Similarity Measures for Automated Comparison of In Silico and In Vitro Experimental Results

Glen E. P. Ropella, Dev A. Nag and C. Anthony Hunt*

Department of Biopharmaceutical Sciences, Biosystems Group,
The University of California, San Francisco, CA 94143, USA

Abstract—**The overwhelming complexity of biological systems prevents exhaustive description of the processes and mechanisms that cause system functionality. There are large numbers of processes to be considered with options for manifold hypotheses describing each. The long-term goal of this project, for a particular biological system, is to put the computer to work weeding out the weaker hypotheses and, even, weeding out posited processes that do not pertain directly to specific functionality. An objective towards this goal is to build a computational framework to host an ongoing competition for the most effective structural description of what goes on inside an organ, in this case the liver. In order to do that, one needs robust algorithms for comparing the data taken from biological experiments with the data taken from the simulation. In this paper, we begin to delineate and survey algorithms by which to compare the output of any given simulation with data taken from experiments.**

Key Words— **Computational biology, computer, in silico, liver, model, modeling methodology, simulation, validation, similarity**

I. Introduction

A new research domain for computational biology is the design and experimental use of representations of biological processes that are constructed using an Object Oriented Programming language. We use software agents that are designed to represent specific biological organisms or their components and their interactions within a framework that, in part, represents the real environment. The in silico interactions are intended to mimic observed or inferred biological interactions. Experiments may then be done on these in silico constructs. Observer agents within the framework collect and report observations. An obvious goal of this exercise is to build these constructs to reflect and test current biological concepts, and to organize and parameterize the constructs so that their operation reflects actual experiments. If the results of in silico experiments and actual biological experiments are indistinguishable, then the parameterized construct may be a useful model of the biological components under study. An expert can inspect the two sets of results and offer an opinion on the degree of their similarity. Automated model generation and refinement requires having one or more well-defined similarity measure to substitute for the expert's judgement. Classical regression approaches do not apply because we are comparing the outputs of two or more experiments. We are not seeking a regression function that can account for one or more of the data sets. Here we explore several similarity measures when the experimental results are comprised of time series data. The biological experiments of interest use an isolated perfused rat liver (IPRL) [1]. The experimental results are the fraction of the administered solute (typically a drug) per unit volume at intervals in the collected perfusate.

II. Background

A. Experimental Indistinguishability

Replicate in vitro IPRL experiments conducted on the same liver provide similar but not identical solute outflow profiles. The same is true for in silico experiments. In vitro, there are two main contributors to intraindividual variability, methodological and biological. For replicate experiments (in the same liver) the coefficient of variation (CV) for fractional solute outflow during the various collection intervals ranges between 10 and 40%. The CV can define a continuous interval bracketing the experimental data. By definition—all things being equal—the results of replicate experiments are indistinguishable. A new set of results that falls within a specified CV interval range and has essentially the same shape seen in prior experiments is defined as being experimentally indistinguishable even if it comes from an in silico IPRL experiment. Under those conditions, based on the data alone, there would be no way to determine whether a data set came from an in vitro or an in silico experiment. These observations provide the basis for selecting and evaluating similarity measures.

The character of the outflow profile is presumed to carry more information than can be extracted by simple statistical estimators. The objective of the similarity measure is to help select amongst various models of the liver, not simply to assume a model and select amongst variations on that model. Hence, the successful similarity measure must target information in the profiles that correlate with the generative structures and building blocks inside the model,

* To whom correspondence should be addressed: *hunt@itsa.ucsf.edu*.

in addition to the more satisfaction of straightforward statistical indistinguishability.

The phenomena of interest are the various features of the outflow profile. Neither an instantaneous, per observation, comparison, nor a whole-curve comparison is appropriate because the three features, the tail, the peak, and the transition between them, clearly show different variances. For these reasons, we believe a multiple observation similarity measure is warranted.

