UNIVERSITY of ALABAMA in HUNTSVILLE AND IXPRESSGENES, INC

ASSOCIATIVE EXPERIMENTAL DESIGN (AED) ANALYSIS USER MANUAL (v1.2)

06.11.2018

This document is prepared to help users to use Associative Experimental Design (AED) analysis program. To use the program, the program needs to be installed and run following the instructions carefully explained below.

1. Setting Up the Input Data File(s)

It is important that screen components be consistently named. Otherwise during the analysis PEG 4K, PEG 4000, polyethyleneglycol 4000, and poly ethylene glycol 4000 will show up in the analysis as four different compounds. We do this by copying each screen into a combined listing in Excel. The condition coordinates (A1, A2…H11, H12) are kept for each screen, and a numerical listing, from 1 to the total number of screen conditions, is added. This listing will be used to regain the screens at the end of this process.

The program input starts with the buffer, then 1'st, 2'nd, 3'rd, etc. precipitant, so some reorganizing from the 'as supplied by the manufacturer' listing may be needed. Salts are divided into their component cations and anions, and non-essential information (i.e., hexahydrate) that may be important for buying the chemical but not to its effect is removed. The combined listing is then sorted, starting with the first column (the buffer). A second sort level can be added here for the pH value. Don't worry about empty conditions – i.e., where there is no buffer, where the precipitant is a polymer or organic (has no counter ion), etc.

Once the listing has been sorted go through and ensure that the naming is consistent. This can most easily be done by doing a copy and paste on the preferred name for your lab for every instance where that component appears in the listing. Note that for the PEG example given above these may be separated by other components and several pastings may be needed. Once a column has been normalized repeat the sort operation on the next components column. If this column has salt precipitants a second sort level, to include both the cation and the anion, will be needed. Again, normalize the naming for all the components. Once done, repeat for the next column of components, and repeat until all names are consistent.

Once all the names have been made consistent, do a final sort on the numerical listing column. This will put the different screens back together again, with all the conditions in their correct order. Aside from attaching it to the correct column headers the files are now correctly set up. If desired the individual screens can be copied and pasted to their own spreadsheets so that they can be recombined in different combinations as used for the screening experiments.

Long range plans are to include software that reads in the manufacturers screen conditions files and, interactively with the user, outputs a normalized and correctly formatted screen data file for use with this software.

2. Scoring the Screen Results

We use 10-point scoring system as outlined in the Dinc et al. (2016) paper. A more granular scoring scale, having just 5 points, has been tried and found to give roughly comparable results. We use trace fluorescent labeling (TFL) (Pusey et al., 2015) for following out plates and for the final stage of scoring. Here a score of 4 is given to conditions having 'bright spots' that are not resolvable as a distinctly crystalline outcome. If TFL (or some variant) is not employed, then one can substitute a granular precipitate for a score of 4.

3. System Requirements

To install and use AED on a computer system, .NET Framework 4.0 or later version must be installed. Although there are no other requirements to run the program, it is recommended to have Microsoft Excel on the current computer system to analyze or organize input/output files.

4. Input File Format

To run the program, the input file must be prepared properly, and the file must contain listed features as in Table 1 (The features do not have to be in order). The type of all features (except scores) needs to be set as "Text," the score columns need to be "General."

Table 1 List of attributes of input file.

Well_Id	C1_M	C3_Cation	C4_Ph
B_Anion	C1_Ph	C3_Conc	C5_Anion
B_Cation	C2_Anion	C3_M	C5_Cation
Ph	C2_Cation	C3_Ph	C5_Conc
B_Conc	C2_Conc	C4_Anion	C5_M
C1_Anion	C2_M	C4_Cation	C5_Ph
C1_Cation	C2_Ph	C4_Conc	S_a
C1_Conc	C3_Anion	C4_M	S_b
			S_c

The data should not have any empty row; a sample case is displayed in Figure 1.

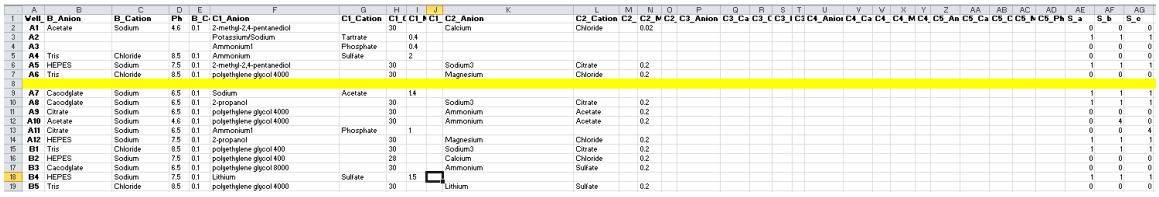

Figure 1: Empty rows in input file (not allowed).

