Evaluating Biosignatures for Life Detection

Andrew Pohorille and Joanna Sokolowska

Supplementary Information

A. Derivation of informedness measure from the expected utility approach

In formulations of SDT based on expected utility theories, all specific outcomes are associated with utilities. Then, the total expected utility is calculated as a sum of individual utilities weighed by the probability of each outcome or the corresponding decision weight.

Assume that the four possible outcomes in SDT (true positive, false negative, false positive and true negatives) have utilities $u_1...u_4$, respectively. These utilities are the same for all biosignatures. In other words, utility depends on the outcome but not on a biosignature. Then, the total expected utility of the biosignature, $U$, is equal to

$$U = u_1 P(B \mid L) - u_2 P(\sim B \mid L) - u_3 P(B \mid \sim L) + u_4 P(\sim B \mid \sim L)$$  \hfill (S1)

where $P(B \mid L)$, $P(\sim B \mid L)$, $P(B \mid \sim L)$ and $P(\sim B \mid \sim L)$ are, respectively, probabilities of true positives, false negatives, false positives and true negatives. Here, we consider $u_i$ as nonnegative. Minus signs in front of the terms representing false negatives and false positives mean that these terms always reduce overall expected utility.

The four probabilities in the equation above are subjects to the constraints:

$$P(B \mid L) + P(\sim B \mid L) = 1$$  \hfill (S2)

and

$$P(B \mid \sim L) + P(\sim B \mid \sim L) = 1$$  \hfill (S3)

This allows for eliminating $P(B \mid L)$ and $P(\sim B \mid \sim L)$ from Eq. 1, yielding

$$U = u_1 + u_4 - (u_1 + u_2) P(\sim B \mid L) - (u_3 + u_4) P(B \mid \sim L)$$  \hfill (S4)

This equation can be rewritten as:

$$U = (u_1 + u_2) \left[ 1 - \frac{u_1 + u_2}{u_1 + u_4} P(\sim B \mid L) - \frac{u_3 + u_4}{u_1 + u_4} P(B \mid \sim L) \right]$$  \hfill (S5)

Define
\[ \alpha = \frac{u_1 + u_2}{u_1 + u_3} \tag{S6} \]

and

\[ q = \frac{u_1 + u_4}{u_1 + u_2 + u_3 + u_4} \tag{S7} \]

Then, Eq. 5 takes the form:

\[ \bar{U} = (u_1 + u_4) \{ 1 - \left[ \alpha P(\sim B|L) + \left( \frac{1}{q} - \alpha \right) P(B|\sim L) \right] \} \tag{S8} \]

This is nearly identical to Eq. (6) for \( J \) in the paper:

\[ J = 1 - \left[ \alpha P(\sim B|L) + (1 - \alpha) P(B|\sim L) \right] \tag{S9} \]

The term \( u_1 + u_4 \) in front of the r.h.s. of Eq. 8 is irrelevant, as it scales expected utilities of all biosignatures by the same value, compared to \( J \). As in \( J \), \( \alpha \) plays the role of mission design parameter, which corresponds to what is often called decision criterion parameter in other decision problems. It weighs relative importance of avoiding false positives and false negatives. The only difference between Eq. 8 and Eq. 9 is in precise weighing of false positives and false negatives, as the weighing term in front of false positives is \( 1 - \alpha \) in \( J \) and \( 1/q - \alpha \) in \( \bar{U} \), where \( q \) is the ratio of utilities of correctly identified outcomes to the sum of all utilities (note the sign convention for utilities specified above).
B. Example of evaluating biosignatures

Background on selected biosignatures. Amino acids exist in two enantiomeric forms L ("left-handed") and D ("right-handed"). In all terrestrial biology, L-amino acids are utilized to build the basic polymers of life – proteins. Studies in synthetic biology indicate that strong enantiomeric preference might not be just a characteristic of terrestrial life but a requirement for "universal biology", although it is not clear whether preference for the L-form has to be universal. This leads to a proposal that the existence of a sufficiently large enantiomeric excess might be a good biosignature in any environment beyond Earth. For an excellent, recent review on this topic see Glavin et al., 2019.

Biosignatures. In this example, we consider two recently proposed, closely related biosignatures, based on enantiomeric excess of amino acids.

1. Biosignature B1 - sufficient (e.g. >20%) enantiomeric excess in amino acid composition (Neveu et al., 2018).
2. Biosignature B2 - large, enantiomeric excess in amino acid composition, simple distribution of amino acids and light stable C-, N- and H-isotopic compositions (Glavin et al., 2019).

