Comment on 'Protein Isoelectric Point as a Predictor for Increased Crystallization Screening Efficiency'

Thomas Huber[1,2] and Bostjan Kobe[2,3]

[1]Department of Mathematics, [2]Department of Biochemistry and
Molecular Biology, [3] Institute for Molecular Bioscience,
The University of Queensland, Brisbane Qld 4072, Australia

A recent article in this journal [1] describes a statistical predictor to increase the efficiency of protein crystallisation screens. The approach is based on the observation that a correlation exists between the calculated isoelectric point of a protein, pI, and the difference between the pI and pH of the solution in which the protein was crystallised. Kantardjjieff and Rupp specifically comment on the lack of any statistically significant correlation between a protein's pI and pH of crystallisation conditions. This has been well documented in the literature [2,3] and is also well understood in condensed matter science, where polymer model systems have been studied theoretically as well as experimentally for a long time [4, 5].

The purpose of this comment is to point out that while there is always a correlation between pI and pH-pI, it is of no significance when no correlation between pI and pH exists. Ignoring this fact has lead to a serious misinterpretation of crystallistion data. Crystallisation of (bio-)polymers is being widely applied in molecular biology and designed protein-specific crystallisation screens are highly desirable to increase the efficiency of protein crystallizations. We believe, it is important to prevent the misconception that simple pI calculations can be used to design such screens.

The linear correlation coefficient $r_{x,y}$ between two variables x and y, such as pI and pH, is conveniently defined by their variances σ_x and σ_y, and their covariance $\sigma_{x,y}$

$$r_{x,y} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x \sigma_y}}$$

where $\sigma_x = N^{-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$, $\sigma_y = N^{-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$ and $\sigma_{x,y} = N^{-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$; $\bar{x}$ and $\bar{y}$ denote the arithmetic averages of the data x and y, respectively. For uncorrelated data $\sigma_{x,y}$ is zero and as a result $r_{x,y}$ is also zero.

From this definition it is now also straightforward to write the correlation coefficient between x and $y - x$

$$r_{x,y-x} = \frac{\sigma_{x,y} - \sigma_x}{\sqrt{\sigma_x(\sigma_x + \sigma_y - 2\sigma_{x,y})}}$$

For uncorrelated data ($\sigma_{x,y} = 0$), the correlation coefficient between x and x-y is negative and of the size $\frac{\sqrt{\sigma_x}}{\sqrt{\sigma_x + \sigma_y}}$.

This is visually illustrated using 10,000 pI and pH data distributed randomly and uniformly over the full pH range (Figure 1 A and B). Figure 1 C and D were produced using the same analysis that Kantardjjieff and Rupp used to produce figure 3 in their paper, and which lead them to the conclusion "It is clear that basic proteins have a tendency to crystallize 0.5-3 pH units below their pI, whereas acidic proteins prefer to crystallize 0-2.5 pH units above their pI." We demonstrated here that this conclusion is based on misinterpretation of the data and should not be used to guide crystallisation experiments until a correlation between pH and pI is demonstrated.

Acknowledgement: BK is an NHMRC Fellow.

References

[1] Kantardjieff, K. and Rupp, B., Protein isoelectric point as a predictor for increased crystallization screening efficiency, *Bioinformatics* , DOI: 10.1093/bioinformatics/bth066 (2004).

[2] Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C., and Lesley, S. A., Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome, *Acta Cryst. D* **59**, 1028–1037 (2003).

[3] Wooh, J. W., Kidd, R. D., Martin, J. L., and Kobe, B., Comparison of three commercial sparse-matrix crystallization screens, *Acta Cryst. D* **59**, 769–772 (2003).

[4] Belloni, L., Colloidal interactions, *J. Phys. Condens. Matter* **12**, R549–R587 (2000).

[5] Frenkel, D., Soft condensed matter, *Physica A* **313**, 1–31 (2002).