The main sheet, which contains the data, should be named as “Sheet1” as in Figure 2.

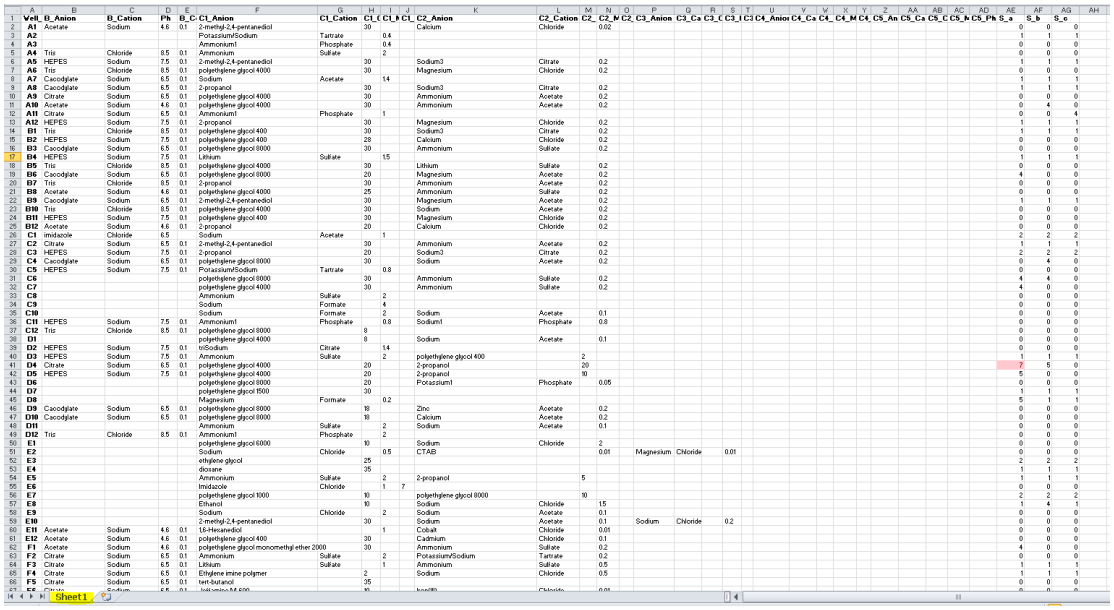

Figure 2: Sample input file

5. Output File Format and Interpretation

When the program finishes the AED analysis, it generates an excel output file with a name provided by the user into the selected directory. The file has two sheets: 1) “LIST_OF_CANDIDATES” that contains list of cocktails generated by AED analysis as in Figure 3) “LIST_OF_SCORES” that contains the list of ranking results of each reagent as in Figure 4. The list of candidate screen sheet’s format is very similar to the input screen sheet format. This sheet contains the following attributes listed in Table 2.

Table 2 List of attributes of output candidate cocktails.

Well_Id	**C1_M**	**C3_Cation**	**C4_Ph**
B_Anion	**C1_Ph**	**C3_Conc**	**C5_Anion**
B_Cation	**C2_Anion**	**C3_M**	**C5_Cation**
Ph	**C2_Cation**	**C3_Ph**	**C5_Conc**
B_Conc	**C2_Conc**	**C4_Anion**	**C5_M**
C1_Anion	**C2_M**	**C4_Cation**	**C5_Ph**
C1_Cation	**C2_Ph**	**C4_Conc**	**Rank**
C1_Conc	**C3_Anion**	**C4_M**	

The rank attribute has been calculated with the same formula in the AED journal paper. That is, the ranks provided in Figure 10 were based on the formula below.

For a candidate cocktail C that consists of precipitant p, buffer b, and salt s as reagents, let δ_p, δ_s and δ_b represent the scores of the cocktails having precipitant p, salt s, and buffer b for a given screen file, respectively. Let Δ represent all scores of the input file. Then, the significance ratio $\rho(\delta_r)$ is computed as, $\frac{\mu(\delta_p)}{\mu(\Delta-\delta_p)}$, $\frac{\mu(\delta_s)}{\mu(\Delta-\delta_s)}$, and $\frac{\mu(\delta_b)}{\mu(\Delta-\delta_b)}$, respectively.

Once $\rho(\delta_r)$ of each reagent is calculated, the program also provides the average rank of 3 components in the last column, in addition to $\rho(\delta_r)$s as in Figure 3.