Knowledge base. A knowledge base given in the table below was constructed for the example biosignatures. This knowledge base is organized according to the structure described in Section 4 of the main text.

| ARGUMENTS PRO | ARGUMENTS CON |
|---------------|---------------|
| Presence in biological systems | | |
| Homochirality is a fundamental property of life | |
| Enantiomerically mixed oligomers form only a very limited number of structures; this might be insufficient to support life | There are examples of enantiomerically mixed functional oligomers |
| Polymers with mixed chirality do not appear to have structure and therefore are non-functional | |
| | Polymers made of analogs of alpha-amino acids are possible and can be functional |
| Survival (biological) | | |
| Significant excess of one enantiomeric form of amino acids might indicate their biological origin | Mechanisms of abiotic synthesis of amino acids in space is poorly understood. Synthesis with enantiomeric excess is possible |
Continuous deposition of biogenic amino acids in the presence of extant life diminishes false negatives.

Racemization of biogenic amino acids is relatively fast on geological time scale and might lead to false negatives.

Racemization of some non-proteinogenic amino acids that are present in biology, e.g. isovaline, is slow and therefore it should be possible to establish that they are of biological origin.

Enantiomeric excesses of biological amino acids should be related to their racemization rates.

Distribution of amino acids could be a good indicator of biology. This distribution should contain 10-40 amino acids of different size and polarity.

### Presence in abiotic systems

| | |
|---|---|
| Urey-Miller experiments are known to produce racemic mixtures of chiral molecules. | |
| Strecker synthesis, a candidate mechanism for abiotic synthesis of amino acids, yields racemic mixtures. | |
| Enantiomeric excess arising from abiotic sources might not be a common phenomenon. | |
| Wide diversity of amino acids (>100) in meteorites suggest that these amino acids are not of biological origin. | |
| No abiotic pathways for synthesis of some biological amino acids considered essential, e.g. histidine and tryptophan, are known. | |

### Survival (abiotic)

| | |
|---|---|
| Asymmetric autocatalysis and amplification during crystallization can increase enantiomeric ratio in abiotic samples of amino acids. | Proteinogenic amino acids from extraterrestrial sources are usually found to be racemic. |
| If isotopic composition of L and D isomers or potentially biological and clearly abiotic amino acids is the same, then the origin was most likely abiotic. | |

**Evaluation of probabilities.** A small group of astrobiologists was asked to evaluate four probabilities for both biosignatures, B1 and B2, on the basis of information provided in the knowledge base. Two probabilities were related to a possible biological source of the biosignature. They were: the probability that the biosignature is present in biological systems in
the target environment, \( P(P|L) \), and the probability that the biosignature from biological sources survives in this environment, \( P(S|L) \). The remaining two probabilities regarded a possible abiotic source of the biosignature: the probability that the biosignature from abiotic sources is present, \( P(P|\sim L) \), and the probability that this biosignature survives, \( P(S|\sim L) \). A biosignature feature generated biologically exists if it is present in biological samples and survives in the environment. This means that

\[
P(B|L) = P(P, S|L).
\]

Applying the chain rule of probability calculus this equation can be rewritten as

\[
P(B|L) = P(S|L, P)P(P|L).
\]

Accepting a reasonable assumption that the presence of a biosignature and its survival are uncorrelated, the last equation simplifies to:

\[
P(B|L) = P(S|L)P(P|L).
\]

Now, the equation for \( P(B|L) \) contains only the quantities evaluated by the participating astrobiologists. A similar equation follows for the probability of the biosignature feature that is of abiotic origin

\[
P(B|\sim L) = P(S|\sim L)P(P|\sim L).
\]

The last two equations provide prescription for calculating true and false positives (signal and noise) from the probabilities estimated by the participants in the pilot study.

The average values of \( P(P|L) \) and \( P(S|L) \) for biosignature \( B_1 \) were 0.85 and 0.65, respectively. The average probabilities related to abiotic sources of \( B_1 \), \( P(P|\sim L) \) and \( P(S|L) \), were 0.25 and 0.35 respectively. This yields \( P(B|L) = 0.552 \) and \( P(B|\sim L) = 0.087 \).

For biosignature \( B_2 \), the average \( P(P|L) \) and \( P(S|L) \) were 0.90 and 0.60 and the average \( P(P|\sim L) \) and \( P(S|L) \) were 0.10 and 0.35. \( P(B|L) \) and \( P(B|\sim L) \) calculated from these values were 0.54 and 0.035, respectively.

