Optimization in Molecular Design and Bioinformatics

Costas D. Maranas[a]

[a]Department of Chemical Engineering, The Pennsylvania State University

This work is an exposition on the application of optimization tools to problems in molecular design and bioinformatics. The specific areas addressed by the author include the design of polymers, surfactants, refrigerants, and enzymes. The goal is to systematically design molecules for the given application with desired performance characteristics. The performance measures of interest in polymer design are mechanical, electrical and thermophysical properties. In case of surfactants properties such as the HLB, emulsivity, detergency, and foaming stability influence the performance significantly. The performance measure in refrigerant selection and cycle synthesis is the balance between operating costs related to energy input and the investment costs. The performance measure in enzyme design is the probability of achieving a given nucleotide sequence target. The role of optimization is to "systematically" search through the alternatives. The research results in each of the applications mentioned above are presented.

Introduction

The competitive edge and market share of many chemical industries manufacturing polymers, refrigerants, solvents, surfactants, enzymes, and biomaterials are ultimately intertwined with the identification of "new" and "better" products. Though the vast number of alternatives presents a designer with an opportunity to find a better product, it also poses the challenge of systematically searching through the alternatives. With the rapid growth in optimization theory, algorithm development and high-performance computing, exciting and unprecedented research opportunities are emerging in molecular design to assist in this endeavor. Research results in polymer design, surfactant design, refrigerant selection and enzyme design are discussed in this work.

Previous work include the computer–aided design of molecular products such as polymers [9,7,5], solvents [5] and refrigerants [4,5] to name a few. The employed search algorithms include enumeration techniques, knowledge–based strategies, genetic algorithms and mathematical programming based methods. A comprehensive review of prior work can be found in Camarda and Maranas [3].

The objective is to find a molecule for a given application which optimally satisfies the desired performance targets.

Polymer Design

In polymer design the problem of identifying the polymer repeat unit architecture so that a performance objective that is a function of mechanical, electrical and/or physicochemical properties is addressed. Since the molecular design problem is posed within an optimization framework, a quantitative representation of the molecule and a quantitative structure-property relation is required. Group contribution methods (GCM) provide popular, versatile and relatively accurate ways for estimating properties based on the number and type of molecular groups participating in a molecule or repeat unit. (GCM) are based on the additivity principle of the groups constituting the molecule under investigation and have been extensively utilized in the estimation of a wide spectrum of polymeric properties including volumetric, calorimetric, thermophysical, optical, electromagnetic and mechanical properties. An extensive compilation of these estimation methods along with the corresponding parameters can be found in van Krevelen [11]. The use of (GCM) makes adequate the molecular representation using $\mathbf{n}=(n_1, n_2, \ldots, n_N)$ where n_i are the number of groups of type i present in the molecule. The problem of identifying the best molecule based on some measure of performance can be expressed as the following mixed–integer nonlinear optimization problem.

$$\begin{aligned} \min \quad & \mathcal{MP}\left(p_j(\mathbf{n})\right) && \textbf{(OMD)} \\ \text{subject to} \quad & p_j^L \;\le\; p_j(\mathbf{n}) \;\le\; p_j^U \\ & n_i \;\in\; \left\{n_i^L, n_i^L+1, \ldots, n_i^U\right\}, \quad i=1,\ldots,N \end{aligned}$$

The following two most widely used measures of performance are considered in this study [7]:

(1) Minimization of the maximum scaled deviation of properties from some target values (*property matching* **(PM)**),

$$\min \; \mathcal{MP} = \max_j \; \frac{1}{p_j^s} \left| p_j(\mathbf{n}) - p_j^0 \right|$$

where p_j^o is the target for property j and p_j^s the corresponding scale. (2) Minimization/maximization of a single property j^* (*property optimization* **(PO)**),

$$\min \,/\, \max \; \mathcal{MP} = p_{j^*}(\mathbf{n}).$$

To maintain structural feasibility of the molecule a number of linear constraints on $\mathbf{n}$ must be included in the problem **(OMD)**. These structural feasibility constraints define the necessary conditions under which a set of molecular groups can be interconnected so that there is no shortage or excess of free attachments. The estimation of most properties pertinent to engineering design is given by the ratio of two linear expressions in n_i. Though the above formulation is a mixed integer nonlinear program (MINLP) in general, the underlying mathematical functionalities of the above property estimation model are utilized to reformulate and solve the problem as a mixed integer linear program (MILP).

