments in Frith et al. [61] indicated that compared with the conventional alignment method (i.e. maximizing an alignment score that corresponds to the ML estimation of pairwise alignment), the γ-centroid alignments substantially reduce the number falsely aligned bases (i.e. FP) in return for the sacrifice of a slight reduction in the number of correctly aligned bases (i.e. TP).

**Example: gene predictions**
In gene prediction [2, 62], the accurate prediction of a boundary (e.g. an exon–intron boundary) in the predicted gene is important. Sensitivity and specificity have been used for evaluation at the gene and exon levels [2, 63]. To design a predictor that is suited to these accuracy measures, the number of correctly predicted boundaries and un-boundaries can be used in the gain function, and the corresponding MEA estimator is called a maximum expected boundary accuracy estimator. Like MEA estimators in RNA secondary structure prediction, the prediction can be efficiently conducted by using a dynamic programming approach. See Gross et al. [63] for the details. In their computational experiments, the MEA estimator based on boundary accuracy measures was superior to existing gene prediction methods.

**Other examples**
In addition to the aforementioned examples, many algorithms in bioinformatics can be classified, from the viewpoint of MEA/MEG, with respect to gain function and predictive space. See [59] for a review of MEA estimators.

**Point estimation with additional information**
Although well-designed MEG/MEA estimators would provide a better single solution that is more appropriate to a given evaluation measure than ML estimators, the probability of the prediction is still low. In fact, it is lower than the probability of the ML estimation. In real applications, it can be useful to provide a single solution along with additional information as we now show (Multiple pieces of information can be simultaneously used as additional information for point estimation).

**Point estimation with marginal probabilities**
Marginal probabilities (compare with the ‘Visualizing Marginal Probabilities’ section) can be often used as additional information for a point estimation as demonstrated by the following examples.

**Example 6 [RNA secondary structure with base pairing probabilities (BPPs)]**
Base pairing probabilities (Example 1) can be used as additional information for a point estimation of RNA secondary structure as shown in Figure 9. This kind of prediction is available by using the CentroidFold web server [38] and the RNAfold web server [64].

**Example 7 [Pairwise alignment with aligned-pairing probabilities (APPs)]**
Aligned pairing probabilities (compare Example 2) can be used as additional information for the point estimation of pairwise alignments (Figure 10). The LAST web server (http://last.cbrc.jp/) returns a pairwise alignment with aligned probabilities. Also, FSA (Fast Statistical Alignment) [65] computes marginal probabilities including gap probabilities (Gap probabilities are the marginal probabilities that a specific position corresponds to a gap, which can be computed by a dynamic programming algorithm).

Kim et al. [66] have extended aligned pairing probabilities to the case of multiple alignments: probabilistic sampling-based alignment reliability (PSAR) calculates the reliability of each column in a given multiple alignment.

**Point estimation with credibility limits**
The credibility limit [67] of α (0 ≤ α ≤ 1) provides a global measure of the reliability of a point estimation, which is computed in the following way.
A probability distribution \( p(y|D) \) for a predictive space \( Y \) is given [e.g. a probability distribution \( p(\theta|x) \) for secondary structures of RNA sequence \( x \) is given].

A distance between two solutions is defined (for example, the base pair distance can be used for RNA secondary structure predictions).

Fix an arbitrary solution \( y \in Y \) that you are interested in (for example, fix the MFE structure of the RNA sequence \( x \)).

Compute the minimum distance \( d \) for which \( 100\% \) of the accumulated probabilities of all solutions are included in the set of solutions whose distance from \( y \) is less than \( d \). This distance is called the credibility limit of \( \alpha \).

The authors [67] have constructed an algorithm to compute credibility limits, based on stochastic sampling of RNA secondary structures or alignments (the ‘Prediction of Suboptimal Solutions’ section). Note, however, that credibility limits for RNA secondary structures and alignments can be computed by using dynamic programming without using sampling techniques [45] because the complete distribution from a fixed solution can be computed by using dynamic programming (see the ‘Visualizing Distributions with One or Two Reference Solutions’ section).