In both cases, the signal was evaluated as substantially stronger than noise. The corresponding values of Youden J statistics (see Eq. (4) in the main text) were 0.465 and 0.505 for \( B_1 \) and \( B_2 \). The Bayesian factor \( K(B) \) (see Eq. (7) in the main text) for \( B_1 \) and \( B_2 \), respectively, was 6.3 and 15.4. This means that \( B_2 \) is expected to be a better biosignature than \( B_1 \). This is largely because the probability of false positives in this case is lower, which is expected considering that \( B_2 \) is more stringent than \( B_1 \). In contrast, probabilities \( P(P|L) \) and \( P(B|L) \) associated with biological origins were evaluated as nearly identical for \( B_1 \) and \( B_2 \). This is inconsistent with the fact that \( B_2 \) is a subset of \( B_1 \) and, therefore, its presence should be less probable. This result, however, is
not unexpected, as it most likely has roots in a common cognitive bias, called conjunction fallacy, described in 5.1.2.

As mentioned at the beginning of Section 5, it might be beneficial to judge arguments provided in the knowledge base rather than estimate probabilities. Evaluating arguments requires an additional step of converting these evaluations to probabilities according to predefined, validated rules. Since such validated rules are not currently available, this approach has not been pursued. Also, considering that the sample size in this pilot study was quite small, no uncertainly quantification (see Supporting Material, Section 3) was carried out.

Evaluation of utilities. The participants were asked to provide judgements needed for evaluating utilities of B1 and B2 (see Section 3.1 and Eq. (14)). As described in the main text, this was done in two steps. In the first step, the participants assigned weights to different criteria - the presence and survival of a biosignature feature from biological and abiotic sources - irrespective of the specific definition of a biosignature. All four criteria received equal weights. Then, utilities of B1 and B2 were evaluated on each criterion on a scale from 0 to 100. To do so, as it turned out, the participants simply used the previously assigned probabilities, e.g. they assigned utility of 80 if the probability corresponding to a given criterion was 0.8. Applying Eq. (14) yielded results that were essentially the same as those obtained for probabilities. This clearly indicates that evaluations of utilities should have been done on a different group of researchers but, unfortunately, such a group was not available to us for this small pilot study. In addition, half of the participants reported difficulties in interpreting the term “utility” associated with individual criteria. Even though the evaluations in the utility-based approach were uninformative, at least they illustrate the procedure according to which such evaluations are done.

References:

Glavin, D. P., Burton, A. S., Elsila, J. E., Aponte, J. C., & Dworkin, J. P. (2019). The Search for Chiral Asymmetry as a Potential Biosignature in our Solar System. Chemical reviews. Neveu, M., Hays, L. E., Voytek, M. A., New, M. H., & Schulte, M. D. (2018) The ladder of life detection. Astrobiology 18:1375-1402.
C. Accounting for uncertainties in estimating probabilities and utilities

Predicted outcomes of probabilistic models are always burdened with uncertainties. It is therefore essential to provide measures of these uncertainties and their impact on outcome prediction. Without such measures it is unknown how reliable are these predictions, which makes them of only limited value. Estimating uncertainties is particularly challenging for unique or rare events, especially when sources of inaccuracies could be of both stochastic and systematic nature, which is the case in evaluating biosignatures for life detection. In this problem, there are two primary sources of uncertainties. One is associated with the inherent lack of sufficient knowledge about biological and abiotic processes that generate biosignatures. The other is due to perceptual and cognitive biases experienced by experts. This section is devoted to discussing how to estimate these uncertainties.

Assume that the probability of a given state or event has been estimated by an expert as $p$. This could be an aggregate probability, such as $P(B|L)$ or $P(B|\sim L)$, or a more elemental probability, as discussed in section 2.3. This estimate $p$, however, is not precise. Although $p$ is the most likely value, a range of other probability values is also possible, but generally are less likely. Probabilities that operate on probabilities are called second-order probabilities. They provide a measure of uncertainty in the assignment of probability values (Gaifman, 1986, Baron, 1987, Paass, 2013).

The simplest treatment of uncertainties is to assign an interval of probabilities instead of precise values. For example, an expert might evaluate survival probability of a biosignature as being in the interval $[0.1, 0.4]$ instead of attempting to assign a precise value of e.g. 0.25. In this treatment, all probability values in the interval are considered equally likely. Equivalently, second order probability distribution within the interval is uniform. Outside the defined interval, the probability is zero, i.e. it cannot be smaller than 0.1 and larger than 0.4. Thus, narrower intervals imply more certainty in the assignment of probabilities.