B. Comparing Time Series

In the context of an IPRL, an in silico experimental time series reflecting the solute outflow fraction is captured and compared with *n* time series taken from a real liver. These *n* nominal time series often look similar, and our goal is to create a quantitative measure for how similar they are to each other and how similar the in silico experimental time series is to that group. Our approach will mostly involve comparing the distance of each of the nominal time series from their mean, applying the same measure to the in silico time series (against the nominal time series mean), and then ranking or comparing the resulting values.

For example, the first similarity measure studied (the global standard deviation measure described below) provided a number for the in silico data (typically between 0.5 and 0.7). However, this quantity only makes sense in the context of the same metric applied to each of the nominal time series (none of which will be 100% within the similarity bands). A similarity score of 0.6, for example, might be higher than any one of the nominal data similarity scores. This relationship can be quantified by further computing a probability density function of similarity scores (on the nominal data) and seeing where, within that, the in silico similarity score falls.

III. Methods

A. Criteria

The similarity metrics should satisfy a number of criteria:

- The measure should roughly agree with a visual sense of similarity on the charted data. Because the purpose of the in silico experimentation is to **invalidate** models, validation will consist of both the results of the measure and careful visual comparison between the graphs of the original data.
- The measure should incorporate the variance of the group of nominal time series. A wider variance in that time series should cause the in silico results to be judged more similar to the nominal cluster.
- The measure should make no strict assumptions about the nature of the series distribution. In other words, measures which assume a Gaussian, Binomial, or Poission distribution may provide misleading similarity values in our context. We need robust, non-parametric algorithms that do not require these kinds of assumptions.

As we explore potential measures, other requirements may come into play. We note that metrics are commonly used to measure similarity. However, they often do not satisfy criteria 1, or if they do, the successful metric is complicated and counter intuitive. [2]

We will denote each of the *n* nominal time series by $\{d_i\}_j$, where *i* ranges over time (from 1 to *T*) and *j* represents the series index (1 to *n*). The nominal series mean is denoted by $m_i = \langle \{d_i\}_j \rangle$, where the mean value is taken with respect to *j*. The experimental in silico data series is a_i. We will assume that the series are normalized with respect to time such that all series have the same cardinality (e.g., *T* is the same for all series).

B. Data as Vectors in a Metrizable Space

One class of measures portrays each time series as a single, *T*-dimensional vector in some metrizable space.

Minkowski: The Minkowski metric is a generalization of the City Block and Euclidean metrics, and in fact has a parameterization that reduces it to those two (city block for $p = 1$, and euclidean for $p = 2$).

$$metric_{Minkowski}(p) = \left(\sum_{i=1}^{T} |c_i - m_i|^p \right)^{1/p} \qquad (1)$$

While the Minkowski metric is interesting, There are apparently no references to usage for $p > 2$. As $p \to \infty$, the dimension *i* which maximizes $|c_i - m_i|$ takes over the calculation (in the limit, the metric becomes that maximum distance).

Mahalanobis: The Mahalanobis distance is similar to the euclidean distance, but it has the advantage of automatically scaling and adjusting for the correlation of different dimensions.

$$metric_{Mahalanibis} = (c_i - m_i)' M^{-1} (c_i - m_i) \qquad (2)$$

where *M* is the covariance matrix, and *v*' denotes the transpose of a vector *v*. When *M* is taken as the identity matrix, the Mahalanobis distance reduces to the Euclidean distance.

The covariance matrix (μ_i = mean x_i) is calculated by taking the pairwise covariances of *N* variates $\{x_i\}_{i=1}^{N}$:

$$M_{a,b} = \mathrm{cov}(x_a, x_b) = \left\langle (x_a - m_a)(x_b - m_b) \right\rangle \qquad (3)$$

In our context, we would assume that each time index in the group of nominal data series represents a different variate, and each variate (time slice) took on a number of values (one for each of the series). The downside of the Mahalanobis metric is that the calculation of the covariance matrix may be somewhat unreliable is data limited. The covariance

matrix would be a 200 by 200 matrix, but each entry would only be calculated from 4-6 different data points, greatly increasing the prevalence of statistical artifacts. If we had a much larger number of nominal data series, the Mahalanobis metric could be more useful.