**HANDLING UNCERTAINTY TO DEVELOP COMPLEX PIPELINES**

**General strategies**

When developing complex algorithms or pipelines in which two or more prediction methods are involved, the uncertainty described in this review should be handled carefully, especially for the intermediate prediction method(s) in the pipeline. For example, let us consider the following pipeline, which is based on two prediction methods (with the example of phylogenetic tree estimation from unaligned sequences in parentheses; see Figure 11):

1. Obtain data \( D \) (e.g. \( D \) is a set of unaligned sequences \( S \)).
2. Predict an intermediate solution \( \theta' \in Y' \) from \( D \) [e.g. predict a multiple sequence alignment (MSA) \( A \) of \( S \)].
3. Predict the final solution \( \theta \in Y \), based on \( \theta' \) (e.g. predict a phylogenetic tree \( T \) based on the multiple alignment \( A \)).

Given the uncertainty of the intermediate estimation problem (in Step 2 aforementioned), predicting a single solution \( \theta' \) should be avoided because point estimation is unreliable. Actually, as discussed in the ‘Introduction’ section, the prediction of phylogenetic trees is greatly affected by the uncertainty of multiple alignments [15]. To handle the uncertainty, it is ideal to consider a joint distribution \( p(\theta', \theta'|D) \) on \( Y \times Y' \). For instance, \( p(\theta', \theta'|D) \) could be a joint distribution for a multiple sequence alignment (MSA) and a phylogenetic tree. Then the distribution is marginalized onto \( Y \):

\[
p(\theta|D) = \sum_{\theta' \in Y'} p(\theta', \theta'|D).
\]

**Figure 9:** Secondary structure of a tRNA sequence with base pairing probabilities produced by the CentroidFold web server (http://www.ncrna.org/centroidfold) [38]. Base pairs with warmer colors have higher base pairing probabilities.

**Figure 10:** A pairwise alignment with aligned-pairing probabilities produced by the LAST web server (http://last.cbrc.jp) [61]. Aligned pairs with colder colors have higher aligned pairing probabilities.
This marginal distribution includes the uncertainty of the intermediate solution $y'$. From this marginal distribution, we can predict either one final solution, by using, for example, ML/MEA/MEG estimators (‘What Should We do When One Prediction is Required?’ section), or several solutions (see the ‘Avoiding Point Estimations’ section). However, considering joint distributions [such as Equation (12)] often leads to huge computational cost, and an approximation could be useful to obtain efficient estimators as shown in the following examples.

Example: prediction of RNA secondary structure by using homologous sequence information

In this section, we consider the problem of predicting an RNA secondary structure by combining homologous sequence information, where the input is the target RNA sequence $t$ and a set of its homologous sequences $H$ (see Figure 12). In this problem, we assume that the target sequence $t$ and each sequence in $H$ share a consensus structure; therefore, the information from the homologous sequences would improve the accuracy of the RNA secondary structure prediction of the target sequence.

One approach to this problem is the following (Figure 12) (Covariance models [69] can also be used to solve the problem).

1. A multiple sequence alignment for input RNA sequences (both target and homologous sequences) is computed by using a multiple aligner for RNA sequences (such as CentroidAlign [70], Locarna [71] and MAFFT [68]).
2. A common/consensus secondary structure for the multiple alignment is computed by using, e.g. CentroidAlifold [72] and RNAalifold [73].
3. The secondary structure of the target RNA sequence is computed from the predicted common secondary structure by mapping the common structure to the target RNA sequence.

Note that intermediate point estimations are used twice (in Steps 1 and 2) in the aforementioned procedure. The results will be affected by the uncertainty of these point estimations, which should, therefore, be avoided (‘General Strategies’ section).

There is a probabilistic model for structural alignments of RNA sequences (known as the Sankoff model [74]), in which both consensus secondary structure and conventional alignments are considered simultaneously (Strictly speaking, Sankoff [74] did not use a probabilistic model for structural alignments. However, one can be easily introduced by using a similar approach to that of McCaskill [9]). As an alternative to point estimation, it is better to obtain a distribution of RNA secondary structures of the target sequence $t$ by marginalizing (and averaging) the distribution of structural alignments, and then predict a secondary structure for the target by using MEG/MEA estimators or the ML estimator with respect to the marginal probability distribution. However, this approach entails huge computational cost for prediction ($O(L^6)$ where $L$ is the length of RNA sequences). Hamada et al. [75] approximate this approach by factorizing the distribution of structural alignments into a distribution of secondary structures and a distribution of usual alignments (This algorithm is implemented in the Centroid-Homfold software [75]). Furthermore, they showed that the approach in Figure 12 (in which the
uncertainty of secondary structures and alignments is not considered) was consistently worse than their method in which the uncertainty of alignments and secondary structures is considered. See [75] for the details.

