Efficient Discovery of Chromatography Equipment Sizing Strategies for Antibody Purification Processes Using Evolutionary Computing

Richard Allmendinger, Ana S. Simaria, and Suzanne S. Farid

University College London, Torrington Place, London WC1E 7JE, UK
{r.allmendinger,a.simaria,s.farid}@ucl.ac.uk

Abstract. This paper considers a real-world optimization problem involving the discovery of cost-effective equipment sizing strategies for the chromatography technique employed to purify biopharmaceuticals. Tackling this problem requires solving a combinatorial optimization problem subject to multiple constraints, uncertain parameters (and thus noise), and time-consuming fitness evaluations. After introducing this problem, an industrially-relevant case study is used to demonstrate that evolutionary algorithms perform best when infeasible solutions are repaired intelligently, the population size is set appropriately, and elitism is combined with a low number of Monte Carlo trials (needed to account for uncertainty). Adopting this setup turns out to be more important for scenarios where less time is available for the purification process.

1 Introduction

Monoclonal antibodies (mAbs) represent the fastest growing category of therapeutic biopharmaceutical drugs due to their unique binding specificity to targets. The manufacturing process for mAbs is costly and time-consuming, and can be divided into two phases (see Fig. 1): *upstream processing* (USP) and *downstream processing* (DSP). In USP, mammalian cells expressing the mAb of interest are cultured in bioreactors. Then the broth moves to DSP, where the mAb is recovered, purified and cleared from viruses using a variety of operations including a number of chromatography steps. Chromatography operations are identified as critical steps in a mAb purification process and can represent a significant proportion of the purification material costs (associated e.g. with the use of expensive affinity resins and large amounts of buffer reagents). Whilst alternatives to traditional column chromatography platforms are emerging, industry practitioners are still reluctant to perform major process changes [1]. At the same time, it is important to determine how best to use existing production facilities for mAbs [2]. This is particularly challenging given the significant improvements in USP productivities that have been accomplished over the past decade with higher mAb concentrations (*titres*) being achieved in cell culture. These improvements have not been matched in purification capacities, leading to concerns over purification bottlenecks and the desire to continuously optimize the design and operation of existing chromatography steps. Hence, to efficiently exploit these cell culture improvements, and account for the increasing demand for therapeutic mAbs, it has become critical to identify cost-effective purification processes [1].

C.A. Coello Coello et al. (Eds.): PPSN 2012, Part II, LNCS 7492, pp. 468–477, 2012.

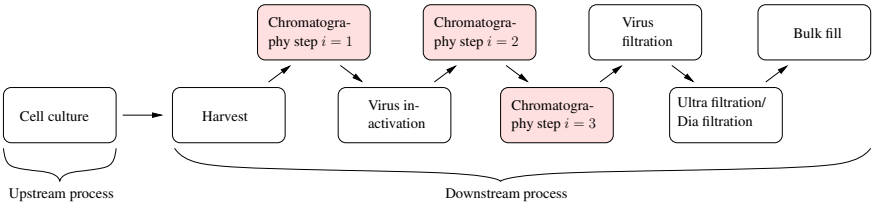


Fig. 1. Typical flowsheet for an antibody manufacturing process

An approach to realize this identification step, which is also adopted here, is to develop simulation models of mAb manufacturing processes and identify promising chromatography setups using computational methods. For example, in [3] the authors present a simulation model to identify windows of operation for the column diameter, bed height and loading flowrate of a chromatography step using productivity and cost of goods (COGs) as performance criteria. A model to find combinations of protein load and loading flowrate that meet yield and throughput constraints has been developed in [4]. The discrete-event simulation framework proposed in [5] allows the selection of optimal chromatography column diameters over a range of titres. The methodology used in [3–5] consists of selecting and evaluating specific values within the full range of variation of the critical parameters. However, such an approach may not be feasible for very large decision spaces as considered here, which drives the need for more efficient optimization methods in this domain.