C. Data as Time Series with Envelopes

Each of the following three approaches attempts to create bands around the mean of the nominal series, providing an envelope of similarity. The actual similarity score can be computed as the number of observations from the candidate series that fall within the envelope. An alternative algorithm could compute a quantity at each time for how far each observation was from the nominal mean, normalized by the envelope boundary – then, an average over time of that difference would provide a single quantity reflecting the average normalized deviation from the nominal mean.

Global Standard Deviation: This is the similarity metric implemented in the current IPRL framework. First, a coefficient of variation is computed on the multiple nominal time series. Then we pool the normalized data and take the standard deviation.

$$\{n_i\}_j = \frac{\left(\{d_i\}_j - m_i\right)}{m_i} \qquad c_{variance} = std(\{n_i\}_j) \qquad (4)$$

The envelope is then defined using: $upper_i = m_i*(1 + c_{variance})$, $lower_i = m_i*(1 - c_{variance})$.

These bands form the similarity envelope around the nominal mean time series. The similarity score is then calculated by counting the number of observations of the candidate time series that fall within the envelope and dividing that by the total observations in the series.

Local Standard Deviation: An alternative to the above is to use the deviation of the group of nominal series at each time slice. The upper and lower bands are simply the nominal mean values plus or minus the instantaneous standard deviation.

$$sd_i = std(\{d_i\}_j) \qquad (5)$$

$$upper_i = m_i + sd_i \quad lower_i = m_i - sd_i$$

Windowed Standard Deviation: An alternative is to use the standard deviation within a limited sample, *H,* of the nominal mean time series. When m_i is changing dramatically, the envelope widens; when the m_i is constant (or close to it), the envelope shrinks. This concept is exploited in stock trading models (Bollinger Bands).

$$ma(H)_i = \frac{\sum_{delay=0}^{H-1} \{m_{i-delay}\}}{H} \qquad (6)$$

$$sd(H)_i = sd\left(\left\{m_{i-delay}\right\}_{delay=0}^{H-1}\right) \qquad (7)$$

where the standard deviation is taken with respect to *delay*. The envelope is defined as: $upper_i = ma(H)_i + sd(H)_i$, $lower_i = ma(H)_i - sd(H)_i$

D. Data Interpreted as Stochastic Series

Pearson Correlation: A simple test for the relationship between two ordered series is their Pearson correlation coefficient. It ranges between -1 and 1, and has a larger magnitude for data that is more closely related (with coinciding variations). The correlation is completely insensitive to scaling in the range dimension.

$$r = \frac{T(\sum_i m_i * c_i) - (\sum_i m_i)(\sum_i c_i)}{\sqrt{(T\sum_i m_i^2 - (\sum_i m_i)^2)(T\sum_i c_i^2 - (\sum_i c_i)^2)}} \qquad (8)$$

As a distance metric, we can use (1 – r)/2, or a similar function of the correlation coefficient. However, the Pearson correlation assumes a normal distribution of the values within the two series. Since this clearly does not apply to the liver data, the utility of this metric may be limited.

Spearman Correlation: The Spearman correlation (AKA, rank correlation) rectifies the distribution assumption of the Pearson correlation. The data vectors are transformed into rank vectors, and the correlation is computed on the rank vectors (rather than the data vectors directly).

The rank vectors are the set resulting from the permutation of the numbers in the ordered set {1, 2, … *T*}, where that permutation applied to the data in the original vectors would order the data from lowest to highest. For example, the data set {2.4, -0.42, 1.3, -4.0} would become the rank set {4, 2, 3, 1}. This Spearman correlation has the advantage of being non-parametric. As above, we could use (1 – r)/2 as a distance metric.

