

# CaliBayes: Integration of GRID-based post-genomic data resources through Bayesian calibration of biological simulators



Darren Wilkinson, Tom Kirkwood, Richard Boys  
Schools of Mathematics & Statistics and Clinical Medical Sciences  
University of Newcastle upon Tyne, UK

BBSRC Bioinformatics and e-Science Programme II  
BBS/B/16650, January 2005 – December 2007



[d.j.wilkinson@ncl.ac.uk](mailto:d.j.wilkinson@ncl.ac.uk)  
<http://www.staff.ncl.ac.uk/d.j.wilkinson/>

## Summary

The primary aim of this project is to capitalise on the development of GRID-based modelling and simulation resources such as BASIS ([www.basis.ncl.ac.uk](http://www.basis.ncl.ac.uk)) by building a higher level computational GRID facility designed for integration of multiple post-genomic data resources. This facility, based on state-of-the-art Bayesian calibration techniques, will call upon GRID-based biological simulators for forward simulation whilst solving the inverse problems facilitated by data integration internally. Thus the project will provide a powerful new tool for the academic community that will enable inferences to be made about parameters and relationships within large network models of biochemical pathways and cell signalling systems.

The poster provides an outline of the distributed computing architecture that we intend to adopt, emphasising the added-value that a GRID-based approach provides. It also gives an overview of Bayesian calibration techniques and explains why they perform much better than more naive optimisation methods in the context of calibrating large and complex biological process simulation models.

## 1 Problem outline

- Large, complex biological simulation models typically contain many parameters whose values are uncertain.
- There are increasing amounts of post-genomic data being made available on-line, that can in principle be used to calibrate simulation models.
- This motivates the creation of a GRID-based inference engine that will utilise both GRID-based simulators and GRID-based data repositories in order to accurately estimate model parameters.

## 2 Computing architecture

- A service-oriented architecture (SOA) will be adopted, and all resources will communicate via SOAP Web Services, using any appropriate extensions to WS-I that are widely adopted by the UK e-Science community (eg. WS-Security).
- Due to the fact that it is not possible to know *ab initio* how simulator output relates to the contents of a data repository, intermediate web services will need to be constructed.
- For each GRID-enabled simulator that is to be used for model calibration, a service is required that will transform the simulator output into the same format as the data that is to be used for calibration purposes.
- For each GRID-enabled data repository that is to be used to calibrate models, an appropriate comparison service will be required in order to measure data similarity. This will use multivariate analysis to compare experimental and simulated data.

- With these interfacing services in place, the main CaliBayes service will carefully select simulator runs with various parameter combinations sequentially on the basis of knowledge of all previous runs in order to find a combination of parameters giving the best match to available experimental data.
- Biological models will typically be encoded in the Systems Biology Markup Language (SBML, [www.sbml.org](http://www.sbml.org)), with unknown parameters marked separately. Minimal information will be extracted from these models by the main CaliBayes engine, and then the full model will be dispatched to an appropriate SBML-aware service for full time-course simulation.
- Using a GRID-based SOA means that we do not have to provide a large simulation facility or mirror large post-genomic data repositories, but simply access processing power and data as and when required. It also means that the service will scale well as new simulators and data repositories come on-line.

