Institution: University of York

Unit of Assessment: 10 - Mathematical Sciences

Title of case study: Statistical pattern recognition applied to protein crystallisation images in the pharmaceutical industry

Period when the underpinning research was undertaken: 2002 – 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Julie Wilson	Professor	Oct 1999 - present
James Foadi	Research Assistant	Feb 2000 – Feb 2001; Jul 2002 – Jun 2004
Chris Walker	Research Assistant	Sep 2004 – Jul 2010
David Watts	Research Assistant	Sep 2004 – Sep 2007
Samarasena Buchala	Research Assistant	Dec 2006 – Dec 2007

Period when the claimed impact occurred: 2018 – 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact (indicative maximum 100 words)

A major bottleneck in drug design and structural biology is the production of high-quality protein crystals, from which their molecular structure can be determined by protein crystallography. Robotic systems used by pharmaceutical companies can perform tens of thousands of crystallisation experiments per day, but manual analysis of the resulting images is monotonous, expensive and error-prone. Wilson's group has pioneered automated image analysis for protein crystallography experiments, culminating in the MARCO (**MA**chine **R**ecognition of **C**rystallization **O**utcomes) method which classifies experimental outcomes with 94% accuracy. MARCO has been incorporated into the information management software of the leading crystallisation robotics manufacturer, Formulatrix, and is now used by global pharmaceutical companies in studies to determine the interactions between proteins and potential drug molecules, saving both time and scarce purified proteins.

2. Underpinning research (indicative maximum 500 words)

Pharmaceutical companies and structural genomics centres routinely use robotics to perform the many experiments required to determine suitable conditions for protein crystallisation. Originally, the results in microlitre droplets were reviewed by human scorers using microscopes, but by the early 2000s automated imaging systems allowed experimental results to be scored on screen. However, visual processing is slow, expensive and prone to error, with inconsistency and idiosyncrasy on the part of crystallographers. Wilson's group has applied statistical pattern recognition techniques of increasing sophistication to show that features extracted from the images can be associated with particular experimental results, producing reliable and consistent classifications.

Wilson's 2002 paper [R1] was one of the first attempts at automatic classification of images from crystallisation experiments, reducing or eliminating human intervention. Her initial work classified individual objects within the droplet of solution in which crystallisation conditions are tested, combining the results to classify the image [R1]. Labelled examples are required for training the classification models, so this approach requires significant effort to produce the training data and later work therefore focused on analysis of the droplet as a whole. In [R2], texture features were extracted using a two-dimensional Fourier transform. The intensities in the Fourier-transformed images were modelled by a power law with parameters determined for different spatial frequency ranges. In addition to these model parameters, statistical measures obtained from deviations from the fitted models were used to provide characteristic variables for classification. The information acquired in the frequency domain was complemented by features extracted from correlations between different scales of a wavelet transform. In a novel application of wavelet decomposition, the distribution of wavelet coefficient values in each sub-band image was modelled with a generalized Gaussian distribution to

provide discriminatory variables, and these were then combined with second-order statistics obtained from joint probability distributions across different wavelet scales to classify the images [R3]. This work was carried out in collaboration with the Oxford Protein Production Facility (OPPF), who integrated the software into their crystallisation pipeline. OPPF used the software to rank experimental results for human inspection, drastically reducing the number of experiments that needed to be visually assessed. Although accuracy was improved by combining multiple classifiers [R4], the limited training set and dependence on the particular imaging system used meant that the software was not widely applicable.

Much more recently, there has been a major breakthrough. The MARCO (MAchine Recognition of Crystallization Outcomes) project is a collaboration involving Google data scientists, major pharmaceutical companies and structural genomics centres, providing a database of over half a million images. Building on Wilson's ideas and combining state-of-the-art machine learning algorithms with a huge data set for training, it has achieved > 94% correct classification on test-set images from multiple imaging platforms [R5]. This is a major step forward both in reliability and portability, avoiding tailoring to specific imaging systems.