E. Data Interpreted as a Probability Distribution Function

Another way of looking at the data is as a raw probability density function (PDF). For example, we can normalize both the nominal mean and candidate time series:

$$m_i' = \frac{m_i}{\sum_i m_i} \qquad c_i' = \frac{c_i}{\sum_i c_i} \qquad (9)$$

such that the normalized series both add to 1. Now, we can apply some of the above measures to these normalized functions (e.g., the Euclidean distance). There are, however, more powerful and interesting tests available.

Kolmogorov-Smirnov: We can take the cumulative distribution functions (CDF) of the two normalized series above:

$$cdf_m(x) = \left\{\frac{i}{T} : i = \text{cardinality}(m_i' \le x)\right\} \qquad (10)$$

$$cdf_c(x) = \left\{ \frac{i}{T} : i = \text{cardinality}(c_i' \le x) \right\} \quad (11)$$

These CDFs are defined everywhere on (-∞, ∞), even for a PDF with compact support. For continuous functions, the CDF is defined:

$$CDF(x) = \int_{-\infty}^{x} PDF(f)df \quad (12)$$

The Kolmogorov-Smirnov (KS) test is the maximum distance between these two CDFs:

$$metric_{KS} = max_x\{cdf_m(x) - cdf_c(x)\}. \quad (13)$$

In addition to being non-parametric, the KS test actually provides a p-value for testing the hypothesis that the two distributions are the same (using a D-statistic table). A KS-value (or p-value) can be calculated for each of the nominal data series (against the nominal series mean) as well as for the candidate. The KS test is also more sensitive to data near the center of the distributions than at the tails. However, this is not harmful in our context because the most interesting behavior does occur at the center of the distribution.

Cramer Von Mises: This test, similar to KS, provides a metric which can be transformed into a p-value.

$$metric_{CVM} = W^2 = \int_{-\infty}^{\infty} [cdf_c(x) - cdf_m(x)]^2 m_i' dx \quad (14)$$

However, the Kolmogorov-Smirnov test seems more widely accepted, and is probably more meaningful to a broader audience.

F. Structural versus Numerical Comparisons

The measures listed so far are numerical. Quantitative properties of the data are abduced from the data and that provides a quantitative comparator. However, there are also techniques by which more qualitative or structural properties of the data can be abduced. Structural pattern recognition [3] provides us with a number of techniques for judging similarity irrespective of circumstantial detail, scale, noise, or even partial matches. Interestingly, the successful and pedigreed compartment model for liver clearance [1] consists of signals described and convolved in the frequency domain and then transformed into the time domain. Because those signals are based, in part, on features of liver clearance profiles, we have good reason to believe that a feature extraction (or other pattern-matching) technique will work on this data, regardless of origin. However, because of limited space we include here by reference the details and observations in [3].

IV. Conclusions

The first version of the in silico IPRL used the "global standard deviation" measure described above. It is clear that this measure does not satisfy requirement 1 because low similarity scores do not account for candidate series that match the nominal series in shape and character in spite of having a large number of observations that fall outside the envelope. It also seems clear that the strictness of the measure should be proportional to the deviation between the nominal series. Despite these shortcomings, however, the current similarity measure does provide a way for the framework to select against models that produce "invalid" results.

The measures listed above represent our initial survey of the possible selection criteria we could use. We will choose a subset of these and those we have yet to find to implement and evaluate based on their practical and theoretical attributes.

Acknowledgment

This work was supported in part by funding provided by the Biosystems Group's Royalty Fund. We thank the members of the Biosystems Group—the Hunt lab—for and helpful discussions.

References

[1] M.S. Roberts and Y.G. Anissimov, "Modeling of Hepatic Elimination and Organ Distribution Kinetics with the Extended Convection-Dispersion Model," *J. Pharmacokin. Biopharm.* vol. 27, no. 4, pp. 343-82, 1999.

[2] S. Santini and R. Jain, "Similarity Measures", *IEEE Trans. Pattern Analysis and Machine Intelligenc*e, vol. 21, no. 9, pp. 871-83, Sept. 1999.

[3] R.T. Olszewski, "*Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*." PhD Thesis CMU-CS-01-108, Carnegie Mellon Univ., 2001.