This study addresses this issue by investigating the application of evolutionary optimization methods for the discovery of *chromatography column sizing strategies* — defined here by the diameter and bed height of a column, the number of columns used in parallel, and the number of cycles a column is run for — that are cost-effective in terms of COGs per gram (COG/g) of product manufactured. This discovery task can be formulated as a combinatorial (single-objective) optimization problem subject to multiple constraints and interacting decision variables, uncertain parameters and expensive fitness evaluations (represented by time-consuming computer simulations). Over the years, evolutionary algorithms (EAs) have proven to be efficient, flexible and robust optimizers for challenging optimization problems of this type — which are commonly referred to as *closed-loop optimization problems* [6, 7].

An industrially-relevant case study is used to investigate how to tune some of the simple EA configuration parameters: population size, degree of elitism, number of Monte Carlo trials (needed to cope with uncertain parameters), and constraint-handling method. The fitness landscape of different scenarios of the case study are analyzed also to observe which landscape features pose a particular challenge when optimizing equipment sizing strategies.

The rest of the paper is organized as follows. The next section describes the chromatography sizing problem considered in this work in more detail. Section 3 outlines the case study, choice of algorithms and the parameter settings considered for tackling the case study. The experimental results are presented and analyzed in Section 4, and Section 5 concludes the paper.

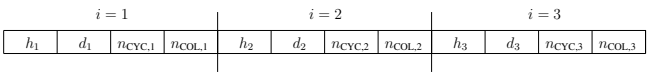


Fig. 2. A candidate solution (sizing strategy) with $k = 3$ chromatography steps. Each step $i = 1, ..., k$ is defined by the bed height h_i and diameter d_i of columns, number of cycles $n_{\text{CYC},i}$ each column is used, and the number of columns $n_{\text{COL},i}$ operating in parallel.

2 Problem Domain: Chromatography Equipment Sizing

The *chromatography equipment sizing problem* can be represented as a combinatorial optimization problem with the task of finding the most cost-effective chromatography sizing setup for a sequence of chromatography steps used in the purification process of mAbs. In the following the decision variables, objective function, constraints, and uncertain parameters to which this problem is subject to are described.

Decision Variables: Fig. 2 shows the encoding used to represent a solution $\boldsymbol{x}$ to the chromatography sizing problem. For each chromatography step $1 \leq i \leq k$ (k is the total number of steps) (e.g. affinity or ion-exchange chromatography) four discrete decision variables were defined related to the sizing and operation of chromatography columns: bed height h_i and diameter d_i of columns, number of cycles $n_{\text{CYC},i}$ each column is used, and the number of columns $n_{\text{COL},i}$ operating in parallel. That is, the problem is subject to $l = k \cdot 4$ discrete variables in total. For each step i, the variables define the (i) total volume of resin V_i available for the purification of a product at that chromatography step, and the (ii) processing time T_i that the chromatography step takes; both parameters are calculated according to standard mass balance equations as follows [8]:

$$V_i = \pi \cdot d_i^2/4 \cdot h_i \cdot n_{\text{CYC},i} \cdot n_{\text{COL},i} \tag{1}$$

$$T_i = n_{\text{CYC},i} \cdot h_i \cdot (CV_{\text{BUFF},i} + CV_{\text{LOAD},i}/n_{\text{COL},i}) \cdot u_i, \tag{2}$$

where $CV_{\text{BUFF},i}$ and $CV_{\text{LOAD},i}$ are the number of column volumes of buffer and product load per cycle, and u_i is the linear flowrate of the resin used at step i.

Objective Function: Our objective f is to find a chromatography sizing setup that yields minimal *cost of goods per gram (COG/g) of product manufactured*. The COGs include both direct (resource) costs (e.g. resin, buffer and labor costs) and indirect costs (e.g. facility-dependent overheads, such as maintenance costs and depreciation), and is divided by the total annual product output P to yield the metric COG/g. The COG/g values are obtained by running a detailed process economics model, which simulates the different purification steps based on mass balance and cost equations as defined in [8].