A similar type of technique was used for sequence alignments of RNA sequences [70]. Moreover, a more general form of estimator was introduced in [56].

Example: variant detection with NGS technologies

In the following pipeline, I try to detect variants of genomes, using next-generation sequencing (NGS) technologies.

1. Determine bases (by ‘base-calling’ [76]) from intensity data produced by sequencers (e.g. Illumina, 454, SOLiD and PacBio RS) and obtain a set of reads (i.e. fragmented short sequences).
2. Map (align) the reads to reference genomes by using, e.g. BWA [77], bowtie [78] or LAST [79].
3. Predict single-nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs) from the mapped reads obtained in Step 2, using tools such as SAMtools [80] or VarScan [81].

In this pipeline, several point estimations are used: Step 1 includes the uncertainty of called bases. By retaining the information about the uncertainty of base-calling, most current sequencers produce reads with quality scores in the FASTQ format [82]. Step 2 includes uncertainty with respect to the locations of mapped reads and the detailed alignment between reads and reference genomes. To the best of my knowledge, there is no method that handles these uncertainties completely, although the approach of probabilistic alignments with quality scores [83] partially handles both uncertainties (the uncertainty arising from locations of mapped reads cannot be resolved completely even if probabilistic alignments with quality scores are used in the pipeline).

DISCUSSION

Usefulness of marginal probabilities

Analyses using marginal probabilities

Because of the uncertainty of single solutions that has been described in this article, considering an ensemble of solutions is useful in the analysis of biological data in bioinformatics. In particular, marginal probabilities (compare the ‘Visualizing Marginal Probabilities’ section) are often used in the analyses because they take into account information about the entire ensemble of solutions.

- In Adachi et al. [52], the analysis of the base paring probability matrix of an RNA aptamer (bound to a cytokine) suggested that a stem involved in the aptamer was unstable, and there was a possibility that it could form several different structures. This was confirmed by biochemical experiments (Figure 13). It should be emphasized that we could not have obtained this conclusion if only a single predicted structure (e.g. MFE structure) had been considered.
In Halvorsen et al. [84], the authors analyzed the conformation change of RNA secondary structures, caused by an SNP, by using the base pairing probabilities. They reported seven disease states and phenotypes in which two or more associated SNPs were found to alter the structural ensemble of the RNA [see Table 1 in [84] for the details].

Iwakiri et al. [Iwakiri et al., submitted for publication] systematically analyzed base pairing probabilities for paired and unpaired nucleotides of RNA secondary structures involved in known protein–RNA complexes taken from the Protein Data Bank (PDB) [85]. Their analyses lead to novel findings that could not be found if only the snapshots of RNA secondary structures in the PDB are considered.

**Algorithms using marginal probabilities: posterior decoding**

It should be also remarked that marginal probabilities are used to construct algorithms, and such an approach is called posterior decoding. Posterior decoding algorithms are widely used in bioinformatics: see [13, 23, 53, 86–92], for example (From a historical viewpoint, a posterior decoding was originally proposed in Miyazawa [53] for alignments, later adopted for HMMs by Holmes and Durbin [86], and first introduced into RNA secondary structure predictions by Knudsen and Hein [87]). Actually, most MEG/MEA estimators ['Maximum Expected Gain/Accuracy (MEG/MEA) Estimators’ section] lead to posterior decoding algorithms because the final algorithm is based on only marginal probabilities. For example, the final algorithm in the \(\gamma\)-centroid estimators for RNA secondary structure predictions [compare the ‘Example: RNA Secondary Structure Prediction (I)’ section] is based on the base pairing probability matrix (BPPM) of the RNA sequence (see [58] for the detailed algorithm).