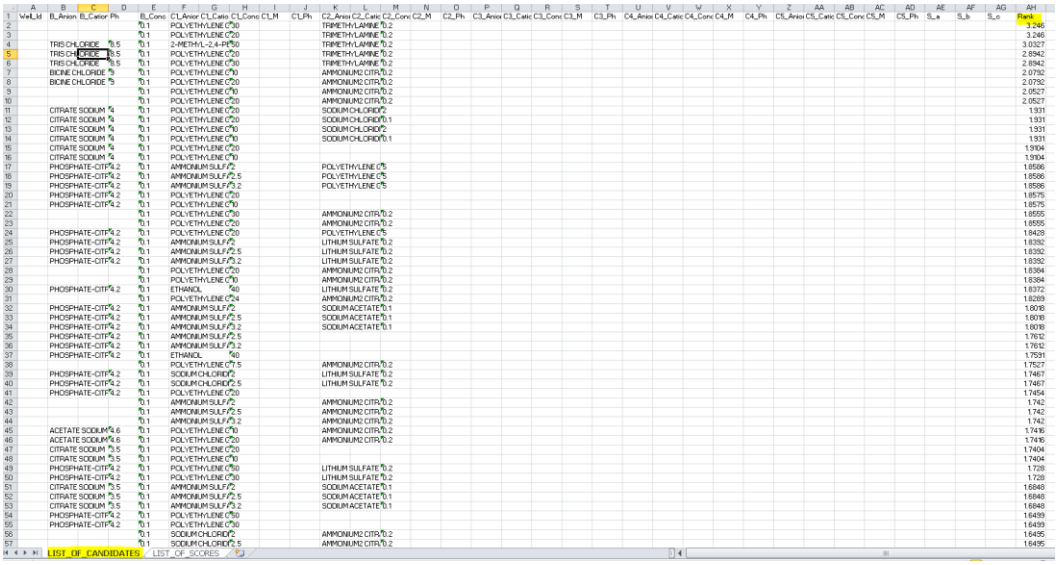

Figure 3 List of candidates

	A	B	C	D	E	F	G	H
1	Ph	Ph_Rank	Precipitant	Precipitant_Rank	Salt	Salt_Rank		
2	4	3.102610	POLYETHYLENE GLYCOL MONOMETHYL ETHER 550	2.27142857142857	TRIMETHYLAMINE N-OXIDE	6.51428571428571		
3	4.2	2.943925	POLYETHYLENE GLYCOL MONOMETHYL ETHER 2000	1.68	AMMONIUM2 CITRATE	2.34285714285714		
4	3.5	2.592760	POLYETHYLENE GLYCOL 6000	1.62857142857143	MAGNESIUM FORMATE	1.47692307692308		
5	9	1.623376	1,2-PROPANEDIOL	1.61632653061224	AMMONIUM SULFATE	1.33953488372093		
6		1.543742	POLYETHYLENE GLYCOL 1500	1.6	AMMONIUM1 PHOSPHATE	1.31428571428571		
7	9.5	1.337924	MAGNESIUM FORMATE	1.47692307692308	GLYCEROL	1.30909090909091		
8	8.5	1.250471	POLYETHYLENE GLYCOL 10000	1.37142857142857	POLYETHYLENE GLYCOL 1000	1.29230769230769		
9	4.5	1.1	AMMONIUM SULFATE	1.33953488372093	LITHIUM SULFATE	1.23428571428571		
10	5.5	1.085882	ETHANOL	1.33333333333333	AMMONIUM FORMATE	1.1265306122449		
11	6.2	0.959493	2-METHYL-2,4-PENTANEDIOL	1.33333333333333	SODIUM ACETATE	1.12207792207792		
12	7.5	0.835897	POLYETHYLENE GLYCOL 1000	1.29230769230769	POTASSIUM BROMIDE	1.07755102040816		
13	7	0.675874	SODIUM CHLORIDE	1.06175115207373	SODIUM CHLORIDE	1.06175115207373		
14	4.6	0.610596	POLYETHYLENE GLYCOL 400	1.00571428571429		1		
15			POLYETHYLENE GLYCOL 8000	0.917972350230415	POTASSIUM THIOCYANATE	0.876190476190476		
16			POLYETHYLENE GLYCOL 3350	0.850285714285714	LITHIUM CHLORIDE	0.685714285714286		
17			POLYETHYLENE GLYCOL 4000	0.555102040816327	MAGNESIUM CHLORIDE	0.649624060150376		
18					SODIUM CITRATE	0.647619047619048		
19					CADMIUM CHLORIDE	0.3		
20								
21								
22								
23								
24								
25								
26								
27								
28								

Figure 4 List of scores.

a. Elimination of Bad Combination of Candidate List

While the final candidate list is being created, if some of the forbidden chemicals appear in the same screen, they are removed from the list. In the current version of AED, the forbidden chemicals that are not allowed to appear together are listed in Table 3. According to the table, any chemical in list 1 cannot be used in a screen with any chemical in list 2.