Constraints: The problem is subject to two types of constraints:

1. Each chromatography step $i = 1, ..., k$ needs to satisfy a *resin requirement constraint* to ensure that the resin volume V_i available for purification at step i is sufficient to process the mass of product M_i entering that step, given the resin's dynamic binding capacity DBC_i and the maximum utilization factor κ. Formally, this constraint can be defined as

$$V_i \geq \frac{M_i \cdot \kappa}{DBC_i} \quad . \tag{3}$$

Solutions violating this constraint *are considered infeasible* and handled using one of the constraint-handling strategies introduced in Section 3.

2. There is also a *demand constraint* to ensure that the amount of product manufactured P is sufficient to satisfy the annual demand D, or $P \geq D$. This constraint may be violated for column sizing strategies with long chromatography processing times T_i. The use of COG/g as the objective function (recall that the product output P is in the denominator of this metric) was found to be sufficient to cope with this constraint. Hence, if a solution violates the demand constraint, then it is *not considered infeasible*.

Uncertainties: Uncertainty related to the product titre can have a significant impact on the annual product output P. As the equipment sizing is a function of an expected titre value for bioreactors through to chromatography columns, titre fluctuations can cause (i) failure to meet demand (if titre is lower than expected) or (ii) product waste (if titre is higher than expected and equipment capacity is insufficient to process the excess). Other sources of uncertainty (e.g. yield) may be present and are realistic but are not considered in this paper.

3 Experimental Setup

This section describes the case study, search algorithms and their parameter settings as used in the subsequent experimental analysis.

Case Study Setup: The case study considered in this work is industrially-relevant and focuses on a single-product mAb manufacturing facility that employs a process sequence as shown in Fig. 1 (with $k = 3$ chromatography steps) to satisfy a total product demand of $D = 500$kg/year with an expected titre of 3g/L. Titre variabilities were modeled using the triangular probability distribution, Tr(2.6,3.0,3.4). Three scenarios of this case study with different ratios of USP:DSP trains were investigated: 1:1, 2:1 and 4:1. The USP train refers to the number of bioreactors operating (in a staggered mode), and an increase in the USP:DSP ratio corresponds to a decrease in the *DSP window*, the time available to perform chromatography. The range of possible decision variable values is 15 cm $\leq h_i \leq$ 25 cm (11 values), 50 cm $\leq d_i \leq$ 200 cm (16 values), $1 \leq n_{\text{CYC},i} \leq 10$ (10 values), $1 \leq n_{\text{COL},i} \leq 4$ (4 values), $i = 1, 2, 3$; i.e. there are $(11 \cdot 16 \cdot 10 \cdot 4)^3 \approx 3.5 \cdot 10^{11}$ sizing strategies in total. The sizing strategy employed in industry is obtained based on empirical rules: a single column $n_{\text{COL},i} = 1$ with a fixed bed height of $h_i = 20$ cm is run for a fixed number of cycles $n_{\text{CYC},i} = 5$ with the diameter size d_i being calculated such that the resulting total resin volume V_i (Equation (1)) satisfies the resin requirement constraint (Equation (3)).

Search Algorithms: To gain insight into the behavior of evolutionary search algorithms on the chromatography sizing problem, four types of search algorithms were considered: a standard generational genetic algorithm (SGA), a genetic algorithm with generation gap (GA-GG), a genetic algorithm with a $(\mu + \lambda)$-ES reproduction scheme

(GA-ES), and a population of stochastic hill-climbers (PHC). All four algorithms began the search with the same initial population containing μ randomly generated solutions. The algorithms used also the same mutation operator, which selected a decision variable value at random from the set of possible values. SGA used uniform crossover and random flip mutation as the variation operators, and binary tournament selection (with replacement) for parental selection; for environmental selection, it replaced the entire current population with the offspring population. GA-GG and GA-ES differ from SGA in the environmental selection step only. With GA-GG, the new population was formed by selecting the fittest μ solutions from the combined pool of the offspring population and the two fittest solutions of the current population. With GA-ES, a greater degree of elitism was employed and the fittest μ solutions from the combined pool of the current population and the offspring population were selected. PHC maintained a population of stochastic hill-climbers, which, at each generation g, independently underwent mutation and replaced their parent if it was at least as fit.

