Journal of
Applied Crystallography
ISSN 0021-8898

Received 17 June 2008
Accepted 9 September 2008

AutoSherlock: a program for effective crystallization data analysis

Raymond M. Nagel,[a] Joseph R. Luft[a,b] and Edward H. Snell[a,b]*

[a]Hauptman–Woodward Medical Research Institute, SUNY at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA, and [b]Department of Structural Biology, SUNY at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA. Correspondence e-mail: esnell@hwi.buffalo.edu

A program, *AutoSherlock*, has been developed to present crystallization screening results in terms of chemical space. This facilitates identification of lead conditions, rational interpretation of results and directions for the optimization of crystallization conditions.

1. Purpose

The High-Throughput Screening (HTS) Laboratory (Luft *et al.*, 2003) at Hauptman–Woodward Medical Research Institute (HWI) uses an array of 1536 different chemical cocktails to screen macromolecular samples for crystallization from over 850 collaborating laboratories worldwide. These crystallization experiments are imaged automatically over a period of several weeks, generating thousands of images to be examined. Currently this examination is performed manually, though an automated analysis is anticipated for the future. Images are analyzed and scored (by perceived outcome), and then the data are interpreted to decide on the direction of optimization. *AutoSherlock* is a new software program developed to aid the data interpretation and subsequent optimization of crystallization conditions by graphically linking the images to scores and presenting them in chemical space (Snell *et al.*, 2008).