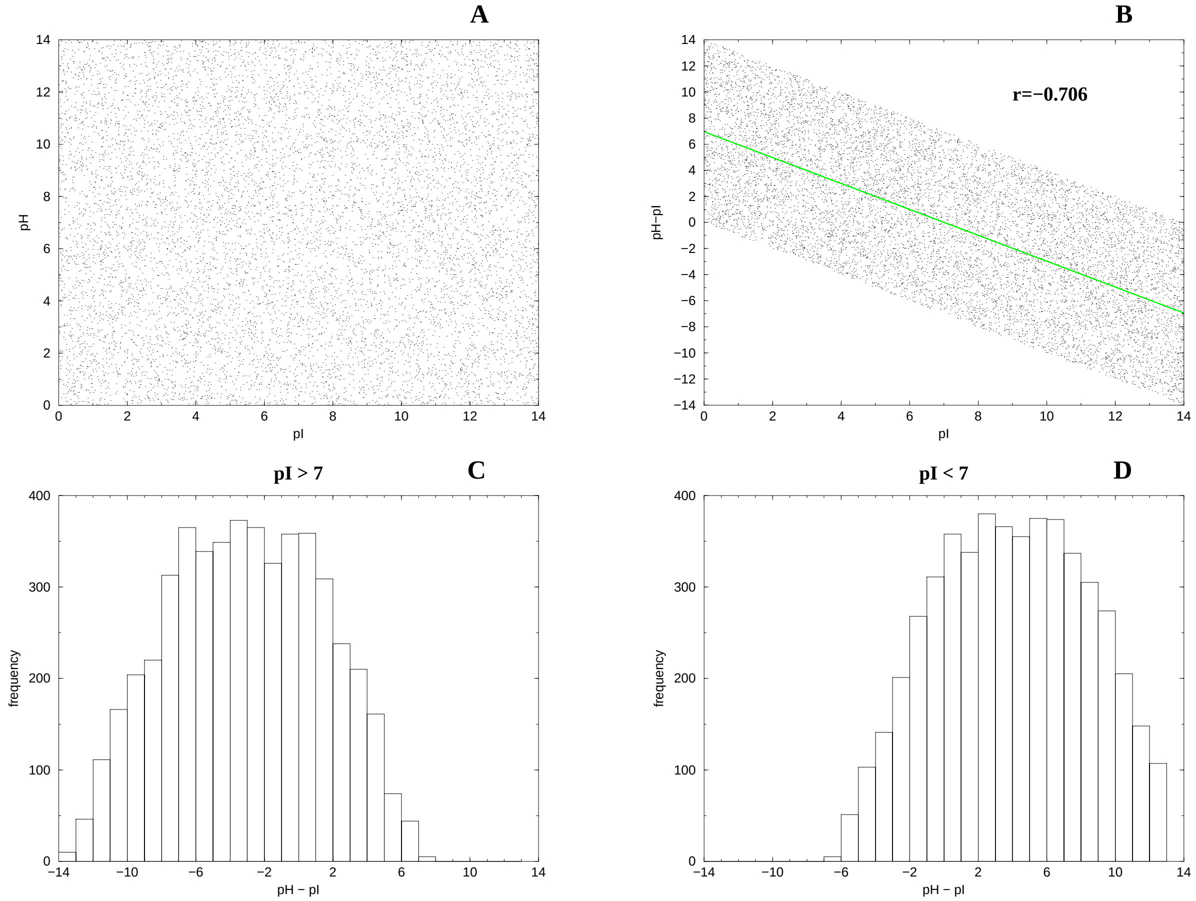

Figure 1: 10,000 uniformly, randomly distributed pI, pH data. A: pH versus pI. B: pH-pI versus pI. Because pI and pH are uniform random over the same range, $\sigma_{pI} \approx \sigma_{pH}$ and $r_{pI,pH-pI} \approx 2^{-\frac{1}{2}}$. C: frequency distribution of pH-pI for data data with pI $>$ 7. D: frequency distribution of pH-pI for data data with pI $<$ 7.