Accounting of Uncertainty: To account for titre variabilities, m Monte Carlo trials (based on the probability distribution Tr(2.6,3.0,3.4)) were performed for each candidate solution. The fitness of a solution was then the average of the COG/g values across the m trials, and this average was updated if a solution happened to be evaluated multiple times during an optimization procedure.

Handling of Infeasible Solutions: Five constraint-handling strategies were analyzed to cope with infeasible solutions (violating Equation(3)). Four of them (RS1, RS2, RS3 and RS4) *repaired* infeasible solutions, i.e. modified the genotype of a solution, while strategy RS5 avoided repairing.

The four *repairing strategies* iteratively increased the values of the decision variables (associated with a particular chromatography step i), one variable at a time, until Equation (3) was satisfied or until the maximum value of a variable was reached, in which case the value of another variable was increased. The sequence in which the variables were modified affected the search. To investigate this effect, different sequences, represented by the strategies RS1 to RS4, were analyzed. The strategy RS1 applied repairing according to the decision variable sequence $d_i \rightarrow n_{\text{CYC},i} \rightarrow h_i \rightarrow n_{\text{COL},i}$ (where i is the chromatography step violating Equation (3)); this sequence represents typical rules applied in equipment sizing scale-up models. The strategy RS2 employed the inverse sequence of RS1. The strategies RS3 and RS4 switch between different repairing sequences during an optimization procedure. While RS3 chooses at random between the two sequences employed by RS1 and RS2, the strategy RS4 chooses at random among all possible repairing sequences (note, there are 4! sequences in total) whenever it needs to be repaired. The approach employed by RS4 is plausible e.g. if no prior knowledge about promising repairing sequences would be available. The strategy RS5 does not apply repairing but penalizes infeasible solutions by degrading their fitness by a large penalty value c.

The experimental study investigated different settings of the parameters involved in the search algorithms. The default settings used are given in Table 1. Any results shown are average results across 20 independent algorithm runs. A different seed was used for the random number generator for each EA run but the same seeds for all strategies. This allows for the application of a repeated-measures statistical test, the Friedman test, to investigate performance differences between algorithmic setups.

Table 1. Default parameter settings of search algorithms

Parameter	*Setting*
Parent population size μ	80
Offspring population size λ	80
Per-variable mutation probability	$1/l$
Crossover probability	0.6
Constraint-handling strategy	RS1
Number of generations G	25
Penalty value c	5000
Monte Carlo trials m	25

4 Experimental Results

Before analyzing the behavior of evolutionary search algorithms on the chromatography equipment sizing problem, an indication of the properties of the fitness landscapes spanned by three case study scenarios is given. For this, the *adaptive walks method* already used in [7] was adopted. This involved performing 1000 adaptive walks (using a fixed titre of 3g/L) on the landscape of each scenario, and recording the length and final fitness of each walk. Figure 3 shows the distribution of both measurements in the form of boxplots. From Figure 3(a) it can be observed that increasing the USP:DSP ratio decreases the average length of an adaptive walk. That is, the landscape becomes more rugged, or, equivalently, the number of local optima increases. This pattern is due to tighter DSP windows, which cause more solutions to violate the demand constraint and thus makes the problem harder to solve. This also causes an increase in the COG/g values as indicated in Figure 3(b). The next section presents an analysis of how the search algorithms fared, for both the deterministic (using a fixed titre of 3g/L) and stochastic scenario.

Deterministic Product Titre: Figure 4(a) analyzes the performance of the search algorithms as a function of the population size μ. The aim of this experiment was to understand whether a large population should be evolved for few generations, or a small population for many generations. This understanding is important when optimizing subject to limited resources, such as limited computational power and time constraints. The figure illustrates that: (i) a population size of around $40 \leq \mu \leq 80$ yielded the best performance for the GA-based algorithms, (ii) GA-ES found the most cost-effective strategies, and (iii) random search outperforms PHC. Small population sizes, or search algorithms employing no elitism, such as SGA, did not perform well due to the high probability of getting trapped in one of the many local optima of the fitness landscape. Large population sizes converged slowly due to the low number of generations available for optimization. PHC was inferior to random search because the hill-climbers could get trapped in local optima, in which case further improvements were unlikely, while random search kept on generating (at random) new and potentially fitter solutions. (The performance of random search is constant for varying μ as it depends only on the number of function evaluations performed.)

