SPEDRE: a web server for estimating rate parameters for cell signaling dynamics in data-rich environments

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1093/nar/gkt459">http://dx.doi.org/10.1093/nar/gkt459</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Oxford University Press</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Sun Jun 18 05:21:30 EDT 2017</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/79751">http://hdl.handle.net/1721.1/79751</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution Non-Commercial</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc/3.0">http://creativecommons.org/licenses/by-nc/3.0</a></td>
</tr>
</tbody>
</table>
ABSTRACT

Cell signaling pathways and metabolic networks are often modeled using ordinary differential equations (ODEs) to represent the production/consumption of molecular species over time. Regardless whether a model is built de novo or adapted from previous models, there is a need to estimate kinetic rate constants based on time-series experimental measurements of molecular concentrations or enzyme activities. Even when time-series experimental measurements of all species are available [such as by using stable isotope labeling by amino acids in cell culture (SILAC) proteomics (2)], estimating the rate constants is still a difficult non-linear optimization problem (3). Many widely used methods, collected into popular software packages such as COmplex PAthway Simulator (COPASI) (4), are applicable to this parameter estimation problem. Methods have traditionally been classified as local, global or hybrid global+local methods (5,6).

The traditional application of parameter estimation for modeling network dynamics has been to small biochemical networks with sparse data sets. With the growing ease of measuring complete proteomes (7) and with the assembly of large network models, needs have expanded to include data-rich approaches to parameter estimation, sometimes called spline-based collocation methods (5,8–10). Spline-based collocation methods exploit complete or nearly complete data sets to interpolate directly the slopes of concentration over time, instead of relying on numerical simulations to compute the derivatives of the ODEs. Spline-based collocation methods have not previously been implemented in any parameter estimation web server. The method of Systematic Parameter Estimation for Data-Rich Environments (SPEDRE) (11) uses a spline-based collocation approach and coarse-grained discretization to provide approximate heuristic search of the global parameter space, with excellent scalability for large networks. However, SPEDRE is designed only for problems with low-degree networks (few reactions per protein, but no limit on the number of proteins). Although there are many algorithms for parameter estimation [reviewed in (12,13)], we are not aware of
any parameter estimation web servers, other than COPASI (14,15) and SPEDRE.

**PROCESSING METHOD**

Parameter estimation in biological networks is challenging because the parameters are interdependent, which makes them impossible to estimate individually. Even if each parameter can take on only a fixed number of possible values, the possible combinations of these parameter values becomes astronomical, growing exponentially with respect to the number of reactions. SPEDRE exploits combinations from the low-dimensional subproblems. We do this heuristically using Loopy Belief Propagation (16), also called ‘message passing’, a probabilistic network inference technique that computes probability distributions for the parameters and sends the distributions as messages across the edges of the network. Loopy Belief Propagation is an iterative heuristic that terminates by convergence or when the specified maximum number of iterations is reached. On termination, Loopy Belief Propagation provides optimized bins for all rate constants. This set of bins provides a starting point for the post-processor to refine, using the Levenberg–Marquardt numerical method of local optimization (17). In essence, SPEDRE is a hybrid global+local optimization method, but unlike other hybrid global+local methods, the global portion (called SPEDRE-base) does not use stochastic sampling. The final output of SPEDRE is a plot of the bins provided by Loopy Belief Propagation, as well as a vector of the optimized rate parameters. Our previous work specified the SPEDRE algorithm in detail (11), whereas the current work aims to describe the web server interface.

Asymptotic analysis of the underlying SPEDRE-base algorithm (11) reveals attractive properties of the method because the time complexity scales exponentially with the network degree, but it scales efficiently (polynomially) with the number of species, the number of reactions and the size of the data set. Correspondingly, the method scales well on biological pathways with a bounded number of reactions per species. Dense networks with hub-like species have high network degree and are unsuitable for the SPEDRE algorithm. Our web server handles these cases simply by running the Levenberg–Marquardt algorithm (17) instead.

The SPEDRE-base algorithm was implemented in C++ with an interface to COPASI (version 4.6, build 32) for the Levenberg–Marquardt algorithm. The web server version of SPEDRE was implemented using the Opal toolkit, as introduced in (18). To display the customizable bin plot of SPEDRE results, Google Chart API (Google Inc.) was used. As the features of Opal and Google Charts API grow over time, the functionality of the SPEDRE web server will grow as well.

**INPUT**

SPEDRE requires two main inputs from users: the concentrations of all the molecular species and the biochemical reactions. The concentrations of the molecular species would come from proteomic experiments, from computational hypotheses that merge published sub-networks or from any source that specifies the abundance of every species, at a set of discrete timepoints. This requirement for input of all species is highly restrictive, but it allows SPEDRE to focus on data-rich problems, which are a specialized but growing segment of parameter estimation work. The concentrations must be specified in a comma separated value (CSV or TXT) file. The file format is compatible with time-course simulation output from the COPASI software, where the first line specifies the headers (time and species names). The first column of each additional row specifies the time value, with the species levels in the remaining columns. If a data point
is found to be invalid, the relevant time point is removed from the system before performing rate constant estimation. If there is greater incompleteness or sparsity in the available measurements, users should use the CopasiWeb (14) service instead.

