Chemometrics and Intelligent Laboratory Systems 149 (2015) 28–38

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Querying Bayesian networks to design experiments with application to 1AGY *serine esterase* protein engineering

Debora Slanzi [a,b,*], Davide De Lucrezia [c], Irene Poli [a,b]

[a] Department of Environmental Science, Informatics and Statistics, University Ca' Foscari, Canareggio 873, 30121 Venice, IT
[b] European Centre for Living Technology, Ca' Minich, S. Marco 2940, 30124 Venice, IT
[c] Explora Biotech, Via della Libertà 9, 30175 Venice, IT

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form 24 September 2015
Accepted 27 September 2015
Available online 9 October 2015

Keywords:
Bayesian networks
Design of experiments
Optimisation
Protein engineering

ABSTRACT

Current experimental research in several scientific areas must deal with the issue of high dimensionality and complexity. In particular, experimental design strategies are hindered by the limited number of points that can be tested due to technical and economic constraints. In this paper we propose a novel approach called QueBN-design (Querying Bayesian network design) derived by coupling conditional probabilistic inference in Bayesian network models and evolutionary principles. As proof-of-principle, we evaluate the performance of our approach in a simulation study achieving very good results also in comparison with other commonly used designs. Further, we address the problem of engineering synthetic proteins, and in particular the 1AGY *serine esterase* protein. Also in this case results indicate that QueBN-design can effectively guide the search in very large experimental spaces testing a very limited number of points, outperforming other evolutionary and classical benchmark designs.

1. Introduction

A challenging problem in current scientific research is the design of experiments for systems characterised by high dimensionality and complex variable interactions. The very large number of variables and the common nonlinear interrelations can put in difficulty classical hypotheses of statistical design and multidimensional modelling. Also testing a very high number of experimental compositions can be too costly and can produce a negative effect on the environment. These issues have been recently addressed from different perspectives, and several approaches have been developed, mostly focusing on specific experimental contexts [1,2,3,4,5,6,7]. Among these, the evolutionary design approach has shown to be very successful in a large set of experimental studies [8,9,10,11,12,13,14], in particular when the experimentation is conducted to search for an optimal value, or region of optimality, in large experimental spaces. This approach has been developed according to the principle of natural evolution: an initial small set of design points is selected and sequentially evaluated and transformed in new generations of points moving in different and frequently unexpected areas of the search space. Evolution has been formulated in several different forms that include: genetic algorithms [15,14], particle swarm optimization [16,17,18], and ant colony optimization [19,20]. The resulting designs exhibit good performance, but frequently the stochastic component that drives the evolution can slow down the process of convergence to the target.

To deal with this problem, recent studies have been developed on evolutionary experimental design, with the key idea that the evolution can be driven by sequential probabilistic models able to identify the relevant variables for representing the system and inferring the dependence relations among them. The design becomes model-based, and the probabilistic inferences can effectively drive the search towards the target by testing a very small set of compositions [10]. We introduce a basic structure of this design in EBN-design [12], where a Bayesian network model is built and estimated on both the design variables and the response variable in order to guide the evolutionary process. The structure of the network emerges from the observed dependence relations (arcs) among variables (nodes), and the strength of the dependence is measured by probability distributions [21,22]. A similar approach is the Estimation of Distribution Algorithms (EDAs), in which estimated BNs enter as part of the evolutionary process [23,24,25]. These procedures use probabilistic models to estimate the relations among variables and computational operators to derive succeeding generations of solutions; the explicit use of the response variable as part of the model, as in EBN-design, is however not part of this class of procedures. EDAs are successfully used in many research areas, even if they require a large number of generations to reach the optimal solutions, which can be sometimes quite problematic.

* Corresponding author at: Department of Environmental Science, Informatics and Statistics, University Ca' Foscari, Canareggio 873, 30121 Venice, IT. Tel.: +39 3469707575.
E-mail addresses: debora.slanzi@unive.it (D. Slanzi), d.delucrezia@explora-biotech.com (D. De Lucrezia), irenpoli@unive.it (I. Poli).

http://dx.doi.org/10.1016/j.chemolab.2015.09.016
0169-7439/

A powerful feature of Bayesian networks consists in providing probability distributions of subsets of variables given the evidence on some other identified variables, which are called evidence variables [26,27]. These conditional probability distributions can play a key role in deriving efficient design strategies and successful search procedures.

In this paper we introduce the Querying Bayesian network design (QueBN-design), an enhancement of EBN-design, where the experimental strategy is sequentially achieved by conditioning probability distributions on the evidence of target values of the system response. The Bayesian network is sequentially queried to discover which are the best design variable combinations given high values of responses, and provided with this relevant information to construct succeeding generations of experiments.

We evaluate the performance of QueBN-design in a simulation study where we compare with other designs its ability to achieve the optimal value of two benchmark functions representing high dimensional and complex responses.