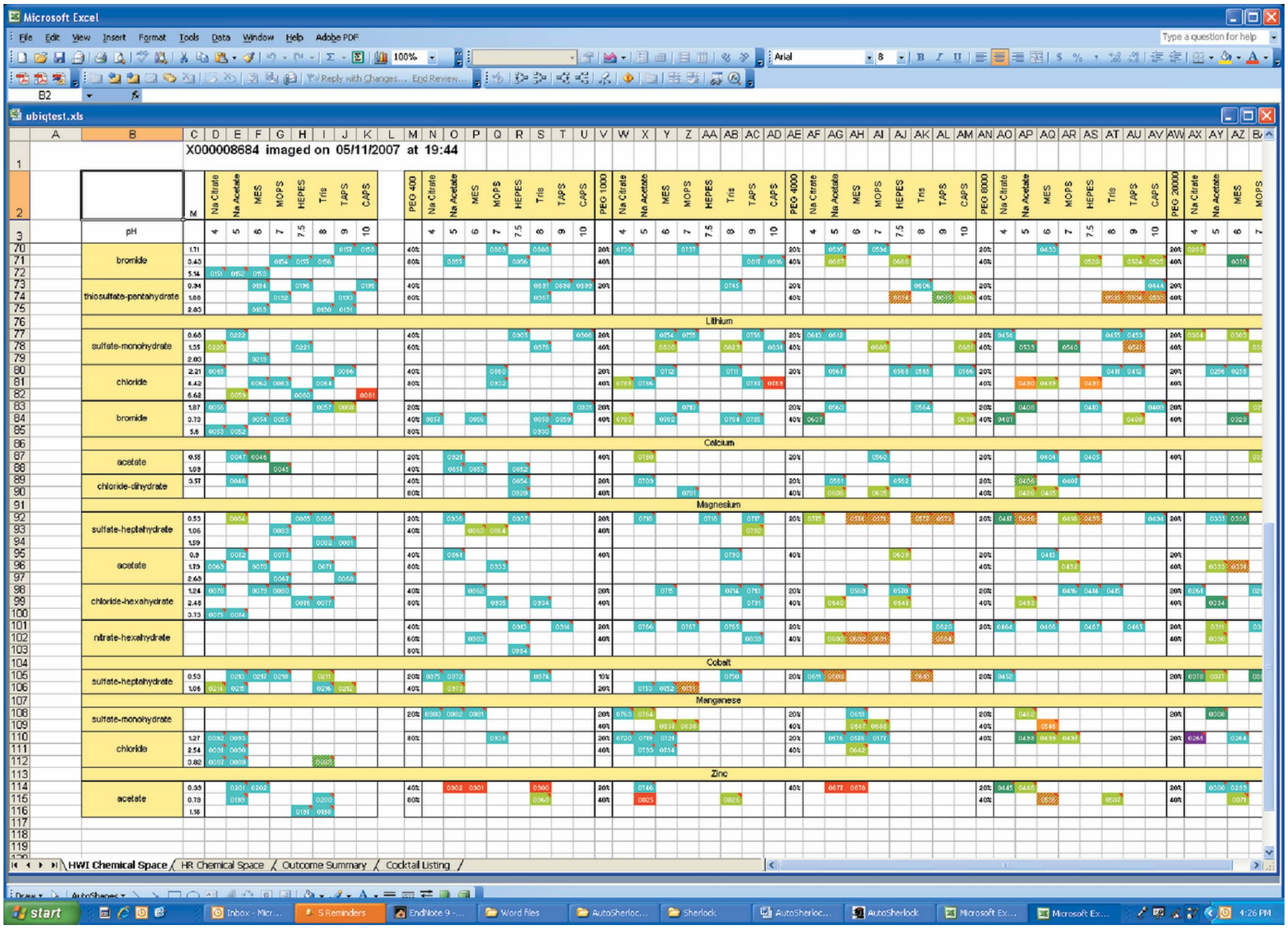
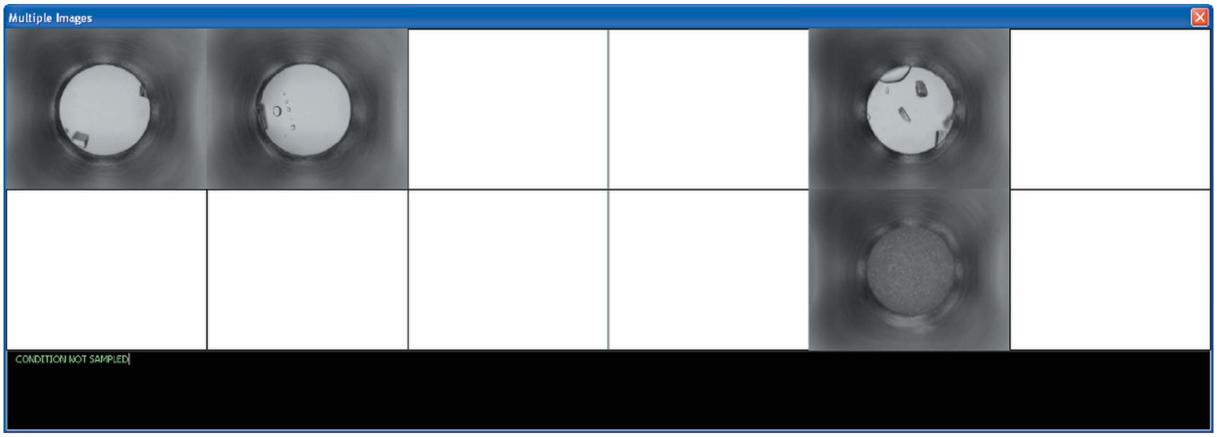


Figure 1
A screenshot showing part of the 'HWI chemical space' worksheet. In this case all the images have been classified and are represented in chemical space by colored blocks on the sheet. Crystal hits are shown as red and other outcomes as different colors. In this case, the data are arranged horizontally by pH and increasing PEG molecular weight, ranging from no PEG (group 1) to PEG 20 K, and vertically by cation, with this subdivided by anion with the lowest concentration first. The cations and anions are ordered by default in an approximate Hofmeister series. All ordering is user controlled.

doi:10.1107/S0021889808028938



CONDITION NOT SAMPLED

Figure 2
Shown here is the window that results from selecting a 2 × 6 section of the worksheet in Fig. 1. Three images classified as crystals and one as precipitate are shown. Also clearly indicated is the experimental space that is un-sampled owing to the incomplete factorial design of the cocktails (*i.e.* the blank spaces). In this case, the cocktail is zinc acetate, with the concentration of PEG 400 increasing top to bottom from 40 to 80%, and the pH varying left to right. Classification and chemical information may be revealed by moving the cursor over each image. Each image may also be selected to show the full-sized image with matching data (see Fig. 3).