One of the limitations of group contribution estimation is that the internal molecular structure of the polymer repeat unit is only partially taken into account. For example, both polypropylene -$CH_2CH(CH_3)CH_2CH(CH_3)$- and head to head polypropylene -$CH_2CH(CH_3)CH(CH_3)CH_2$-, have the same molecular group representation. These shortcomings are alleviated with the use of property correlations involving topological indices as structural descriptors. These indices are numerical values which uniquely identify the polymer repeat unit and contain information about the atomic and electronic structure. Specifically, Bicerano [1] used the zeroth- and first-order molecular connectivity indices to correlate a wide range of polymer properties, including density, glass transition temperature, bulk modulus, and heat capacity. The functional form of the topological indices used are given in Camarda and Maranas [3]. The following additive property predictive form is utilized:

(Property Prediction) = (Basic Group Contribution) + (Connectivity Indices Contribution)

Though in general the above problem is a nonconvex MINLP, it is reformulated and solved as a convex MINLP utilizing the mathematical functionality of the connectivity indices.

So far it has been assumed that the properties are uniquely determined by the types of groups present in the molecule and their interconnectivity. However, in reality there are discrepancies between predicted and observed values. These can be reconciled by recognizing that the parameters of the property model vary around their nominal values. This can be expressed mathematically by utilizing probability distributions to describe the likelihood of different realizations for the model parameters. The probabilistic description of performance objectives and constraints is described in Maranas [6]. This formulation involves probability terms whose evaluation for each realization of the deterministic variables requires the integration of multivariate probability density distributions. This is accomplished without resorting to computationally intensive explicit or implicit multivariate integration. This is done by transforming the stochastic constraints into equivalent deterministic ones. Furthermore, it is shown that for probabilities of interest this formulation is a convex MINLP which can be solved to global optimality using commercial packages. The objective of using this formulation is to construct a trade-off curve between performance target and the probability of meeting the target. This aids the designer in choosing the optimal level of risk in selecting the molecule. Next, the surfactant

design problem is briefly discussed.

Surfactant Design

The design of surfactant solutions is an important problem in many industries since they are extensively utilized in diverse applications such as detergents, emulsifiers, and to ensure film coating and waterproofing. In the design of surfactant solutions the performance measures of interest are HLB, emulsivity, detergency, and foaming stability. Though this problem is also addressed within the general molecular design paradigm discussed previously, this problem presents additional unique challenges. The macroscopic properties of interest are related to structural descriptors of surfactants through fundamental solution properties such as critical micelle concentration (CMC) and area of a surfactant molecule within a micelle. Though this has the same flavor as relating property of polymers to connectivity of the molecule through topological indices there is an important difference. In polymer design, connectivity indices could be determined from the connectivity by simple evaluation. In the case of surfactants, determination of fundamental solution properties involves the minimization of free energy.

Therefore the problem of identifying the molecular structure of a surfactant with optimal values for the desired macroscopic properties is posed as a two-stage optimization problem [2]. The inner stage identifies the CMC and other micellar properties by minimizing the free energy μ_g, while the outer stage optimizes over the surfactant structural descriptors. A conceptual optimization formulation of the problem is as follows:

$$\max / \min \quad f(\text{macroscopic properties})$$

subject to

$$\begin{pmatrix} \text{macroscopic} \\ \text{properties} \end{pmatrix} = g(\text{fundamental properties})$$

$$\begin{pmatrix} \text{fundamental} \\ \text{properties} \end{pmatrix} = \begin{matrix} \arg\min \\ \text{descriptors} \end{matrix} \mu_g(\text{structural})$$

