

A predictive evolutionary approach to design biochemical experiments

Un approccio predittivo e evolutivo al disegno sperimentale

Michele Forlin¹

Dipartimento di Statistica, Università Ca' Foscari, e-mail: mforlin@unive.it

Davide De March

European Center for Living Technology, e-mail: davidedemarch@unive.it

Debora Slanzi

Dipartimento di Statistica, Università Ca' Foscari, e-mail: debora@unive.it

Irene Poli

Dipartimento di Statistica, Università Ca' Foscari,
e European Center for Living Technology, e-mail: irenpoli@unive.it

Riassunto: La costruzione di un sistema previsionale in grado di valutare lo spazio non noto di esperimenti in ambito biochimico viene affrontato in questo lavoro con il calcolo evolutivo. La scelta del disegno sperimentale, e perciò dei punti che lo compongono, viene eseguita secondo il paradigma dell'evoluzione e derivata da algoritmi genetici. Questo approccio viene proposto e sviluppato per l'alta dimensionalità dello spazio di ricerca di tali esperimenti. L'informazione derivante dai modelli, che vengono scelti in base al grado di accuratezza delle previsioni, partecipa poi all'evoluzione del disegno, consentendo la formulazione di un sistema integrato.

Keywords: evolutionary predictive models, experimental designs.

Getting insights into the dynamical processes of biological organizations, in terms of relevant molecular components and their interactions, is essential in many areas of research of nanobiotechnology. Studying these processes we frequently encounter problems characterized by high dimensionality and poor knowledge or intuition on the number and the role of the components of the systems. It is well known that high dimensionality may represent a serious obstacle in conducting any reliable analysis in designing the experiments and formulating predictive models (Cawse, 2003). In this paper we introduce an evolutionary approach in designing high dimensional experiments and achieve the design by building a genetic algorithm. This procedure searches in the experimental space the set of compositions (factors, levels, and protocols) with the best predictive accuracy. Briefly, the experiments consist in mixing a set of amphiphilic molecules under particular environment conditions and technology protocols, and observing a self-assembling process that may originate chemical entities such as vesicles or micelles. The response of the system is a stochastic measure of the formation of these entities, and the experimental factors are the elements of the set of amphiphilic molecules which have to be chosen in a large set (~ 90) and in different levels of concentration. In the statistical literature these experiments are known as mixture experiments (Cornell, 2002). To select the design we

¹Address of correspondence: Dipartimento di Statistica, San Giobbe, Cannaregio 873, Venezia

adopt a genetic algorithm: a population of possible candidates of different mixtures is randomly generated and evolved in different generations through a set of genetic operators, e.g. stochastic selection, crossover, mutation, innovation. With this algorithm we build a population-based stochastic design, whereby each member of the populations, representing an experiment (e.g. a mixture), is coded as a genetic sequence, and evaluated in terms of the system response and in terms of its predictive capacity. More specifically, in a first step of the analysis, we limit the number of candidates amphiphiles to be considered in the experimentation to 16 (selected according to theoretical knowledge of the molecule structure). We proposed and evaluated 540 experiments in 5 different generations (180 different mixtures with 3 replications) where each mixture could contain at most 5 different amphiphilic molecules with a total volume of 100ul (Theis et al. 2006). This setting of the experiments is known in the literature as a $\{16, 5\}$ Simplex Lattice Design and the full exploration of the space would require 15504 experiments. The results of this approach exhibit a very good performance of the genetic algorithm: we achieve a set of experiments with a much higher response with respect to a previous set of results obtained without the application of the genetic algorithm. Modelling these experimental results with a polynomial structure, which involves main effects, binary and ternary interactions, we achieve the following formulation

$$T_j = \sum_{i=1}^{16} \beta_i x_{ij} + \sum_{i=1}^{16} \sum_{k, i < k} \beta_{ik} x_{ij} x_{kj} + \sum_{i=1}^{15} \sum_{k, i < k} \beta_{ik2} x_{ij} x_{kj} x_{2j} + \\ \sum_{i=1}^{15} \sum_{k, i < k} \beta_{ik9} x_{ij} x_{kj} x_{9j} + \sum_{i=1}^{15} \sum_{k, i < k} \beta_{ik13} x_{ij} x_{kj} x_{13j} + \epsilon_j$$

where x_{ij} represent the molecule x with the level i in the experiments j , and T_j represent the response of the j -experiment. From this estimated model, with high predictive power (Predictive Mean Square Error = 0,01), we learn which particular interactions of molecules and with which levels we could reach high level response. Based on these results we also conducted simulated studies to evaluate the performance of a model-based algorithm, i.e. a genetic algorithm driven in the evolution by the model information. The results show a much better performance of the algorithm, and a much faster research in the experimental space.

References

- Cawse J.N. (2003) *Experimental Design for Combinatorial and High Throughput Materials Development*. John Wiley Sons.
- Cornell J.A. (2002) *Experiments with Mixtures: Designs, Models and the Analysis of Mixture Data*. John Wiley Sons.
- Theis M., Gazzola G., Forlin M., Poli I., Hanczyc M., Bedau M. (2006) Optimal Formulation of Complex Chemical Systems with a Genetic Algorithm. *ComPlexUs*, in press.