2. Strategy

The cocktails used for crystallization screening consist of three groups: salts, different molecular weight PEGs and commercially available screens. The salts and PEG groups, 1 and 2, were constructed using an incomplete factorial design (Audic *et al.*, 1997). These are described in detail elsewhere (Luft *et al.*, 2003; Kempkes *et al.*, 2008). Images are currently manually classified pending automated image analysis developments using the definitions clear, phase separation, precipitate, skin, crystal, junk and unsure. Images can have multiple classifications with the exception of clear, and each classification is converted into a numerical score for use by *AutoSherlock*. The image scores and chemical information for each experiment are written to a text file (*.mso). The filename contains a textual signature which links it logically to the 1536 images associated with the sample. *AutoSherlock* processes the scored file and produces a Microsoft *Excel* workbook containing tools and reports for analyzing the data. Each cocktail from the experiment is color-coded by score and plotted into the chemical space grid by its combination of anion, cation, pH and PEG concentration. This representation allows the user to quickly gain an overview of the experiment results and obvious crystallization trends, and also to reference the image for the results from any cocktail in the context of its chemical 'neighbors'.

3. Computer language and requirements

AutoSherlock was written in Visual Basic 2005, and requires Microsoft .NET Framework 3.0 (or later), Microsoft *Excel 2003* and the Microsoft *Office 2003* primary interop assemblies (PIA) redistributable. The resulting workbook files are separate from the *AutoSherlock* application and require only Microsoft *Excel 2003*. Macros must be enabled for most features to work properly.

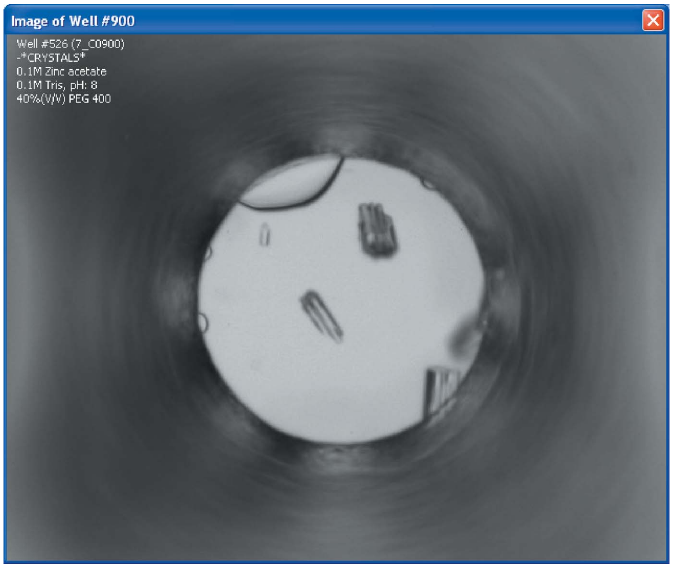
4. Input requirements

AutoSherlock requires a results file (*.mso) generated by our image-review software, *MacroScope*. This file contains chemical information for each cocktail used, along with any scores that have been assigned by the *MacroScope* user. Although *AutoSherlock* requires that these files be XML-formatted, it contains a built-in interpreter to automatically convert the current plain-text files into XML.

Optionally, if the corresponding JPEG images are available on disk, they can be 'linked' physically to the workbook when *AutoSherlock* is run. This allows the user(s) to take full advantage of the workbook by viewing the images in the context of their scores and chemical composition. Because the JPEG images are normally RAR-compressed, a small utility program, *BackPacker*, is bundled with *AutoSherlock* to provide a convenient GUI for decompressing RAR archives.

5. Results

AutoSherlock produces a single *Excel* workbook that contains four worksheets. These show the results from the incomplete factorial (HWI) cocktails; results from commercial screens; a global overview of the results; and a listing of the image analysis scores, image names and biochemical conditions. Fig. 1 shows an example of part of the



Well #526 (7_C0900)
-*CRYSTALS*
0.1M Zinc acetate
0.1M Tris, pH: 8
40%(V/V) PEG 400

Figure 3
Selecting any image from the window shown in Fig. 2 results in a full-sized image, which is shown here. In addition, this view displays chemical information and a score.