This formulation is solved using a truncated newton method. Since this problem may possess multiple local minima the problem is solved with multiple starting points. The structural descriptors include the number of carbon atoms in the surfactant tail n_c, the cross-sectional area of the head a_h, the charge separation for an ionic head group δ, and the dipole separation for dipolar surfactants d. These descriptors provide a concise description of the surfactant molecular topology and polarity. They are theoretically related to fundamental solution properties determining the shape, size and concentration of the surfactant micelles. The fundamental solution properties include the equilibrium area per molecule in a micelle a, the micellar shape, and the concentration at which micelles form (known as the critical micellar concentration or CMC). These properties are related through local regression models to macroscopic surfactant properties characterizing the suitability and effectiveness of the surfactant for a particular application (e.g., hydrophilic-lipophilic balance number (HLB)). Details of the functional relation of free energy to surfactant molecular structure and solution properties is given in Camarda *et.al.* [2]. This methodology is applied to identifying a nonionic surfactant with hydrophilic-lipophilic balance (HLB) of 13.8. HLB is a widely used measure of the emulsifying ability of a surfactant. High value for HLB implies high water solubility, and suitability for detergent or emulsifier. A local regression model is constructed which relates HLB to CMC as follows:

$$\ln HLB = 2.76 + 0.04 \ln CMC$$

The truncated-Newton algorithm was started from a number of initial starting points, and in each case, the algorithm converged to the same optimal solution involving a head cross-sectional area of 0.54977 nm and 5.997 carbons in a straight-chain tail. The CMC for this surfactant was found to be 0.034 mM. A search over tabulated surfactant properties reveals that a surfactant with a dimethyl phosphene oxide head group and a six carbon tail is compatible with those structural descriptors.

Refrigerant Selection and Cycle Synthesis

The focus now shifts from designing a molecule (refrigerant) to selecting a molecule from a prepostulated set of potential candidates. This still poses a challenge when placed within the context of synthesizing refrigeration cycles. The

combinatorial problem of appropriately assigning refrigerants to different locations in the refrigeration cycles requires the use of optimization tools. The problem addressed is stated as follows [10]: Given a set of process cooling loads, heat sinks at different temperatures and a set of available *pure* refrigerants, find the refrigeration cycle topology, operating conditions and refrigerants, selected from the list, that optimize a weighted sum of the investment and operating costs for the refrigeration system.

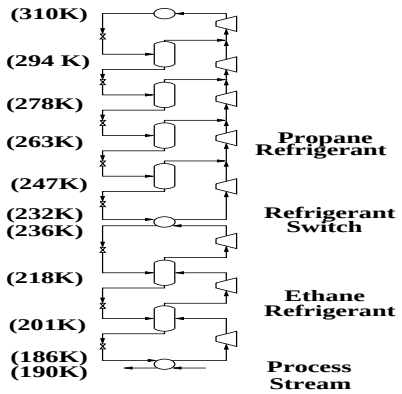

Figure 1: Vertical Cascade for pure refrigerant system

The proposed model involves a superstructure representation for both the synthesis and the refrigerant selection problems. The model allows for the identification of the number of stages, their operating temperature ranges, the type of refrigerant participating in a stage, the temperature where a switch between two refrigerants occurs, the use of economizers, presaturators or heat exchangers between intermediate stages. The objective to be optimized considers both investment and operating costs.

These alternatives are compactly represented as a network. The operating temperature range of each potential refrigerant is discretized and these discretized levels are the nodes of the network. The alternatives corresponding to (i) operation of vapor compression cycle between temperature levels of a particular refrigerant (ii) heat intake from a cooling load (iii) switch between refrigerants are represented by the arcs of the network. The process configuration is obtained once the optimal energy flows in the network are identified. The optimization problem is solved as an MILP. An example of the optimal configuration generated by this procedure for pumping 100kW of heat from 190K to 310K using a ethane-propane refrigeration system is shown in Figure 1. Examples demonstrating the advantage of simultaneous refrigerant selection and cycle synthesis over a sequential approach are given in Vaidyaraman and Maranas [10].

Enzyme Design

DNA recombination techniques provide the backbone of directed evolution experiments for engineering improved proteins and enzymes. The setup of directed evolution experiments is vital to the rapid and economical production of enhanced enzymes since screening a large number of proteins for the desired property is expensive and time consuming. The goal is to develop predictive models for quantifying the outcome of DNA recombination employed in directed evolution experiments for the generation of novel enzymes. Specifically, predictive models are outlined for (i) tracking the DNA fragment size distribution after random fragmentation and subsequent assembly into genes of full length and (ii) estimating the fraction of the assembled full length sequences matching a given nucleotide target. Based on these quantitative models, optimization formulations are constructed which are aimed at identifying the optimal recombinatory length and parent sequences for maximizing the assembly of a sought after sequence target [8].

