Optimization of Solvent Blend Properties Using COSMOtherm and a Simulated Annealing Heuristic
Isaac Dimitrovsky (independent software developer)

Solvent blend selection – a useful application of COSMO-RS and other property prediction methods such as HSP

“For all of the processes ... it is highly desirable to choose an optimum solvent or solvent mixture. Criteria for such an “optimal” solvent can be the cost of the process technology ... as well as the cost aspects of the environmental, health and safety properties of the process.”
- Fast Solvent Screening via Quantum Chemistry: COSMO-RS Approach

The combinatorial challenge

COSMOtherm may take seconds to do a single property prediction. This can quickly make blend selection by exhaustive search impractical. For example, selecting a blend of up to 3 solvents from a set of 45 with fractions restricted to multiples of 5% gives 2,45M possible blends.

Gradient-descent methods can help if the property function is convex or has relatively few local minima, but may not work well if a property has many local minima.

An alternative that may be useful is the class of heuristic search methods. These are generally not guaranteed to find a global minimum, but in practice often find one or more “good” answers, typically within a small percentage of the global minimum. Given the small but significant errors in COSMO-RS predictions, this kind of result might be useful in practice; predicted near-optimal blends might be optimal due to prediction error.

Simulated annealing – a simple but effective heuristic search method

“A first test case – optimizing blends for similarity to a target sigma-potential

Target solvents: dimethyl formamide and dimethyl acetamide, cited as undesirable solvents without known good replacements in Green Chemistry Tools to Influence a Medicinal Chemistry ...

Alfonsi et. al., Green Chemistry 10, p. 31 (2008)

HSP Sphere should produce an excellent solvent when mixed 50:50. This method might have some practical use in finding solvent replacements in cases where the solvent must work with a variety of solutes, or where the solutes are difficult to characterize with COSMO-RS. A similar approach was taken using HSP to find alternatives to xylene in protective coatings - Selecting Effective Xylene Replacements for Protective Coatings, Eastman Chemical Co., 2013

Caching past results also seems worthwhile, both to avoid recalculation and possibly to help guide the search.

A typical implementation of SA:

- start with a run with a random guess
- at each step, randomly modify the guess and compare the new cost with the previous one
- if the cost has improved, always switch to the new guess
- otherwise, switch to the new guess with probability based on the increase in cost (larger increases make switching less likely) and on the step number (as the run progresses, we lower the probability of switching to worse guesses)

For the classic version of SA, we use a virtual temperature T that is lowered as the run progresses, and takes the form of the following expression:

\[ P = e^{-\frac{\Delta F}{T}} \]

This is analogous to annealing by controlled cooling of a metal where the cost function represents a state's energy; this expression then reflects the Boltzmann distribution of the energy difference of the two states.

However, this physical analogy and indeed the particular expression used to control switching to worse guesses is not really essential:

"Forget about this molten metal business. Simulated annealing is effective because it spends much more of its time working on good elements of the solution space than on bad ones, and because it avoids getting trapped repeatedly in the same local optima." Skiena, ibid.

The relatively high cost of setting up and running COSMOtherm scripts suggests a parallel version of SA, with multiple simultaneous runs.

The first algorithm tested uses similar concepts to those in Three Parallel Algorithms for Simulated Annealing, Czech, PPMAM 2001 p. 210, with the addition of caching. A number of runs are calculated in parallel; each is independent, except that at some points the best results are propagated between runs. Caching is just used to avoid recalculations; a cached result is immediately accepted or rejected using the same criterion as for computed results.

I've also experimented with a second algorithm that more drastically modifies SA. In this version we keep a pool of all past guesses and redraw the new set of guesses on each round, using a distribution that is increasingly biased toward lower-cost guesses as the run progresses. This algorithm appeared somewhat more efficient on the tests below.

A first test case – optimizing blends for similarity to a target sigma-potential

This method might have some practical use in finding solvent replacements in cases where the solvent must work with a variety of solutes, or where the solutes are difficult to characterize with COSMO-RS. A similar approach was taken using HSP to find alternatives to xylene in protective coatings - Selecting Effective Xylene Replacements for Protective Coatings, Eastman Chemical Co., 2013

<table>
<thead>
<tr>
<th>Target solvent</th>
<th>Algorithm</th>
<th># of guessed blends</th>
<th>% of search space</th>
<th>selected results</th>
</tr>
</thead>
<tbody>
<tr>
<td>dimethyl formamide</td>
<td>Algorithm 1, 2</td>
<td>4,545</td>
<td>0.9%</td>
<td>96% similarity: 40% propanone, 37% dimethylsulfoxide, 23% thiophene, 58% butanone, 42% dimethylsulfoxide, 92% similarity: 52% ethylacetate, 48% dimethylsulfoxide</td>
</tr>
<tr>
<td>dimethyl acetamide</td>
<td>Algorithm 1, 2</td>
<td>3,081</td>
<td>0.6%</td>
<td>96% similarity: 55% dimethylsulfoxide, 26% triethylamine, 19% butanone, 84% similarity: 70% dimethylsulfoxide, 30% cyclohexane</td>
</tr>
</tbody>
</table>

Blends were selected from among 45 possible input solvents. The best pure input solvents had sigma potential similarity of 59.8% to dimethyl formamide and 57.8% to dimethyl acetamide. Fractions in increments of 10% were used for most of each run; the last few iterations were in increments of 1% to refine the best solutions.