In addition, Wilson is pursuing other aspects of the crystallisation optimisation problem. The pH of an experiment is an important parameter and various buffers are used to maintain a specific pH. It is common practice to search the Protein Data Bank (PDB; http://www.rcsb.org) for conditions in which similar proteins were crystallised, but the pH of the experiment is usually recorded as that of the buffer solution and can be highly inaccurate. In collaboration with the Bio21 Collaborative Crystallisation Centre (C3) in Australia, Wilson has shown that a better estimate of the true pH can be predicted by modelling the effects of other chemicals in the crystallization solution [R6].

3. References to the research (indicative maximum of six references)

[R1] *Wilson, J. Towards the automated evaluation of crystallisation trials. 2002. Acta Cryst., D58, 1907- 1914. DOI:10.1107/S0907444902016633

[R2] *Walker, C.G., Foadi, J. and Wilson, J. 2007. Classification of protein crystallisation images using Fourier descriptors. J. Appl. Cryst. 40, 418-426. DOI:10.1107/S0021889807011156

[R3] *Watts, D., Cowtan, K. and Wilson, J. 2008. Automated classification of crystallisation experiments using wavelets and statistical texture characterization techniques. J.Appl. Cryst., 42, 8-17. DOI:10.1107/S0021889807049308

[R4] *Buchala, S. and Wilson, J. 2008. Improved classification of crystallization images using data fusion and multiple classifiers. Acta Cryst., D64, 823-833. DOI:10.1107/S0907444908014273

[R5] *+Bruno, A.E., Charbonneau, P., Newman, J., Snell, E.H., So, D.R., Vanhoucke, V., Watkins, C.J., Williams, S. and Wilson, J. 2018. Classification of crystallization outcomes using deep convolutional neural networks. PloS one, 13(6). DOI:10.1371/journal.pone.0198883

[R6] *Wilson, J., Ristic, M., Kirkwood, J., Hargreaves, D. and Newman, J., 2020. Predicting the effect of chemical factors on the pH of crystallisation trials. iScience, p.101219. DOI:10.1016/j.isci.2020.101219

*= peer reviewed publication; +=returned to REF2021

4. Details of the impact (indicative maximum 750 words)

Pharmaceutical companies aspire to automate the entire molecular structure elucidation process, from protein expression and purification to structure solution. Molecular structure determination allows interactions between target molecules and potential drugs to be understood and is both a critical step and a major bottleneck in drug discovery and development.

Streamlined methods for obtaining X-ray crystallographic structures – the gold standard of molecular structural analysis – are therefore of vital importance to the pharmaceutical industry. A central requirement is to identify optimal conditions for crystal formation. Automated analysis allows enhanced reproducibility, better quality control and continuous improvement of experiments to optimise crystallisation conditions. A typical crystallisation experiment involves the screening of 768 variants of a given sample in different conditions (temperature, pH and combinations of different additive chemicals). Fewer than 5% of crystallisation attempts produce useful results, with a success rate as low as 0.2% in some contexts. Pharmaceutical companies compensate for the low success rate by sampling chemical space with a throughput of up to 1000 individual experiments per hour, using automated imaging systems to record the results [E1]. Accurate automated image analysis is essential to manage the myriad of experimental results; the critical importance of this task can be seen from the participation of global pharmaceutical companies GlaxoSmithKline, Merck and Bristol-Myers Squibb in the MARCO project [E2, E3]. Building on research by Wilson's group, the MARCO classifier provides automated image analysis that can be used with different imaging systems, with varying resolution and fields of view, from different experimental set-ups (e.g. microbatch under oil and hanging drop). The impact of Wilson's work, both directly and through MARCO, is illustrated by the following examples.