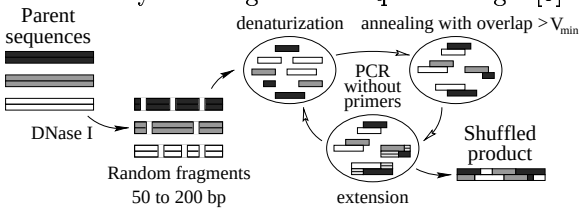

Figure 3: The three steps of DNA shuffling.

A flowchart of DNA shuffling is shown in Figure 3. First an initial set of parent DNA sequences is selected for recombination. The parent sequences undergo *random fragmentation*, typically by DNase I digestion. The fragment length distribution Q_L^0, which describes the fraction of fragments of length L found in the reaction mix-

ture after fragmentation is calculated to be as follows

$$Q_L^0 = \begin{cases} P_{cut}\exp(-P_{cut}L) & \text{for } 1 \le L \le B-1 \\ \exp(-P_{cut}B) & \text{for } L = B \end{cases}$$

Next, the double-stranded fragments within a particular size range (i.e., 50-200 base pairs) are isolated and *reassembled* by the *Polymerase Chain Reaction* (PCR) without added primers. This step is quantified using a fragment assembly model that tracks the fragment length distribution through a given number of annealing/extension steps. This is used to estimate how many shuffling cycles will be needed before full length genes are assembled.

A sequence matching model is developed to aid in the goal of optimizing experimental parameters to maximize the probability of obtaining a desired sequence. This model quantitatively predicts the probability of having a randomly chosen full length sequence, assembled through DNA shuffling, match the given nucleotide sequence target. This model recursively calculates the probability P_i of a reassembled sequence matching the target sequence from position i to position B (length of parent sequence). The probability P_1 represents assembly of the entire target sequence. The recursive expression for evaluating P_i is shown below.

$$P_i = \begin{cases} 1, & i > B \\ \frac{\Delta_{B,B}}{K}, & i = B \\ \sum_{L=L_1}^{L_2} Q_L^0 \sum_{V=V_{min}}^{L-1} A_{L-V,L} & \\ \quad \times \left(\frac{\Delta_{i,i+L-V-1}}{K}\right) P_{i+L-V}, & 1 \le i < B \end{cases}$$

In the above expression, K is the number of parent sequences, (L_1,L_2) are the range of fragment lengths retained for shuffling, Q_L^0 is the probability that a fragment is of length L, $A_{L-V,L}$ is the probability that a fragment of length L will anneal with an overlap V, and $\Delta_{i,i+L-V-1}$ is the number of parent sequences that match the target between the positions i and $i+L-V-1$.

The above predictive sequence matching model enables the formulation of mathematical programs for optimizing the recombinatory fragment length and parent sequence set. The objective of the desired mathematical program is the maximization of the probability of matching the target sequence (P_1). Two variations of this model are considered. In the first variation, the parent sequences which are chosen occur in equal relative concentration. This problem is combinatorial in nature due to the need to choose the optimum set of parent sequences. The binary decision variables are y_k denoting if parent k is chosen and x_L indicating if the recombinatory length is L. This results in a MILP formulation. The second variation allows all the parent sequences to be present but optimizes the relative concentration C_k of each parent sequence. This is solved as a bilinear NLP once for each L in the range being considered to find the optimal recombinatory fragment length.

The importance of using the MILP and bilinear formulations are illustrated through an example. The goal is to shuffle six parent sequences each with a variety of mutations and to produce a sequence containing all twelve of the mutations. The sequences are $B = 151$ nucleotides long and are shown in Figure 4.

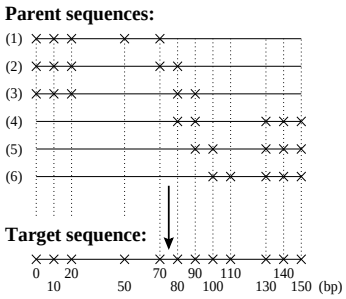

Figure 4: The six parent sequences and the target sequence utilized in the example.