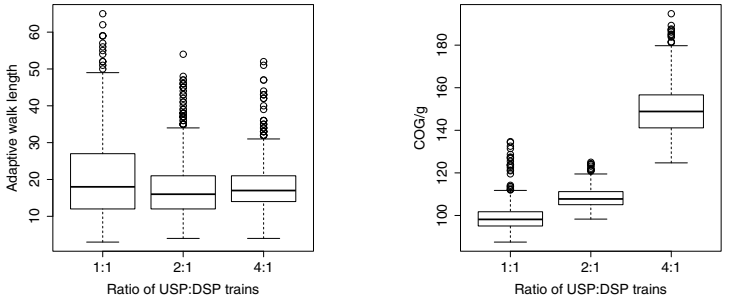
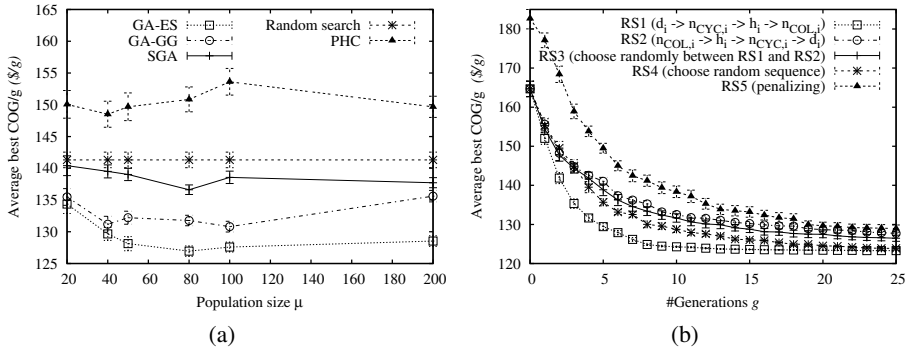


Fig. 3. Boxplots showing the distribution of the (a) length and (b) final fitness (COG/g) of 1000 adaptive walks for different USP:DSP ratios. The box represents the 25th and 75th percentile with the median indicated by the dark horizontal lines. The whiskers represent the observations with the lowest and highest value still within $1.5 \cdot$ IQR of the 25th and 75th percentile, respectively; solutions outside this range are indicated as dots.

Figure 4(b) investigates the performance impact of the constraint-handling strategies RS1 to RS5 when augmented on GA-ES (a similar performance impact was present for the other search algorithms). It demonstrated that the constraint-handling strategy employed had an effect on the convergence speed and the final solution quality. It also indicated that a repairing strategy (RS1, RS2, RS3 and RS4) should be preferred over a non-repairing one (RS5). The superior performance of RS1 is due to the fact that the variable d_i is modified (increased) first to repair a solution. Unlike to the other variables, an increase in d_i is often sufficient to just satisfy the resin requirement constraint without increasing the processing time. From the performance obtained with RS2, RS3 and RS4 it can be concluded that if d_i cannot be changed, then either the variable $n_{\mathrm{CYC},i}$ or h_i should be modified to meet the resin requirement constraint.