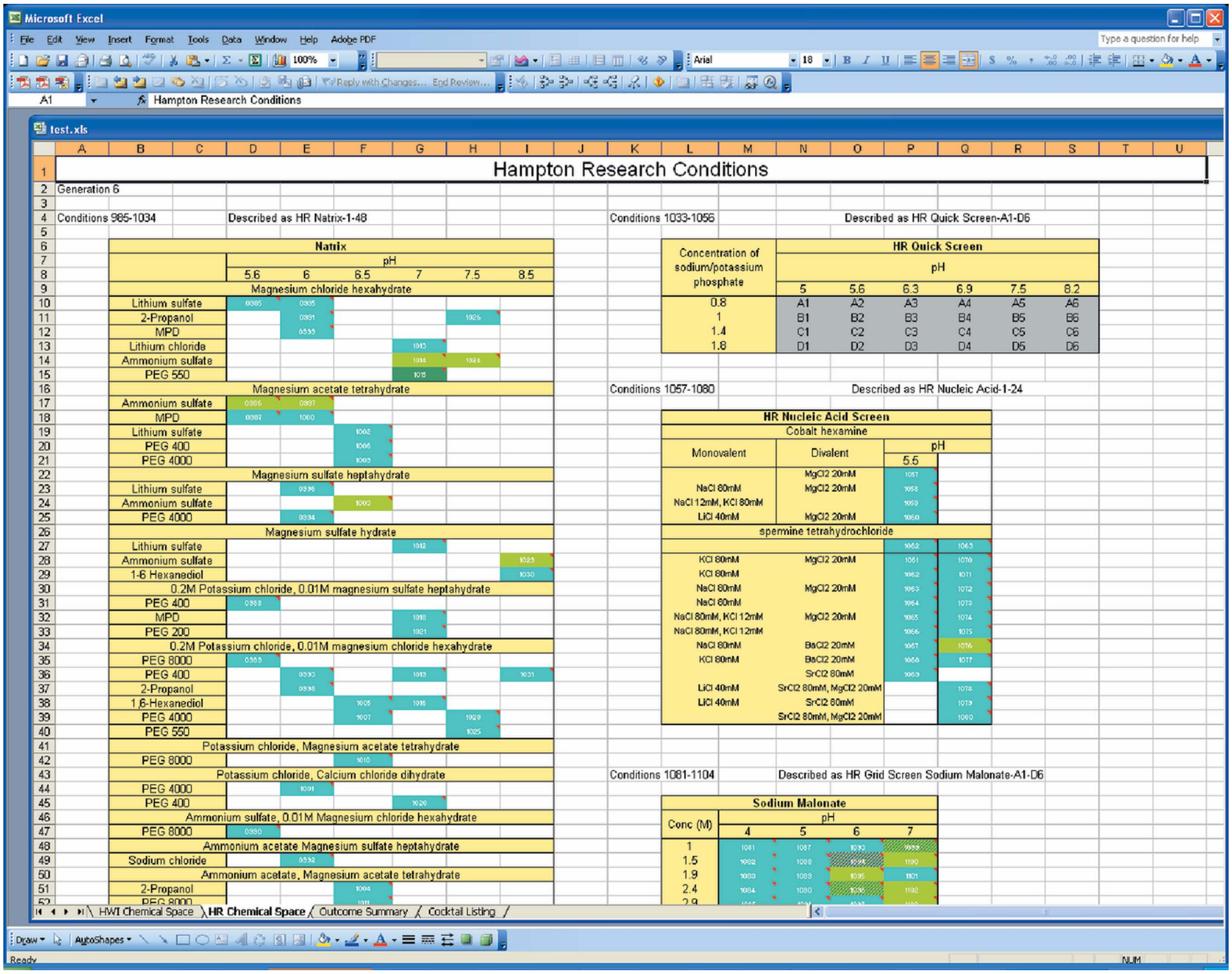
HWI chemical space produced. The results are color-coded according to the classification, with red signifying crystal(s) and turquoise being clear. Any unclassified experiments are colored in gray. The white spaces represent chemical space that is un-sampled. In Fig. 1, the space is arranged horizontally with groups 1 and 2 in order of increasing molecular weight as a function of pH. Vertically, it is sorted by cation then anion in an approximate Hofmeister series. Not all the space is visible in the figure. In Fig. 2, a 2 × 6 area of the spreadsheet has been selected and four images associated with the scores are seen. In this case the horizontal axis represents increasing pH while the vertical axis represents increasing PEG 400 concentration from 40 to 60%. Selecting any of these images enlarges it for closer examination (Fig. 3). It is also clear which conditions have not been sampled. An example of how these data can be used for optimization is presented elsewhere (Snell *et al.*, 2008).

