Operations Research and the Life Sciences – A Healthy Combination

Gunnar W. Klau GUNNAR.KLAU@CWI.NL
CWI, Life Sciences Group, Amsterdam, the Netherlands

Led by Jan Karel Lenstra, a couple of years ago, initiatives have been started to set up a Life Sciences Group at CWI. As a direct beneficiary of these initiatives I am very thankful to Jan Karel Lenstra and to those who helped him with this task.

Here, I will illustrate some links between Operations Research (O.R.), Jan Karel's home turf, and the life sciences. Indeed, now in 2011, a large number of people in CWI's Life Sciences Group are busy with developing O.R. techniques to solve problems from biology. In the following I will describe a general link between O.R. and the life sciences and explain why the same techniques can be used for problems arising in economy and in biology. I will focus on one such example, namely the protein design problem and the problem of assigning frequencies to radio links in telecommunication, which turn out to be close mathematical cousins. Finally, there are some personal remarks.

O.R. and the Life Sciences

A 2004 *Boston Globe* interview with Mark Eisner from Cornell University contains a nice definition of the field of Operations Research (O.R.): *He [Eisner] defines O.R. as "the effective use of scarce resources under dynamic and uncertain conditions." That may sound arcane, but it's pretty much the problem of living—and certainly the central problem of economic life. O.R. isn't economics, however, though most economists have some O.R. training. It's applied mathematics.*

There are three things I like in this definition:

- It clearly states that O.R. is applied mathematics.
- The definition is less dry than more "official" ones like "interdisciplinary field that develops advanced analytical methods to come up with optimal or near-optimal solutions to complex decision-making problems". The issue of invisibility that O.R. shares with other mathematical disciplines is also commented on by Eisner in the same interview by saying that O.R. is "probably the most important field nobody's ever heard of. Indeed, it's not one that's likely to come up at dinner parties."
- It provides a beautiful link to the life sciences by saying that it is "pretty much the problem of living". Going one step further one might say that problems cells face are not so different from the problems that companies face. This analogy then allows for using similar mathematical techniques to solve these problems.

There is already a rich history of applying O.R. to problems in healthcare and applied medicine. Examples include radiotherapy treatment design, robotic surgery, location of healthcare facilities, design of clinical trials, medical ressource allocation, and the optimal scheduling of vaccines.

Recently, however, O.R. is likewise emerging as a crucial component of basic research in modern biology and biomedicine. Microarray technology, next-generation sequencing and advances in proteomics make it now possible to investigate and compare entire genomes or proteomes under different physiological conditions. This leads to new mathematical challenges and has already spurred the development of new algorithms and statistical techniques.

For instance, the shotgun approach to genome assembly leads to a large-scale Traveling Salesman-like graph optimization problem the solution of which has been key to the draft release of the human genome. In phylogenetics, researchers now have moved from the study of trees to networks that allow to model evolutionary events like recombination and lead to new problems that should be addressed with O.R. techniques. More O.R. challenges arise from the study of genetic variations and from personalized medicine in general. In sequence analysis, Bayesian modeling and inference and optimization-based machine learning techniques such as support vector machines as well as data-mining techniques like classification and clustering are popular approaches. Many problems in structural biology and drug design such as protein structure prediction, docking, side-chain placement in protein design and

the comparison of 3D structures benefit from a variety of O.R. techniques.

Last, but not least, systems biology is a recent approach where O.R. has started to play a prominent role. Biologists have focused on studying interactions between components in a biological system. These interactions are frequently described as biological networks that exhibit highly dynamic and interactive behaviour. Examples are metabolic, signal transduction, genetic, and protein-protein interaction networks. Numerous computational challenges arise in the field. While some of them can already be addressed with traditional O.R. techniques like network flows or linear programming, many new problems ask for entirely novel techniques.

Protein Design and Frequency Assignment

This section highlights one example where O.R. links two seemingly unrelated problems—protein design in synthetic biology and radio link frequency assignment in telecommunications. It is a particularly suitable example for this booklet, because it also links work that has been done in CWI's Life Sciences Group with work that has been done by Jan Karel Lenstra.

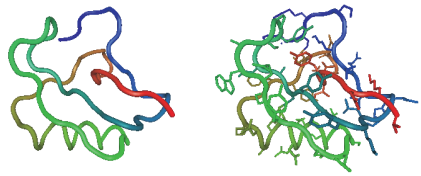
Computational protein design aims at constructing novel or improved functions on the structure of a given protein backbone. Since proteins are key players in virtually all biological processes, the ability to design proteins is of great practical interest, e.g., to the pharmaceutical and biotechnological industries. Experimental methods are time- and money-consuming. Computational approaches are an attractive alternative.

Given the modeled backbone of a protein, the amino acid side-chains have to be placed on this backbone in the energetically most favorable conformation. In protein design instances the candidate side-chains at a certain residue position correspond to more than one amino acid. The choice of the side-chain then determines the amino acid for the design. Two assumptions are commonly made: (i) side-chains adopt only statistically dominant low-energy side-chain conformations, the so-called rotamers, and (ii) the energy of a protein is the sum of intrinsic side-chain energies and pairwise interaction energies. These assumptions lead to the following NP-hard discrete optimization problem: For each residue position choose a rotamer such that the total energy of the protein is minimum. See Figure 1 for an illustration.