A more accurate, but also more complex approach is to model second order probabilities as nonuniform distributions. For example, a domain expert might judge survival probabilities of 0.1 and 0.4 to be less likely than the probability of 0.25 by a factor of three. Probabilities of 0.05 and 0.5 would be also considered possible, but even less likely than 0.1 and 0.4. In general, nonuniform distributions are more realistic than intervals with constant probabilities, but their implementation raises a question: what are suitable mathematical forms of these distributions? The simplest choice would be the Gaussian distribution centered at the most likely probability value. In the example above, this value is 0.25. Then, the variance of the distribution is the single measure of uncertainty. The main drawback of the Gaussian distribution is that it is symmetric, which might not be appropriate in many instances, especially for small or large probabilities. Assume that, according to experts, the most likely survival probability of a biosignature is 0.01, but the chances that this probability is 0.1 is non-negligible. Clearly, this situation cannot be described by way of a symmetric distribution.

An asymmetric function frequently used in probability theory is the beta distribution. It is defined on the interval $[0,1]$, which is precisely the allowed range of probability values. Two parameters
in the distribution control its shape, which could be monotonically increasing, monotonically decreasing, or have a maximum or a minimum. Compared to the Gaussian distribution, a more realistic description of uncertainties is achieved at a price of losing intuitive connection between values of the parameters and the shape of the distribution. For this reason, a computational support algorithm is usually required to extract approximate values of the parameters from a less formal description of the distribution provided by a domain expert.

Another approach does not rely on modeling uncertainties. Instead, uncertainties are assigned on the basis of the variability in evaluations of probabilities made by domain experts. Estimates of probabilities collected from individual experts are treated as a statistical sample in the same way as one would treat repeated measurements of the same quantity. Then, standard statistical methods are used to propagate uncertainties assigned to the elemental probabilities to obtain uncertainties of the global probabilities needed for SDT. Even though some assumptions of the underlying statistical theory are not met in this case, this approach to quantifying uncertainties has been used with some success (Hudomiet and Willis, 2013).

In MAU, uncertainties are due to precision in estimating weights of different criteria and utilities of a given biosignature evaluated on these criteria. Since weights are the same for all biosignatures, so are uncertainties in their assignment. In contrast, uncertainties in utilities have to be assigned for each biosignature separately. Most other considerations are similar to those discussed in the context of probabilities.

In summary, no established theoretical framework exists for assigning uncertainties associated with evaluating probabilities or utilities for single events when only limited background knowledge is available. A considerable degree of arbitrariness is involved in such instances. Simple approaches are based on coarse approximations. More complex ones hold potential for delivering more reliable estimates of uncertainties, but this potential might not be possible to realize because sufficient information is lacking. These problems drove some researchers to question whether assigning uncertainties, especially based on second order probabilities, is at all useful (Hansson, 2008). This point of view, however, is not generally shared. Uncertainty quantification is an indispensable part of analyses in science and engineering. Difficulties in carrying it out reliably should not be the reason to abandon it, but instead it should motivate further research in this area.

References:

Baron, J. (1987) Second-order probabilities and belief functions. Theory and Decision 23:25-36.
Gaifman, H. (1988) A theory of higher order probabilities. In Causation, Chance, and Credence edited by Brian Skyrms & William L. Harper 1:191-219, Kluwer Academic Publishers.
Hansson, S. O. (2008) Do We Need Second-Order Probabilities? Dialectica 62:525-533.
Hudomiet, P., & Willis, R. J. (2013) Estimating second order probability beliefs from subjective survival data. Decision Analysis 10:152-170.
Paass, G. (2013) Second order probabilities for uncertain and conflicting evidence. arXiv preprint arXiv:1304.1139.
D. Tools for representing the relation between stimulus and response

**Decision trees.** Processes represented in a decision tree have to be complete and disjoint. Disjoint processes are such that do not have common elements or causes. Completeness means that all processes of significance are accounted for in the tree. If some of them were omitted there would be no probabilities assigned to them and, consequently, the aggregate probabilities calculated from the model would be burdened with systematic errors.

Decision trees are commonly used in business, social, political and medical decision-making. A conventional application of decision trees is to analyze possible outcomes of investment strategies or placement of novel products with associated risks and benefits. They are particularly suited to represent unique events for which no statistical data are available. They are easy to interpret and understand, but they might become complex and unstable leading to different conclusions depending, for example, on the number of levels included in the decision model.