**Uncertainty in hypothesis testing**

In several problems in bioinformatics such as homology search [93], a hypothesis testing approach is
frequently used, in which the log odds score of two probabilities relative to a null model is computed instead of the maximum score of the solution. For instance, in homology search, the following (Viterbi) score is used:

\[
V = \log \frac{\max_x p(x, z|H)}{p(x|R)}
\]

(13)

where \(x\) is a target sequence (we would like to judge whether this sequence is homologous to a query sequence or not), \(R\) is a random model (e.g. a one-state HMM) and \(z\) is an alignment between the target and query sequences. The score might be affected by uncertainty because it uses only the maximum score of the optimal alignment. To address this issue, Eddy [94] proposed the use of forward scores,

\[
F = \log \frac{\sum p(x, z|H)}{p(x|R)}
\]

(14)

in homology searches of biological sequences. In the forward score, the sum of the probabilities of all solutions (alignments) is used, which mitigates the influence of the uncertainty [94].

Uncertainty introduced by internal parameters in models

In the aforementioned section, I have focused on the uncertainty of solutions where the probabilistic model (i.e. the probability distribution on the predictive (solution) space) is fixed. On the other hand, a kind of uncertainty also appears with respect to changes of internal parameters in the probabilistic model (e.g. a predicted sequence alignment might be substantially changed by slight changes of the substitution matrix or gap costs in alignments). Several interesting studies have been conducted to address this issue.

In [17, 95, 96], parametric approaches are used to enumerate the optimal solutions for all possible changes of internal parameters, which enables the analysis of the parametric behavior of maximum a posteriori (MAP) inference computations. For example, a parametric alignment procedure [96, 97] can efficiently find all the optimal alignments for all possible parameters in a pair hidden Markov model (pHMM), which might be one remedy for the uncertainty of solutions introduced by parameter changes. It should be emphasized that the parametric methods can be explained from a unified viewpoint by using a general mathematical theory called tropical geometry [98].

How to provide probability distributions for solutions

In this review, I have assumed that there exists a probability distribution on a predictive space in the estimation problem (Purpose of this Review’ section). However, the existence of the probability distribution is not trivial, and research on the probability distribution (probabilistic model) for each problem is also important, and much research has been conducted in this area. For example, probabilistic models have been developed for RNA secondary structures [9, 57, 99–101], alignments [1, 53, 102], gene prediction [63, 103], phylogenetic trees [30] and RNA common secondary structures [72, 73].

Future directions

By using the approaches described in this review, issues arising from the uncertainty of solutions can be addressed, in part, in several areas of bioinformatics, including RNA secondary structure predictions and sequence alignments. However, it is obvious that uncertainty will raise further serious issues in other fields of bioinformatics. Therefore, in the future, further general methods for fighting against uncertainty, which are applicable to various problems, should be developed: for example, more efficient methods for visualizing or sampling solutions in high-dimensional spaces, methods that mitigate the uncertainty of point estimations, and so forth.

CONCLUSION

In this review, I focused on the uncertainty of solutions (i.e. the fact that the probability of any solution is low), which often give rise to serious issues when developing algorithms and analyzing biological data, and introduced several approaches to handle this problem appropriately. I have presented many actual examples in this review, and although most of the examples are related to sequence analyses, such as RNA secondary structure predictions, sequence alignments and phylogenetic tree estimations, I believe that the basic concepts and ideas described in this review will be useful for other problems in many areas of bioinformatics.
SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points
- Many bioinformatics problems, such as sequence alignment, gene prediction, phylogenetic tree estimation and RNA secondary structure prediction, are often affected by the uncertainty of a solution, that is, the probability of the solution is extremely small (e.g., $10^{-10}$).
- In the analysis of biological data or the development of prediction algorithms, this uncertainty should be handled carefully and appropriately.
- This review provides several approaches to combat this uncertainty, presenting examples in bioinformatics.
- In particular, point estimations should be handled carefully when designing a pipeline.
- The basic concepts and ideas described in this review will be useful in various areas of bioinformatics.

Acknowledgements

The author is grateful to Dr Martin Frith, Dr Dave duVerle and Prof. Kiyoshi Asai for valuable comments. Thanks are also due to three anonymous reviewers for useful suggestions, which helped to improve the article.

FUNDING

MEXT KAKENHI (Grant-in-Aid for Young Scientists (A): 24680031) (in part).

References