A formal graph-theoretic reformulation of the side-chain placement problem is as follows: Given a k-partite graph $G = (V, E)$, $V = V_1 \cup \ldots \cup V_k$, with node costs c_v, $v \in V$, and edge costs c_{uv}, $uv \in E$, determine an assignment $a : \{1, \ldots, k\} \to V$ with $a(i) \in V_i$, $1 \leq i \leq k$, such that the cost

$$\sum_{i=1}^{k} c_{a(i)} + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} c_{a(i)a(j)}$$

of the induced graph is minimum.



(a) Modelled protein backbone without side chains (b) Specific side chain placement

Figure 1. Computational protein design. The task is to determine the placement of side-chains that results in the lowest overall energy. The placement of side-chains determines the amino acids to be used for designing the protein. The right figure shows one feasible solution

Here, each node set V_i corresponds to the candidate rotamers for the set of possible residues at position i. Node costs model self energies of rotamers and edge costs model interaction energies between pairs of rotamers. A solution is given by selecting for each residue position i, $1 \leq i \leq k$, exactly one rotamer $a(i)$. Clearly, the choice of the rotamer determines also the amino acid at this position.

In (Canzar et al., 2011) we propose a novel exact method for the side-chain placement problem that works well even for large instance sizes as they appear in protein design. Our main contribution is a dedicated branch-and-bound algorithm that combines tight upper and lower bounds resulting from a novel Lagrangian relaxation approach. This makes it possible to optimally solve large protein design instances routinely. The project has been supported by a CWI internship to make the stay of PhD student Nora Toussaint possible.

Interestingly, the side-chain placement problem is mathematically very related to the frequency assignment problem in telecommunications as we will see in the following. The presentation of the frequency assignment problem is mainly taken from the report on the CALMA project (Aardal et al., 2002), in which Jan Karel Lenstra and his colleagues investigated many of its problem variants.

The problem appears when setting up and configuring a wireless communication network. Frequencies have to be assigned to radio links in order to enable communication. However, due to the dramatical increase in wireless communication, frequencies have become a scarce resource. Thus, same or similar frequencies have to be reused for different radio links. This may lead to interference problems, for example, when the involved links are close to each other. There are many variants of frequency assignment problems. In the following, we will focus on the minimum interference problem.

In the minimum interference problem we are given a finite set L of radio links, where each link i has to receive a frequency from a finite domain D_i. For pairs of interfering links the assigned frequencies must differ by more than a given distance d_{ij}. This constraint may be soft or hard, that is, its violation may either be penalized in the objective function or not allowed at all. For simplicity of the presentation we omit preassigned frequencies and special treatment of so-called parallel links. The interested reader is referred to (Aardal et al., 2002).

Formally, this simplified minimum interference problem is as follows: Find an assignment of frequencies $f_i \in D_i$ for each $i \in L$ with $|f_i - f_j| > d_{ij}$ for each pair of links $\{i, j\}$ with a hard interference constraint such that

$$\sum_{C} c'_{ij} \delta(|f_i - f_j| \leq d_{ij})$$

is minimum. Here, $\delta(\gamma) = 1$ if condition γ is true and $\delta(\gamma) = 0$ otherwise, and C denotes pairs of $\{i, j\}$ with soft interference constraints, where c'_{ij} is the cost for violating such a constraint.

Lemma 1. *The simplified minimum interference problem is a special case of the side-chain placement problem.*

Proof. Let L, D_i, for all $1 \leq i \leq |L|$, $C \subseteq \binom{L}{2}$, d_{ij} for all $\{i, j\} \in \binom{L}{2}$, c'_{ij} for all $\{i, j\} \in C$, be an instance of the simplified minimum interference problem. We can easily transform it into an instance of the side-chain placement problem as follows:

We set $k := |L|$ and $V_i = D_i$ for all $i = 1, \ldots, k$. We set all node weights c_v to zero for all $v \in V$. We now set the edge weights c as follows:

$$c_{ij} = \begin{cases} 0 & |f_i - f_j| > d_{ij} \\ c'_{ij} & |f_i - f_j| \leq d_{ij} \text{ and } \{i, j\} \in C \text{ (soft)} \\ \infty & |f_i - f_j| \leq d_{ij} \text{ and } \{i, j\} \notin C \text{ (hard)} \end{cases}$$

It is easy to see that the simplified minimum interference problem has a complete assignment if and only if the side-chain placement problem has a solution of total cost $< \infty$. In this case, solution and solution value of the two problems coincide. □

As a consequence of Lemma 1, we could use the algorithm developed for protein design instances to solve instances of the simplified minimum interference problem. It is interesting future work to see how our algorithm performs on instances arising from telecommunications and how to integrate the constraints that have been omitted in the simplified problem statement.

Personal Remarks

There are many more success stories of O.R. in the life sciences and many more will certainly follow in the future. I am happy and proud that we, the Life Sciences Group, have been given the opportunity to contribute to this exciting area at CWI, and I am personally very grateful for Jan Karel's support.

I also want to thank Jan Karel Lenstra on this occasion for the elusive conditions and for the great atmosphere we have at this institute. Among many other examples I could write about I refer to the one shown in Figure 2.

Figure 2. Jan Karel Lenstra, my son Theo, and myself at the CWI midsummer barbecue 2009. Theo has just won the first prize in his life: free entry to Artis for one year.

References

Aardal, Karen I, Hurkens, C, Lenstra, Jan Karel, and Tiourine, S. Algorithms for radio link frequency assignment: The CALMA project. *Operations Research*, pp. 968–980, 2002.

Canzar, Stefan, Toussaint, Nora C, and Klau, Gunnar W. An exact algorithm for side-chain placement in protein design. *Optimization Letters*, March 2011.