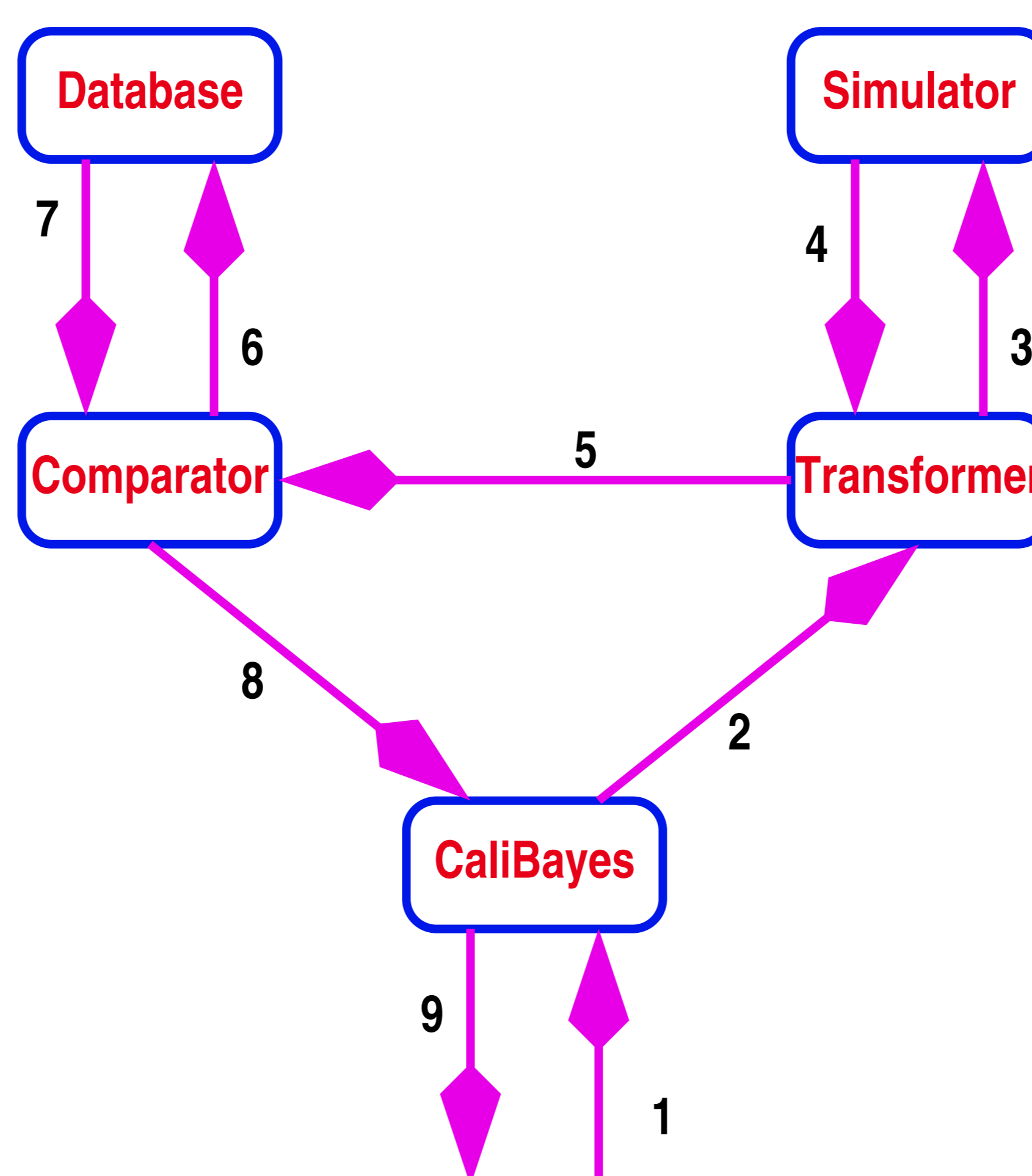


Figure 1: Basic architecture depicting key message flows between the main services, with flows 2–8 repeated until convergence

## 3 Bayesian calibration

### 3.1 The problem

- Traditional optimisation techniques (eg. steepest descent) are very wasteful of information as they use only the last one or two function evaluations to form the next “guess” at the optimum parameter set. They are also “myopic” in terms of their search strategy.
- Some modern optimisation techniques (eg. simulated annealing, genetic algorithms, etc.) are

impractical if function evaluations are expensive (eg. the simulator takes a long time to run), as they require vast numbers of runs.

### 3.2 The solution

- Bayesian calibration techniques are particularly efficient in terms of the required number of function evaluations, as all available runs are used in order to learn about the relationship between the parameters and the goodness-of-fit of the simulator output to available experimental data.
- New function evaluations are chosen carefully in order to minimise uncertainty in the region of the predicted optimum.
- The search strategy is not “myopic”, so a new run will not typically use the current predicted optimal parameter set, as this will typically not give the most information about the location of the true optimum parameter combination.
- The predicted optimum parameter combination will not typically correspond to any run that has actually taken place.
- The goodness-of-fit function is estimated by the main CaliBayes service, using Bayesian Gaussian process regression.
- Adopting a Bayesian approach provides a coherent framework for simultaneous estimation of both the form of the Gaussian process and its associated properties.
- The Bayesian approach also provides a natural framework for allowing incorporation of prior information regarding realistic parameter settings that will help to narrow down the search space.

