Inching towards a new metric

Janet Newman[1], Thomas S Peat[1] and Brian L Lawson[2]

1. CSIRO Division of Molecular and Health Technology, 343 Royal Parade Parkville VIC 3052 Australia
2. Political Science Department, Santa Monica College, 1900 Pico Boulevard Santa Monica California 90405 USA

In protein crystallisation, we often observe that a protein crystallises out of similar conditions in a screen. We wondered if we could formalise this vague concept of ñsimilar conditionsòinto a metric which *quantifies* the similarity or lack thereof between two crystallisation conditions. A crystallisation condition metric could be used not only to ñtighten upôthis loose, but well recognised idea of ñlikenessò, but may also be used as the basis for screen to screen comparisons. The screen to screen comparison enables one to easily choose a bank of commercial screens for initial screening which tests as many different crystallisation conditions as possible.

Distance metrics in crystallisation space

A *metric* is a distance between two points in a space. In normal Euclidian space, distance is given by

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

We define a dissimilarity metric between two crystallisation conditions to be:

$$D_{ij}=1 \text{ (if no species in common)}$$
$$D_{ij}=1/T\sum_{t=1}^{T}\left[s_{ti}\right]-\left[s_{tj}\right]/\max\left[s_t\right]$$

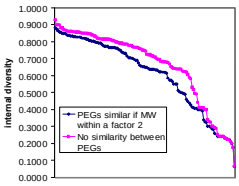
- T is the number of distinct chemical species in conditions *i* and *j*
- [*sti*] is the concentration of species *t* in condition *i*
- max[*st*] is the maximum concentration found for that chemical within chemical space

Thus for two conditions, a distance of 0 indicates that the two conditions are identical, and a value of 1 indicates that the two conditions have nothing in common (*i.e.* are dissimilar).

From this, we describe a screen diversity algorithm ï normalised so that screens with different numbers of conditions can be compared; this algorithm returns a value between 0 and 1 for screen diversity. The algorithm is insensitive to the position of the condition within the screen, and the order of the chemicals within each condition

$$score_{s,t}=\frac{1}{2}\left(\frac{1}{cond_s}\sum_{i=1}^{cond_s}\min_{j\in[1,cond_t]}\left(metric(c_{si},c_{tj})\right)+\frac{1}{cond_t}\sum_{i=1}^{cond_t}\min_{j\in[1,cond_s]}\left(metric(c_{si},c_{tj})\right)\right)$$

A similar approach can be used to look at how much diversity is seen within any one screen.



A pairwise comparison of conditions within a screen gives an indication of how diverse that screen is. This graph shows that there are no screens which consist only of maximally diverse conditions.

Commercial Crystallisation Screens

How do we best choose a starting set of conditions as our entry point for a new crystallisation project? Given that there are over 120 commercially available screens, it is unfeasible to try them all. Without a good understanding of the field, even pinpointing the duplication within this ñcommercial crystallisation spaceòis difficult.

- 7532 conditions
- 4214 are unique (~44 screens worth)
- 3833 distinct conditions (not considering pH)

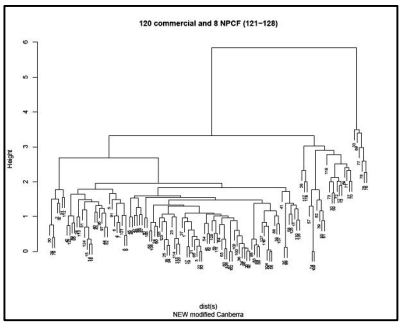
Successful Crystallisation Conditons

The phenomenally successful set of conditions described by Jancarik and Kim in 1991 was based on known crystallisation conditions. In 1991, there were about 700 structures deposited in the Protein Data Bank (PDB). Today there are closer to 46,000 structures in the PDB. Presumably, a set of 50 conditions based on the information we have today would be even better than the screen based on less than 2% of that information. However, ñsuccessful crystallisation spaceòis difficult ï there is no clean (*e.g.* consistent spelling), complete listing of successful conditions, although attempts have been made to obtain one.

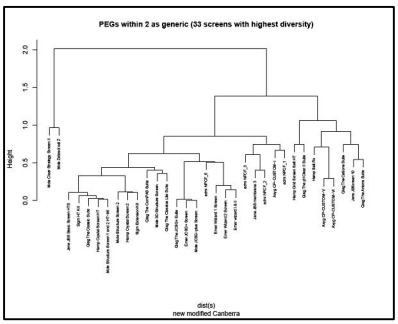
- About 700 of 15000 successful conditions from the PDB are identical to a commercial condition

Problems with the metric:

- No pH data used
- No ions used
- How to capture similarity in PEGs
  - Assume that PEGs that are within a factor of two in molecular weight are similar, and assume that others are dissimilar. This modification will reduce the diversity measured by the metric.
- Only based on chemical names and concentrations (not on ñfundamentalòproperties ï pH, conductivity, ionic strength, water activity, dielectric etc.)



Using the screen distance algorithm, based on the modified Canberra metric described above, we can grow a tree which represents the distance between commercially available screens. The leaves of the dendrogram are screens, and leaves at the same height are equally dissimilar.



In this expanded portion of the dendrogram, we can immediately recognise that the 5 screens in the lower left (Jena Basic HTS, Sigma HT kit, Qiagen The Classics Suite, HR Crystal Screen HT and MD Structure screen 1 and 2 HT are essentially the same.

Bio21 Melbourne Australia

For further information
Contact Dr Janet Newman
Phone +61 3 9662 7326
Email janet.newman@csiro.au
Web www.csiro.au/c3

References:

Jancarik J, SH Kim SH *J. Appl. Cryst.* (1991). **24**, 409-411

Peat TS, Christopher JA, Newman J *Acta Cryst.* (2005). D**61**, 1662-1669

Thanks to:

OpenEye Scientific Software for the use of the BDP database