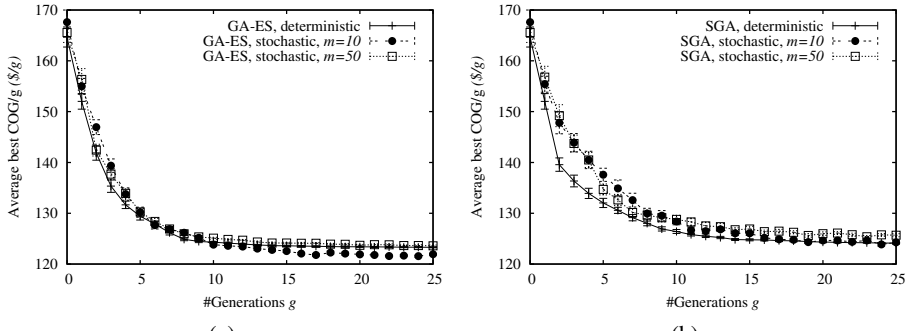
Stochastic Product Titre: The performance of the algorithms was then investigated in the presence of uncertain product titres. Figure 5 indicates that uncertainty impacts negatively the convergence speed and under certain circumstances also the final solution quality. This impact tends to be less severe as the degree of elitism employed by an algorithm increases (i.e. the performance of GA-ES is less affected than the one of SGA). Elitism can help circumventing this issue as it causes a population to converge (quickly) to a (local) optimal region and then exploit this region. However, on the other hand, too much elitism (Figure 5(a)) may disturb and prevent the discovery of innovative solutions; here, optimization in a stochastic environment using relatively small values of m can yield better performance than optimization in a deterministic environment due to the greater randomness in the search. When the optimizer does not employ elitism (Figure 5(b)), however, any additional randomness in the search may be a burden (as it can cause a population to oscillate between different search space regions, preventing or slowing down convergence towards promising regions).

Figure 6 shows the sizing strategies for the most expensive chromatography step ($i = 1$) found by GA-ES for the USP:DSP ratios 1:1 (Figure 6(a)) and 4:1 (Figure 6(b)) at the end of the search across 20 independent algorithmic runs. For both scenarios,



(a) (b)

Fig. 4. (a) Average best COG/g (and its standard error) obtained by different search algorithms as a function of the population size μ; the total number of fitness evaluations was fixed to 2000, i.e. the number of generations is $G = \lfloor 2000/\mu \rfloor$. (b) Average best COG/g, as a function of the generation counter g, obtained by GA-ES using different repairing strategies. Both experiments were conducted on a chromatography equipment sizing problem featuring a ratio of USP:DSP trains of 4:1. For each setting shown on the abscissa, a Friedman test (significance level of 5%) has been carried out: In (a), GA-ES performs best for $\mu > 40$, and in (b), RS1 performs best in the range $1 < g < 15$.



(a) (b)

Fig. 5. Average best COG/g (and its standard error) obtained by (a) GA-ES and (b) SGA in a deterministic and stochastic environment (using different values for the number of Monte Carlo trials m) as a function of the generation counter g. For each setting shown on the abscissa, a Friedman test (significance level of 5%) has been carried out: In (a), GA-ES with $m = 10$ performs best for $g > 15$, and in (b), SGA, deterministic, performs best in the range $1 < g < 6$.

the solutions shown have COG/g values that do not differ by more than 3% of each other. Comparing the most cost-effective sizing strategy found by GA-ES (filled bubble) with the strategy used in industry (filled diamond), GA-ES is able to reduce the COG/g for 1USP:1DSP and 4USP:1DSP by up to 5% (mainly through sizing strategies featuring smaller h_1 and/or d_1 in combination with more cycles $n_{\mathrm{CYC},1}$) and 20% (through sizing strategies exhibiting fewer cycles $n_{\mathrm{CYC},1}$ and larger d_1), respectively. Another advantage of EAs is that the result of an optimization procedure is a set of