First, when all parents are selected for recombination (achieved by fixing all $y_k = 1$), the optimal recombination probability P_1 of 0.0105% for a fragment length L of 37 nucleotides is confirmed. However, when the complete MILP is solved for both x_L and y_k, the subset of parent sequences 1, 3 and 6 is revealed to be the optimum recombinatory choice with a recombination probability of 0.0294%, an almost three-fold improvement. Note that the new optimal length is

$L = 70$ nucleotides, almost twice the length of the previous optimum implying that the selection of the optimal fragment length L strongly depends on the selection of the parent set. A plot of P_1 versus L for different parent sequence recombination sets is shown in Figure 5. These results suggest a surprising complexity in the shape and form of the P_1 versus L plots for different parent choices. Specifically, the multimodal characteristics of these curves reveal narrow fragment length regions for which favorable recombination results are obtained.

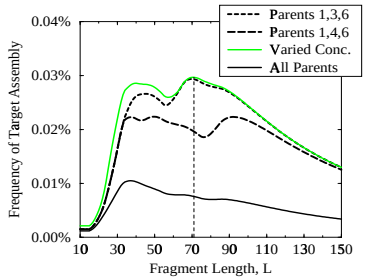

Figure 5: A plot of recombination probability P_1 versus recombinatory fragment length L for different parent sequence sets.

Next, the bilinear formulation is solved, producing the result shown in Figure 5. The optimal recombination probability is equal to 0.0297% at $L = 71$. The optimal parent sequence concentrations for this fragment length are $C_1 = 0.362$, $C_3 = 0.339$, $C_6 = 0.299$, with all other $C_k = 0$, which are fairly close to the equal relative concentration solution. These results indicate that utilizing these formulations can produce a substantial increase in recombination probability.

Summary

This paper discussed the application of optimization techniques to the area of molecular design and bioinformatics. Key issues in the area of polymer design, surfactant design, refrigerant selection, and enzyme design were identified. In each case it was shown how the use of optimization techniques helped in "homing in" on the desired alternatives in a systematic way as opposed to time and labor intensive trial and error approach.

REFERENCES

1. J. Bicerano, Prediction of Polymer Properties. Marcel Dekker, New York, 1996.
2. K. V. Camarda, B. W. Bonnell, C. D. Maranas, and R. Nagarajan, Design of Surfactant Solutions with Optimal Macroscopic Properties. Comput. Chem. Eng., Suppl:S467, 1999.
3. K. V. Camarda and C. D. Maranas, Optimization in Polymer Design Using Connectivity Indices, Ind. Eng. Chem. Res., 38(5):1884, 1999.
4. A. P. Duvedi and L. E. K. Achenie, Designing Environmentally Safe Refrigerants Using Mathematical Programming, Chemical Engineering Science, 51:3727, 1996.
5. R. Gani, B. Nielsen, and A. Fredenslund, A Group Contribution Approach to Computer-Aided Molecular Design, AIChE Journal, 37(9):1318, 1991
6. C. D. Maranas, Optimal Molecular Design under Property Prediction Uncertainty, AIChE Journal, 43(5):1250, 1997.
7. C. D. Maranas, Optimal Computer–Aided Molecular Design: A Polymer Design Case Study, Ind. Chem. Eng. Res., 35(10):3403, 1996.
8. G. L. Moore, C. D. Maranas, K. R. Gutshall, and J. E. Brenchley, Modeling and Optimization of DNA Recombination, Comput. Chem. Eng., 24(2/7):693, 2000.
9. R. Vaidyanathan and M. El-Halwagi, Computer–aided design of high performance polymers, Journal of Elastomers and Plastics, 26(3):277, 1994.
10. S. Vaidyaraman and C. D. Maranas, Optimal Synthesis of Refrigeration Cycles and Selection of Refrigerants. AIChE Journal, 45(5):997, 1999.
11. D. W. van Krevelen, Properties of Polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions, Elsevier, Amsterdam, 3rd edition, 1990.