Unravelling Combinatoric Chemistry

Prof RC Ball (Physics) collaborating with Cambridge Chemistry.

February 3, 2008

How do you discover which molecules will assemble in which way to give some desired feature or structure? This project addresses the case where the possible ingredients are allowed to assemble freely and all their different assemblies are screened together. This saves all the chemistry effort of making each assembly separately, but leaves a major challenge of data analsyis: *what is the chemical identity of the good candidates observed?*

Here is the problem in more mathematical terms. Suppose we consider aggregates freely assembled out of (for sake of example) four basic chemical *monomers* A, B, C and D. At typical example aggregate might be AB_2C for an aggregate built of one A, two$B$'s and one C. Under equilibrium conditions we then have:

$$[AB_2C] = K_{AB_2C}[A][B]^2[C] \tag{1}$$

where $[X]$ is notation for the concentration of species X and K_{AB_2C} is the equilibrium constant of this compound, a key piece of themodynamic information about it. The chemist has detected an interesting amount $[X]$ of various compounds, typically by chromatography, but for each one they *require to infer* the chemical formula (e.g. $X = AB_2C$) and the equilibrium constant K_X.

The problem is rendered soluble by repeating the whole experiment using various different proportions of the base momomers A, B, C and D. Their concentrations as free species in each mixture are less than the amounts added because much gets bound up in aggregates, so you may still not know the actual values of $[A]$ etc.

In simple cases the free monomer concentrations *are* measured, so then it is a simple exercise in linear regression after take the logarithm of equations such as (1) to write

$$\ln([X]) = \ln(K_X) + n_A ln[A] + n_B ln[B] + n_C ln[C] + n_D ln[D] \tag{2}$$

and find the composition numbers ($n_A$etc) and equilibrium constant of each aggregate as regression coefficients.

The central challenge of this project is to address the case where the free monomer concentrations are *not* known. Writing $n_{X\alpha}$ as the unknown composition numbers of aggregate X, where α ranges over values correpsonding to A, B, C and D, and $\lambda_{\alpha i}$ as the logarithm of

concentration of monomer α in the i'th experiment, all the equations (2) can be written as

$$\Lambda_{Ki} = \kappa_X + \sum_{\alpha} n_{X\alpha} \lambda_{\alpha i} \tag{3}$$

The above has almost a classic data factorization form. One additional constraints is easy to use, that the composition numbers $n_{X\alpha}$are constrained to be integers. A harder one is that typically some experiments have some monomers absent: their information is crucial to unravelling ambiguities in the problem, but unhelpful to incorporate directly in (3) where infinite logarithms would threaten to appear.

The student will be tasked to:

- understand progress in hand on this problem,
- make it more consistent in tretament of quantitative experimental errors and
- try out on real experimental data,

with further challenges to

- try to assimilate qualitative errors such as the signal for one supposed aggregate being actually a compbination of two and
- tackle more general partial 'cross-talk'.

This project could run in either miniproject period and could lead into a PhD.