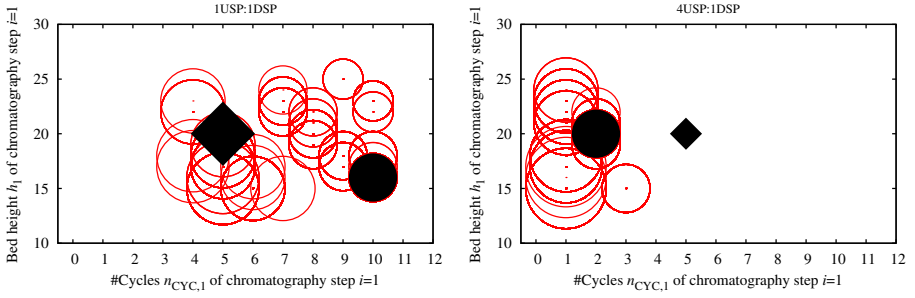


Fig. 6. Column sizing strategies for the most expensive chromatography step ($i = 1$) found by GA-ES at the end of the search across 20 independent algorithm runs (within an uncertain optimization environment) (bubbles) for the scenarios (a) 1USP:1DSP and (b) 4USP:1DSP. The size of a bubble is proportional to the variable d_1; all solutions feature the setup $n_{COL,1} = 1$. The fitness values of all solutions found by the EA for a particular scenario are within 3% of each other. For each scenario, the filled bubble represents the optimal setup found by the EA. The setup used by industry is indicated with a filled diamond and was not part of the solution set found by the EA.

cost-efficient sizing strategies (rather than a single strategy), providing flexibility and freedom to account for facility space restrictions and user preferences when it comes to selecting a final sizing strategy. Note, the EA finds more similar solutions for the scenario 4USP:1DSP than for 1USP:1DSP because the problem is harder to solve, as already indicated in the landscape analysis conducted previously.

5 Conclusion and Future Work

This paper has considered a real-world problem concerned with the discovery of cost-effective equipment sizing strategies for purification processes (with focus on chromatography steps) of biopharmaceuticals. This application can be formulated as a combinatorial closed-loop optimization problem subject to (i) expensive fitness evaluations, (ii) multiple dependent decision variables, (iii) constraints, and (iv) uncertain parameters.

The study revealed that EAs can identify a diverse set of equipment sizing strategies that are more cost-efficient than the strategies used in industry. In particular, the analysis demonstrated that an EA performs best when elitism is used in combination with a small number of Monte Carlo trials (to cope with uncertain parameters), infeasible solutions are repaired using a non-trivial strategy, and (when resources are limited) a medium-sized population (a size between $30 \leq \mu \leq 80$) is evolved for a relatively large number of generations.

Future research will look at extending the equipment sizing problem considered here with decision variables related to the sequence of a purification process employed. This will make the optimization tool developed more versatile, and also help gain more insights into the working of EAs.

Acknowledgement. Financial support from the EPSRC Centre for Innovative Manufacturing in Emergent Macromolecular Therapies with a consortium of industrial and government users is gratefully acknowledged.

References

1. Langer, E.: Downstream factors that will continue to constrain manufacturing through 2013. BioProcessing Journal 8(4), 22–26 (2009)
2. Kelley, B.: Industrialization of mAb production technology: The bioprocessing industry at a crossroads. MAbs 1(5), 443–452 (2009)
3. Joseph, J.R., Sinclair, A., Tichener-Hooker, N.J., Zhou, Y.: A framework for assessing the solutions in chromatographic process design and operation for large-scale manufacture. Journal of Chemical Technology and Biotechnology 81, 1009–1020 (2006)
4. Chhatre, S., Thillaivinayagalingam, P., Francis, R., Titchener-Hooker, N.J., Newcombe, A., Keshavarz-Moore, E.: Decision-Support Software for the Industrial-Scale Chromatographic Purification of Antibodies. Biotechnology Progress 23, 88–894 (2007)
5. Stonier, A., Smith, M., Hutchinson, N., Farid, S.S.: Dynamic simulation framework for design of lean biopharmaceutical manufacturing operations. Computer Aided Chemical Engineering 26, 1069–1073 (2009)
6. Knowles, J.: Closed-Loop Evolutionary Multiobjective Optimization. IEEE Computational Intelligence Magazine 4(3), 77–91 (2009)
7. Allmendinger, R.: Tuning Evolutionary Search for Closed-Loop Optimization. PhD thesis, University of Manchester, Manchester, UK (2012)
8. Farid, S.S., Washbrook, J., Titchener-Hooker, N.J.: Modelling biopharmaceutical manufacture: Design and implementation of SimBiopharma. Computers & Chemical Engineering 31, 1141–1158 (2007)