In Fig. 4, the commercial screen spreadsheet is shown. In this illustration, there are few hits visible and the Quik Screen has not been classified. Selecting a Grid Screen such as the Quik Screen produces rapid visual information on the sensitivity of the sample to fine changes in biochemical conditions. The summary worksheet is shown in Fig. 5. The outcomes are sorted by classification, and selecting any single outcome or multiple outcomes brings the associated images up for examination. This provides a rapid means of checking the accuracy of classification.

Finally, the last worksheet is a cocktail listing that provides the cocktail number, the image filename and the classifications associated with that cocktail (if applicable). If the cocktail is a commercial cocktail then the commercial name is also listed. The cocktail components are given in terms of the name, chemical formula, concentration, units and pH. In its present form, up to ten chemical components can be accommodated, but the maximum number currently used in the 1536 screen is six. This last worksheet is intended for export into laboratory database systems for data mining and record keeping, providing not only data on success, but also the outcome of the failure to aid optimization (Snell *et al.*, 2008).

6. Continuing development

AutoSherlock grew out of a need for displaying large quantities of crystallization screening data in a manner that could be rapidly interpreted. It is in daily use in the laboratory and feedback is directing future developments. Currently, images are classified manually using an in-house image-viewing application, *MacroScope*.



A1 Hampton Research Conditions

Hampton Research Conditions

Generation 6

Conditions 985-1034 Described as HR Natrix-1-48

Natrix	pH 5.6	6	6.5	7	7.5	8.5
Magnesium chloride hexahydrate						
Lithium sulfate	0985	0995				
2-Propanol		0991			1026	
MPD		0999				
Lithium chloride				1013		
Ammonium sulfate				1014	1024	
PEG 550				1015		
Magnesium acetate tetrahydrate						
Ammonium sulfate	0986	0997				
MPD	0987	1000				
Lithium sulfate			1002			
PEG 400			1006			
PEG 4000			1009			
Magnesium sulfate heptahydrate						
Lithium sulfate		0996				
Ammonium sulfate			1003			
PEG 4000		0994				
Magnesium sulfate hydrate						
Lithium sulfate				1012		
Ammonium sulfate						1029
1-6 Hexanediol						1030
0.2M Potassium chloride, 0.01M magnesium sulfate heptahydrate						
PEG 400	0988					
MPD				1018		
PEG 200				1021		
0.2M Potassium chloride, 0.01M magnesium chloride hexahydrate						
PEG 8000	0989					
PEG 400		0993		1019		1031
2-Propanol		0998				
1,6-Hexanediol			1005	1016		
PEG 4000			1007		1028	
PEG 550					1025	
Potassium chloride, Magnesium acetate tetrahydrate						
PEG 8000			1010			
Potassium chloride, Calcium chloride dihydrate						
PEG 4000		1001				
PEG 400				1020		
Ammonium sulfate, 0.01M Magnesium chloride hexahydrate						
PEG 8000	0990					
Ammonium acetate Magnesium sulfate heptahydrate						
Sodium chloride		0992				
Ammonium acetate, Magnesium acetate tetrahydrate						
2-Propanol			1004			
PEG 8000			1011			

Conditions 1033-1056 Described as HR Quick Screen-A1-D6

Concentration of sodium/potassium phosphate	HR Quick Screen pH 5	5.6	6.3	6.9	7.5	8.2
0.8	A1	A2	A3	A4	A5	A6
1	B1	B2	B3	B4	B5	B6
1.4	C1	C2	C3	C4	C5	C6
1.8	D1	D2	D3	D4	D5	D6