**Influence diagrams and Bayesian networks.** Decision trees are not the only tool that can be used to represent life detection models. Other tools, such as influence diagrams and Bayesian networks (Kjaerulff and Madsen, 2008) can be also used for this purpose. Both are directed acyclic graphs that have nodes aimed at capturing possible decision options, factors that influence decisions and decision maker’s objectives regarding decision outcomes, for example increased emphases on avoiding false positives or false negatives. In fact, influence diagrams and Bayesian networks are closely related. Bayesian networks are often used in medical fields to analyze probabilistic relationships between a disease and its symptoms, a problem that is in many respects similar to the relationship between life and biosignatures. Bayesian networks provide a structured formalism to evaluate aggregate probabilities and typically involve less nodes than decision trees but are less intuitive to build and understand. All approaches discussed here can be used not only to help evaluate probabilities but also utilities within the utility framework discussed in Section 3 in the main text.

**References:**

Kjaerulff, U. B., & Madsen, A. L. (2008) Bayesian networks and influence diagrams. *Springer Science+ Business Media.*
E. Comparative and non-comparative measurements

Non-comparative measurements. Non-comparative measurements are carried out independently for each object on a response scale that corresponds to a well-defined numerical range, e.g. between 0 and 10 or 0 and 100. Such response scales are considered to be interval scales, i.e. they are well-ordered and keep equal intervals. This means that, for example, the difference between 0 and 1 is the same as the difference between 9 and 10. These properties guarantee that all evaluations remain in exactly the same relation if a scale is subjected to a linear transformation. Scales could be either discrete or continuous. For example, a discrete scale between 0 and 10 involves 11 categories. Continuous scales are usually implemented by way of sliders with only the ends labeled. Probabilities or utilities associated with biosignatures evaluated on discrete scales are defined only in a range determined by a number of categories. The same quantities evaluated on a continuous scale are given precise values. This, however, does not imply that continuous scales are more accurate, as discussed below.

A considerable body of work has been directed at constructing and characterizing non-comparative scales. A number of questions have to be answered. How many categories should be used? Should there be an even or an odd number of categories? An area of particular interest is the effect of the number of response categories on reliability and validity. The former measures consistency of responses whereas the latter establishes whether the scale actually measures what it intends to measure. Lozano et al. (2008) have shown that when the number of response categories increases, both reliability and validity also increase, but this relationship is not linear. If the number of categories increases above the bare minimum of four, reliability and validity initially increase rapidly, but then levels off at approximately 7 categories. Above 11 categories, there is hardly any further improvement. This means that a sufficiently broad discrete scale and a continuous scale should be similarly reliable and valid.

Comparative measurements. Comparative scales are not burdened with such problems as the number of categories or labeling since objects are directly compared with each other. The most common comparative measurement technique is ranking objects in a descending or ascending order. All that has to be decided is, for example, whether biosignature A is more likely to survive than biosignature B, or that argument 1 is stronger than argument 2, which in turn is stronger than argument 3. Distances between objects, however, remain unknown. If rankings from a number of experts are available, it is possible to estimate these distances and place the evaluations on an interval scale. First, the average rank and the corresponding variance are calculated for each biosignature. Then, the distance between two biosignatures or arguments having consecutive average ranks is assigned on the basis of the difference between these average ranks and the variances. The larger the difference and the smaller the variances, the larger is the distance. Even though ranking is in principle easier to carry out than non-comparative evaluations, it suffers from a disadvantage when applied to biosignatures. Their nature is so diverse that experts are likely to find it very challenging to rank all of them competently even with the aid of a thorough knowledge base. In contrast, each expert can perform non-comparative evaluations only for a subset of biosignatures within own range of expertise.

Another way to elicit comparative evaluations is through pairwise comparisons of biosignatures or arguments. Again, the distances between objects cannot be determined directly,
but a similar technique to the technique described for ranking can be used to estimate these distances. Pairwise comparisons do not suffer from the same disadvantage as ranking, as each expert can carry out comparisons on a subset of objects within her/his expertise, but they have problems of their own. Specifically, the number of comparisons is quite large. Also, transitivity (see Section 3 in the main text), a basic requirement of formal logic, is not guaranteed.

References:

Lozano, L.M., Garcia-Cueto, E., Muniz, J. (2008) Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 4:73-79.