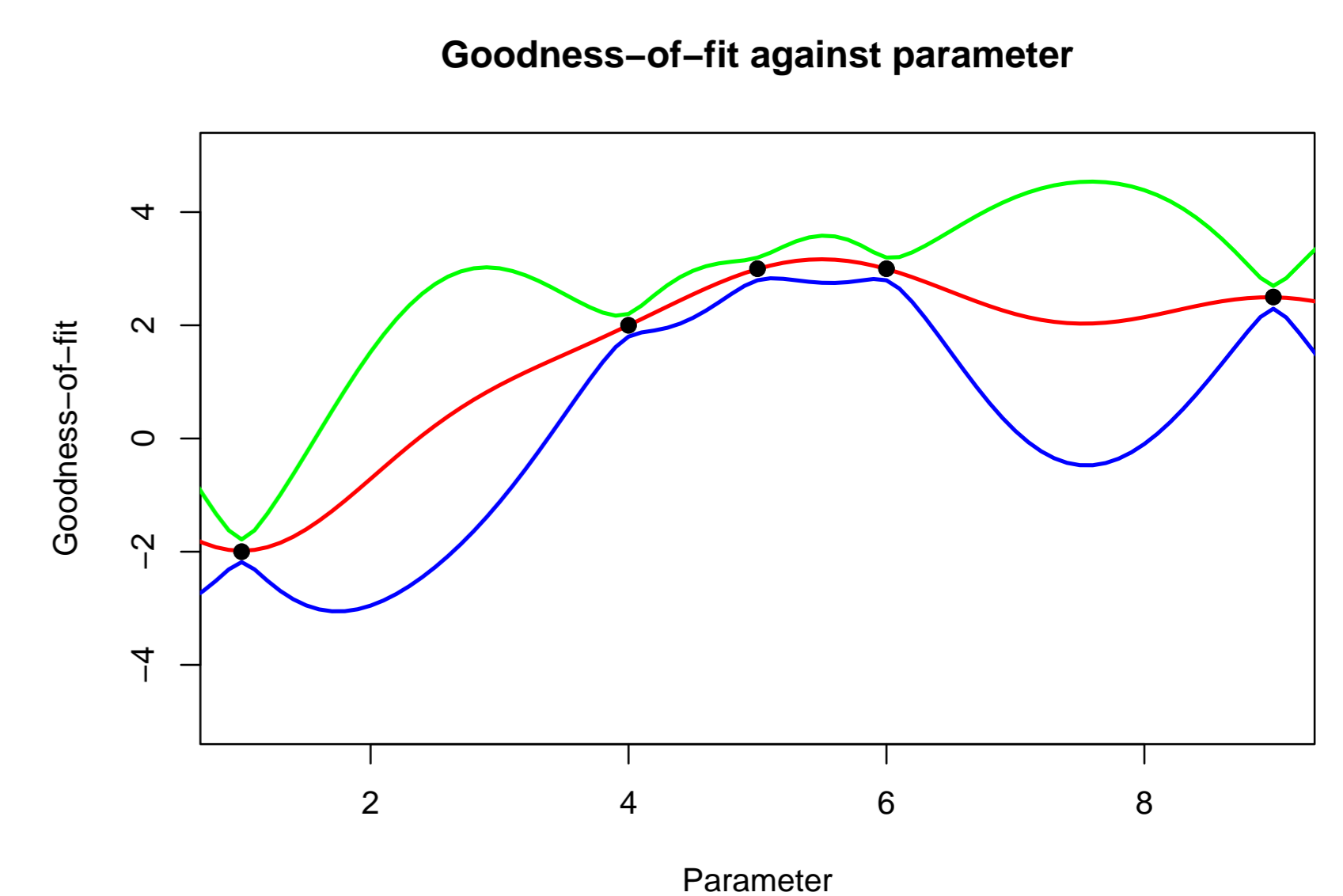


Figure 2: Illustration of calibration concepts with a single parameter and five simulator runs. Goodness-of-fit is only known for five parameter settings, but can be estimated elsewhere, together with a measure of uncertainty. The estimate and confidence bounds are displayed as solid lines on the graph. The current predicted optimum parameter is around 5.5. However, if choosing another run, it may be best to use a parameter around 7.5, as there is currently considerable uncertainty regarding the value of the simulator in this area of the parameter space.