The biochemical reactions must be specified as an XML file in COPASI ML format (4), which can be obtained from SBML format using a link to the conversion service in CopasiWeb (14). This format includes ‘rate laws’ with predefined reaction types, which are necessary for constructing the correct types of kinetic parameters. SPEDRE is currently restricted to the following reaction types: ‘Michaelis-Menten catalysis’, ‘Mass action (irreversible)’ and ‘Enzyme simple’ [rate_constant × (enzyme) × (substrate)]. In other words, a reversible reaction must be represented using two separate reactions, one in each direction; reactions with high-order combinations must be re-expressed as a series of subreactions. Future work may automate the conversion process. The SBML conversion service provided at CopasiWeb (14), may disrupt the names of the reaction types (‘rate laws’), in which case users must change the names of the reaction types manually.

The web server homepage provides descriptions and illustrations for several published pathways, including the Akt pathway (19), the MAPK pathway (20) and a pathway of Actin Filament Assembly-Disassembly (21). Also available is a spectrum of artificial networks (circular or tree-shaped) with widely varying sizes.

A set of default parameters can be modified using the submission form, if users wish to adjust how SPEDRE executes. The main options are the number of bins for discretizing the parameter ranges, and the number of iterations for Loopy Belief Propagation. In addition, the bin spacing can be set to linear or logarithmic scaling. The upper and lower bounds of the rate constants can be specified globally, or individual rate constants can override the upper bound and lower bound if these are specified in the network connectivity input file. The maximum number of iterations can be set to zero if users wish to perform a standalone local search. The anticipated error rate is an option for theoretical calculations, and it allows users to add Gaussian noise to the observed data. Another option for specialized users, called ‘samples per voxel’, allows each voxel of parameter space (each set of parameter bins) to be evaluated by sampling multiple random points in the voxel, instead of using the voxel midpoint.

**OUTPUT**

An example execution based on the MAPK cascade is shown in Figure 2. Using the web interface, users can submit input files that follow the specified formats (Figure 2, top box), and SPEDRE performs the computation task while simultaneously displaying the execution page (Figure 2, middle box).

Different execution specifications may result in different runtime performance, and some jobs may require several hours to complete. Users may wish to bookmark the location of the output page for a later visit. A status page is linked to the execution page and shows the percentage of the overall task that has been completed. Users are advised to consult the asymptotic analysis of SPEDRE-base algorithm (11) when adjusting the SPEDRE execution options beyond the default values. On completion, SPEDRE returns an estimated range for each rate constant, in the form of a bin plot, as shown in Figure 2 (bottom box). A bin plot is a visual representation of the resulting voxel in high-dimensional space, which gives users an impression of the exponential number of possible combinations of rate constants, even in a coarsely discretized search space. Each bin indicates a range in which the estimated rate parameter lies. Finally, the bin midpoints are used as a starting point for local optimization, and the refined set of parameters are output as a vector of floating point numbers.

The bin plot in Figure 2 was generated using the Google Chart API (Google Inc.), which imposes certain constraints, including the maximum URL length of 2048 characters (for plot formatting) and maximum plot size of 300 000 pixels. Users may encounter cluttered plots
for large network size (>30 rate constants). As the API is actively developed with large user base across the industry, this current limit will be overcome with new updates of the API.

**PERFORMANCE**

Depending on execution configuration, SPEDRE can achieve various performance outcomes. Table 1 shows the web server’s performance on five test cases using noiseless and 20% noise data. The weighted sum-of-square error (SSE) represents how well a model with estimated parameters could fit the input time-series data.

These test cases used coarse discretization to run quickly at the expense of accuracy. For the PI3K/Akt cascade and MAPK cascade, the objective function was low, indicating good match with data. (A good match with data means the parameterized model gives plausible explanations of the data, but alternative models or parameters may also exist). The runtime of SPEDRE base (i.e. SPEDRE without Levenberg–Marquardt) was close to the total SPEDRE runtime, indicating that the hybrid global-local approach incurs low additional runtime cost compared with global search alone. The runtime measures also provide an empirical demonstration that SPEDRE runtime scales efficiently with the size of the input network and poorly with the degree of the network. Specifically, the high-degreded Actin Filament Assembly/Disassembly pathway has only 14 species and 25 reactions, whereas the low-degreded circular network has 80 species and 80 reactions; yet, execution on the latter network completes ~23 times faster than the former.

**DISCUSSION**

SPEDRE has been implemented as a web-based service for performing rate constant estimation on biochemical networks, such as cell signaling pathways and metabolic networks. SPEDRE uses a spline-based collocation approach, requiring extensive data as input and providing efficient coverage of enormous parameter spaces. The computational power of a web server makes it suitable for intensive rate constant estimation jobs. The server has dynamic display of the bin plot, as shown in Figure 2 (bottom box), which is customizable using JavaScript. SPEDRE performs preprocessing of user inputs to eliminate missing or invalid data points from the data file. In the scenarios involving a dense network or other features that violate the requirements of SPEDRE, the web service will perform Levenberg–Marquardt optimization only, as an automatic ‘rescue’ for the parameter estimation problems.

This service is not predictive because measured rate constants are not yet available for pathways of significant size. For users who wish to address the accuracy of parameter estimation as a purely mathematical problem, artificial data sets are available and the weighted SSE is displayed.

**FUNDING**


**Conflict of interest statement.** None declared.

**References**