Given the successful performance of QueBN-design in simulation, we addressed the quite hard problem of engineering new synthetic proteins. Protein engineering is the design of new synthetic proteins with some desirable functionalities. The construction of synthetic proteins in several different fields involves the formulation of statistical designs on huge spaces of ordered amino-acid sequences [28,29,30]. Experimentation is generally conducted to find a sequence that folds into a desired structure and then accomplishes a particular biological function. In this context experimentation could involve testing an extremely large set of sequences and this might be technically infeasible or economically unsustainable. Several computational techniques in molecular biology, such as Rosetta design, have been developed [31,32,33] to design synthetic proteins without the exploration of the entire sequence space. These techniques are commonly used, but are computationally very intensive and require the explicit computation of sequence minimum-energy, of all-atom refinement and/or of 3D data that are difficult to obtain. In this paper we choose to engineer proteins with similar functionalities to the natural 1AGY *serine esterase* protein from the fungus Fusarium solania [34,35]. We develop QueBN-design in two different experimental series and achieve very good results in both.

This paper is organised as follows: in Section 2, after a brief introduction on experimental design for optimisation and on Bayesian networks, we derive the Querying Bayesian network design for experimentation. In Section 3 we describe a simulation study to test the performance of QueBN-design and to compare it with several evolutionary and other alternative experimental designs. In Section 4 we build the QueBN-design to address the construction of synthetic proteins with similar functionalities to the natural 1AGY *serine esterase* protein. In Section 5 we provide some concluding remarks.

2. Methods

2.1. Introduction to the experimental design for optimisation

In designing experiments, we define a response variable Y and a set of design variables $\mathbf{X} = (X_1, X_2, \ldots, X_d)$. Assuming that each design variable can take a finite set of mutually exclusive values, the whole experimental space Ω is defined as the set of combinations of all possible design variable values (namely experimental points, design points or tests). We assume a relation among response and design variables in the form of

$$Y = g(\mathbf{X}) + \eta \tag{1}$$

where g is an unknown function representing the behaviour of the system and η is a stochastic error term with a particular probability distribution. Data are sampled from the experimental space according to a chosen design

$$\xi_n = \mathbf{X}_n = (\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_n)'$$

consisting of a set of n experimental points, where each point $\mathbf{x}_k = (x_{k,1} x_{k,2} \ldots x_{k,d}) \in \Omega$, $k = 1, \ldots, n$, is a d-dimensional vector, representing the observed values on the design variables. For each experimental point, we then derive the corresponding response value $y_k = g(\mathbf{x}_k)$. The set of data from experimentation, namely $(\mathbf{X}_n, \mathbf{y}_n)$ with $\mathbf{X}_n$ as an $(n \times d)$ -matrix and $\mathbf{y}_n$ as an n-vector, represents the information for inferring the dependence relations among variables and the estimation $\hat{g}$ of the function g, and for identifying the design point that gives the optimum value of the response variable, i.e. $\mathbf{x}^\star$ such that $g(\mathbf{x}^\star) \geq g(\mathbf{x})$ for all the possible experimental points $\mathbf{x} \in \Omega$ (in maximisation problem).

To select an efficient design ξ_n, several strategies have been proposed in the literature; most of them assume polynomial models to infer the form of g [36,37,38] or adopt surrogate models (emulators) to obtain response surface approximations [39,40,41]. However the high dimensionality of the experimental space, in terms of number of design variables and/or number of possible mutually exclusive values, makes these approaches very hard to use because they require a huge number of experimental points to estimate g or complex surrogate models prohibitive to simulate.

We address this problem adopting the evolutionary design, where just a small set of experimental points are considered. In this approach, the design is evolved across generations according to a particular function measuring the goodness of the design in reaching the objective of the optimisation. To guide the evolution, we propose to combine the rules of evolutionary paradigm with the information achieved by Bayesian network models. Bayesian networks can in fact represent the information in a system by means of dependence and independence relations, finding a minimal structure which explains the joint action of system variables in affecting the system response. The Bayesian networks estimated in each generation of the evolutionary procedure can be seen as source of information to better explain and understand of the underlying structure of the problem and to enhance the performance of the optimisation procedure.

2.2. Bayesian networks

Bayesian networks (BNs) are graphical models for reasoning under uncertainty [26,42,21,22].

Formally, a BN is represented by a directed acyclic graph (DAG), composed of nodes and arcs, and a probability distribution (P). DAG represents the structure of the BN model: nodes are random variables $\mathbf{X} = (X_1, X_2, \ldots, X_d)$, each of which can take a value in a finite set of possible mutually exclusive variable values, and arcs between nodes, in the form of $X_i \rightarrow X_j$, indicate direct probabilistic dependencies between the corresponding variables. In the BN, variables take roles of parent and child according to the dependence relation that links the corresponding nodes: X_j is child of X_i if X_j directly depends on X_i. The absence of an arc between two nodes involves an independence relationship between the variables given the value of any intermediate node. The correspondence between the structure of the DAG and the conditional independence relationships is derived by the d-separation criterion [26]. The Markov property then follows from d-separation: each variable is probabilistically independent of all its non descendant given its parents. From the Markov property, the joint probability distribution P can be written as follows:

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{d} P(X_i = x_i | Pa(x_i)) \tag{2}$$

where $\mathbf{X} = \mathbf{x}$ indicates that the set of variables $\mathbf{X} = (X_1, X_2, \ldots X_d)$ is observed at specific values $\mathbf{x} = (x_1, x_2, \ldots x_d)$, and $Pa(x_i)$ is regarded as the particular value realisation of the parent set of X_i. The set of