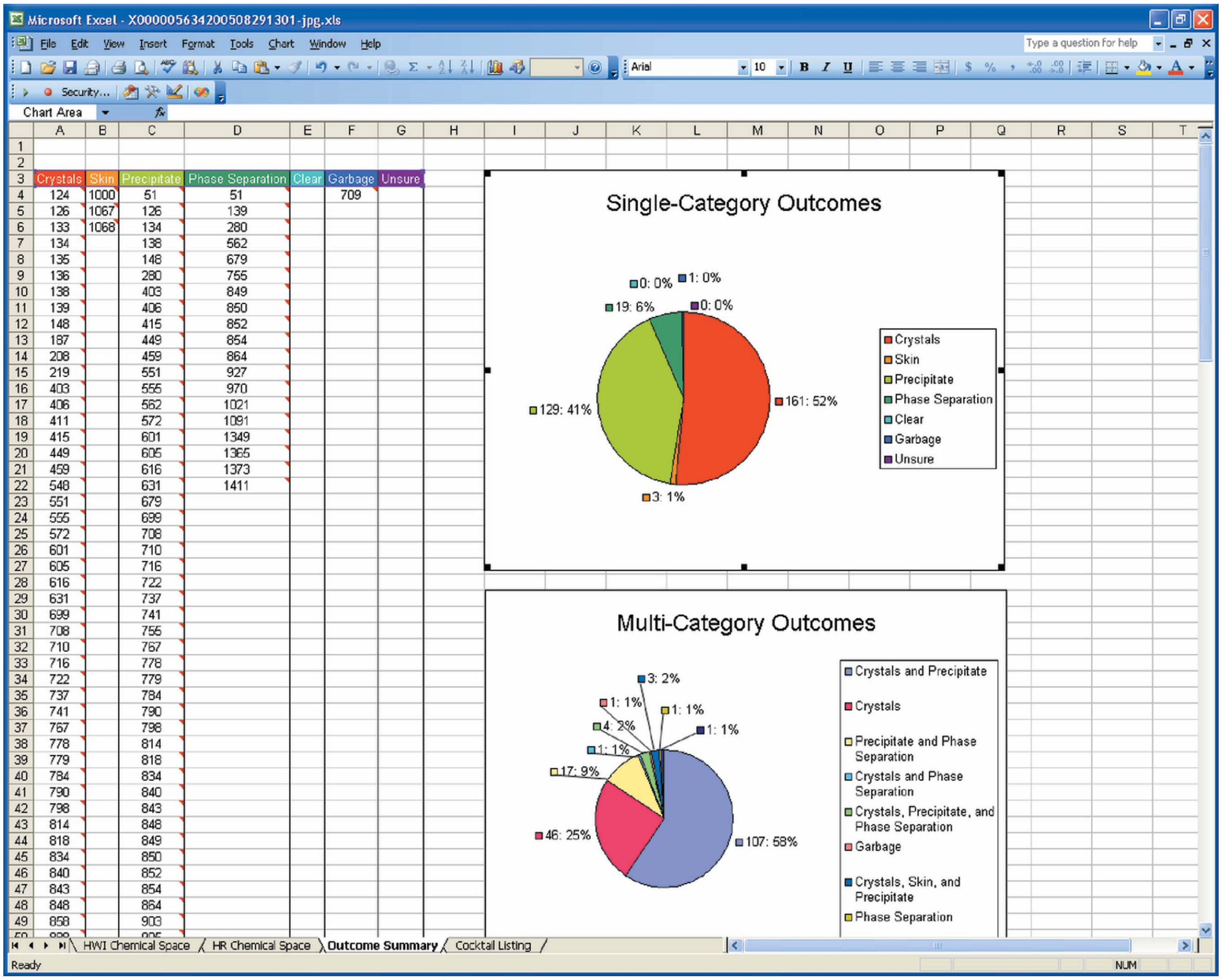
Conditions 1057-1080 Described as HR Nucleic Acid-1-24

HR Nucleic Acid Screen Monovalent	Divalent	pH 5.5	
Cobalt hexamine			
	MgCl2 20mM	1057	
NaCl 80mM	MgCl2 20mM	1058	
NaCl 12mM, KCl 80mM		1059	
LiCl 40mM	MgCl2 20mM	1060	
spermine tetrahydrochloride			
		1062	1063
KCl 80mM	MgCl2 20mM	1061	1070
KCl 80mM		1062	1071
NaCl 80mM	MgCl2 20mM	1063	1072
NaCl 80mM		1064	1073
NaCl 80mM, KCl 12mM	MgCl2 20mM	1065	1074
NaCl 80mM, KCl 12mM		1066	1075
NaCl 80mM	BaCl2 20mM	1067	1076
KCl 80mM	BaCl2 20mM	1068	1077
	SrCl2 80mM	1069	
LiCl 40mM	SrCl2 80mM, MgCl2 20mM		1078
LiCl 40mM	SrCl2 80mM		1079
	SrCl2 80mM, MgCl2 20mM		1080