Table 3 Forbidden combination list.

Chemical List 1	Chemical List 2
Phosphate	Calcium
Citrate	Magnesium
Citric	Zinc
Succinate	Iron
Malic	Cadmium
Malonic	Cadmium
Malonate	Manganese

6. Making Use of the Output Data

Despite our efforts, to date that computers have yet to learn chemistry, so some user effort is required at this point. The approach taken in our laboratory is given below. Other approaches may be employed, as long as the work for you.

The conditions are output as a listing of new condition mixtures, with those having the highest rankings at the top. The rankings are given in the column to the right. A second output sheet has the rankings for the individual components. Developing a new screen involves a lot of cycling between the two listings, if for no other reason than to ensure that the top-ranking components are all included and to a lesser extent those that are bottom ranked are not.

Selection of conditions is moderated by chemical reality. As our new screen conditions are formulated by mixing of stock solutions to a final volume of 0.5 or 1.0 mLs, we cannot, for example, titrate to a specific buffer pH. Thus a solution of 0.1M Sodium Acetate buffer, pH 4.6, plus 1.0 M trisodium phosphate, will NOT have a pH of 4.6 but will have a pH closer to that of the trisodium phosphate. This listing would be discarded.

Stock sodium citrate solution is 1.6 M. Precipitant conditions calling for 1.6 M sodium citrate at 0.1 M buffer, plus (for example) 10% PEG 400, cannot be made. The sodium citrate concentration would be reduced in this case to fit everything in.

Mixtures that are likely to result in salt crystals are discarded from consideration, as well as mixtures likely to result in phase separations such as high PEG and salt concentrations.

On occasion essentially, the same conditions will appear in different parts of the listing. For example, 0.1 M citrate buffer and 1.6 M citrate as precipitant vs. 1.6 M citrate as precipitant 1 and 0.1 M citrate as precipitant 2. While these would not be possible on a preparative basis they are also redundant and could be simple replaced by 1.6 M citrate.

Starting at the top of the list, conditions are selected after review on the basis of the above criteria. Typically, our output screens are 32 'families' of conditions where the concentration of the primary precipitant is varied over three conditions (32/3). Some guidance to the range of the primary precipitant concentrations is given by the range given in the output listing. One output condition is chosen, as an example 0.1 M buffer, 2.0 M ammonium sulfate, 5% PEG 4000. This would be used to generate three solutions, having 0.1 M buffer, 5% PEG 4000, and 2.0, 1.6, and 1.2 M ammonium sulfate. The variations in the ammonium sulfate concentration are not fixed, but varied depending on the experimenters judgement.

Our rule of thumb is to go the 32 families/3 conditions per family route for most proteins. If a protein is perceived as being more difficult we would go with 96 distinct conditions, or possibly 48 families of 2 conditions/family. Other variations are of course possible, such as 24/4, 16/6, 12/8, etc. An important consideration for going the limited grid screen, such as 32/3, route is that if crystals are obtained in all three conditions for a family then one can conclude that those are relatively robust, repeatable, crystallization conditions.

Whichever the approach the selected conditions are transferred to a new worksheet (we keep the associated priority scores with them) and expanded to the number of conditions as determined by the experimental requirements. Conditions are selected from the output table down to priority scores of > 1.0.

Once all the conditions have been filled additional columns are inserted and a pipetting table is generated, using the desired condition component concentrations and the component stock concentrations. This process is also used to catch those conditions that cannot be made. The final column calculates the amount of dH2O to be added (final volume-sum of the component volumes). If this is a negative volume, then adjustments are needed to one or more of the component concentrations.

References:

(Dinc, 2016) İ. Dinç, M. L. Pusey and R. S. Aygün, "Optimizing Associative Experimental Design for Protein Crystallization Screening," in IEEE

Transactions on NanoBioscience, vol. 15, no. 2, pp. 101-112, March 2016. https://doi.org/10.1109/TNB.2016.2536030

(Dinc, 2015) Dinç İ., Pusey M.L., Aygün R.S. (2015) Protein Crystallization Screening Using Associative Experimental Design. In: Harrison R., Li Y., Măndoiu I. (eds) Bioinformatics Research and Applications. ISBRA 2015. Lecture Notes in Computer Science, vol 9096. Springer, Cham https://doi.org/10.1007/978-3-319-19048-8_8

(Pusey, 2015) Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, *71*(7), 806-814.