AstraZeneca has a world-leading industrial crystallography group that has implemented its own classification system in collaboration with Wilson. The team leader in Data Science and Imaging at AstraZeneca said "The research… by Prof Wilson and colleagues directly led to the implementation of the method within AstraZeneca for routine crystal screening. This reduces subjectivity and increases throughput, saving both time and money and accelerating the drug discovery process" [E4]. The reliable results obtained via automated image analysis provide data on both successful and unsuccessful experiments, allowing data mining approaches to relate conditions to properties of the protein that are known or can be measured (*e.g*. molecular weight, amino acid sequence, pI, *etc.*) before crystallisation trials begin. AstraZeneca have designated a "searchable, explorable, and inferable" AstraZeneca Crystal Atlas as a core research focus [E4].

The ***Bio21 Collaborative Crystallisation Centre*** (C3) at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) has implemented a MARCO pipeline and now has ~10 million MARCO scored images [E5]. The Facility Manager at C3 says that users of the C3 crystallisation service [E6] "find MARCO scores really useful", saving "an astonishing amount of time" and that C3 "are finding MARCO gives a much more accurate description of the outcomes than humans do. This speeds up the optimisation process allowing diffraction quality crystals to be obtained in fewer experiments" [E5]. Furthermore, C3 are now building a database with "consistent scoring for all of the droplets ever set up in C3" – this consistency being essential for data-mining which will "inform future crystallization strategies, including novel chemical screens and mutagenesis programs" [E5].

Formulatrix are a US based company providing highly specialised robotics and software to manage the entire protein crystallization process, from design and dispensing of chemicals to image viewing and analysis. The MARCO scoring algorithm is incorporated into the Formulatrix ROCK MAKER® Laboratory Information Management System (LIMS), version 3.15, and is a notable selling point for the product, highlighted on the product's webpage [E7]. The ROCK MAKER Software update 3.15 promotes the MARCO functionality as "saving you time and removing the guesswork from scoring images" [E7]. Formulatrix are the market leaders, supplying software and robotics for automating protein crystallization to leading pharmaceutical companies and academic research institutions around the world. The Product Manager and Application Scientist at Formulatrix confirmed that ROCK MAKER 3.15 incorporating the MARCO software is now being used by the pharmaceutical industry. He stated that "MARCO is now an important feature of our software and … may well represent an important turning point for our company" [E8].

MARCO usage has rapidly become standard for major pharmaceutical companies. For example, an Associate Principal Scientist at ***Merck*** said "we are using MARCO. It's kind of routine now, so that you don't have to check whole crystallisation plates, but pick the drops with crystals", while a Senior Scientist and Preclinical Program Manager at ***GlaxoSmithKline*** confirmed that GSK is using MARCO via ROCK MAKER 3.15 [E9].

In summary Wilson's work has led to the development of systems now used routinely by major pharmaceutical companies to accelerate a critical step in the drug discovery and development pipeline, increasing accuracy, saving time and resources, and with prospects to play an important role in forming future crystallisation strategies.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[E1] Yibin Lin (2018) What's happened over the last five years with high-throughput protein crystallization screening? Expert Opinion on Drug Discovery, 13:8, 691-695, DOI: 10.1080/17460441.2018.1465924

[E2] MARCO website: https://marco.ccr.buffalo.edu/ (accessed 7/11/20)

[E3] Open source classifier available from: https://github.com/tensorflow/models/tree/master/research/marco/ (accessed 7/11/20)

[E4] Letter of support from Associate Director in Data Science at AstraZeneca.

[E5] E-mail from Facility Manager of Bio21 Collaborative Crystallisation Centre at CSIRO, Melbourne, Australia.

[E6] Collaborative Crystallisation Centre (C3): https://research.csiro.au/crystal/ (accessed 7/11/20)

[E7] Website for Formulatrix describing the ROCK MAKER® Laboratory Information Management System with the incorporation of MARCO image scoring.

[E8] Corroborating letter from Formulatrix.

[E9] E-mails from Senior Scientist and Preclinical Program Manager at GlaxoSmithKline Pharmaceuticals Associate Principal Scientist at Merck.