Conditions 1081-1104 Described as HR Grid Screen Sodium Malonate-A1-D6

Sodium Malonate Conc (M)	pH 4	5	6	7
1	1081	1087	1093	1099
1.5	1082	1088	1094	1100
1.9	1083	1089	1095	1101
2.4	1084	1090	1096	1102
2.9	[illegible]	[illegible]	[illegible]	[illegible]

Figure 4
A screenshot showing part of the ‘HR chemical space’ display. Here the commercial screens used in the 1536 cocktails are arranged in a pre-defined grid according to the most appropriate chemical space.



	A	B	C	D	E	F	G
1							
2							
3	Crystals	Skin	Precipitate	Phase Separation	Clear	Garbage	Unsure
4	124	1000	51	51		709	
5	126	1067	126	139			
6	133	1068	134	280			
7	134		138	562			
8	135		148	679			
9	136		280	755			
10	138		403	849			
11	139		406	850			
12	148		415	852			
13	187		449	854			
14	208		459	864			
15	219		551	927			
16	403		555	970			
17	406		562	1021			
18	411		572	1091			
19	415		601	1349			
20	449		605	1365			
21	459		616	1373			
22	548		631	1411			
23	551		679				
24	555		699				
25	572		708				
26	601		710				
27	605		716				
28	616		722				
29	631		737				
30	699		741				
31	708		755				
32	710		767				
33	716		778				
34	722		779				
35	737		784				
36	741		790				
37	767		798				
38	778		814				
39	779		818				
40	784		834				
41	790		840				
42	798		843				
43	814		848				
44	818		849				
45	834		850				
46	840		852				
47	843		854				
48	848		864				
49	858		903				

Figure 5
A screenshot showing part of the 'outcome summary' worksheet. The outcomes are presented according to score with the cocktail number displayed. Pie charts are generated showing the breakdown of single and multiple category outcomes. Selecting any cell will bring up the associated image. Selecting multiple cells will bring up multiple images. This provides a rapid means to review image classification.

Scoring is a weakness of many high-throughput laboratories and a development effort is underway to automate this aspect. In the future, *MacroScope* will be combined with *AutoSherlock* in an application, independent from Microsoft *Excel*, that can both score and display screening results. The new application will be written in platform-independent Java code.

7. Distribution and documentation

AutoSherlock supports generation 5–8 of the HWI screening cocktails in addition to the current generation 8a. It is available on request to previous and current users of the HWI HTS Laboratory. Full documentation is available for the software. The authors would welcome adaption to other high-throughput crystallization systems.

The authors would like to acknowledge the support of the staff in the HWI High-Throughput Screening Laboratory for testing and suggestions for the software. In particular we would like to acknowledge Jennifer Wolfley and Meriem Said. Melvin Parker helped with initial coding ideas, and Ann Wojtaszczyk provided scored crystal examples. Dr Anna Gardberg at Oak Ridge National Laboratory is thanked for initial testing. Finally, Dr George DeTitta at HWI is acknowledged for insightful comments and support with this analysis. The John R. Oishei Foundation, the Cummings Foundation, NIH grants U54 GM074899-01, U54 GM07495-04 and GM-64655, and Erie County are acknowledged for funding support.

References

Audic, S., Lopez, F., Claverie, J. M., Poirot, O. & Abergel, C. (1997). *Proteins Struct. Funct. Genet.* **29**, 252–257.

Kempkes, R., Stofko, E., Lam, K. & Snell, E. H. (2008). *Acta Cryst.* D**64**, 287–301.

Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K. & DeTitta, G. T. (2003). *J. Struct. Biol.* **142**, 170–179.

Snell, E. H., Nagel, R. M., Wojtaszcyk, A., O'Neill, H., Wolfley, J. L. & Luft, J. R. (2008). *Acta Cryst.* D**64**, 1240–1249.