International Journal of
Molecular Sciences

MDPI

Article

Renaissance Distribution for Statistically Failed Experiments

Roman Popov, Girish Karadka Shankara, Clemens von Bojničić-Kninski and Alexander Nesterov-Mueller *

Institute of Microstructure Technology, Karlsruhe Institute of Technology (KIT),
76344 Eggenstein-Leopoldshafen, Germany
* Correspondence: Alexander.Nesterov-Mueller@kit.edu

Received: 21 May 2019; Accepted: 30 June 2019; Published: 2 July 2019

Abstract: Much of the experimental data, especially in life sciences, is considered to be useless if it demonstrates a large standard deviation from the mean value. The Renaissance distribution, as presented in this study, allows one to extract true values from such statistical data with large noise. To obtain proof of the Renaissance distribution, high-throughput synthesis of deep substitutions for a target amino acid sequence was performed, and the known epitope was identified in assay with human serum antibodies. In addition, the Renaissance distribution was shown to approach the epitope affinity maturation by the deep alanine substitution. The Renaissance distribution may have an impact in the development of novel specific drugs.

Keywords: peptide arrays; statistics; deep substitutions

1. Introduction

Experimental science is based on well-established statistical tools used to estimate the quality of the data measured. The determination of the mean value with a relatively small standard deviation is considered a successful measurement. Nevertheless, there is much experimental data, especially in life sciences, that has statistically failed but may be "interesting". This failure is often caused by poorly defined experimental conditions or some processes that are barely understood. Such results are frequently observed, for example, in the case of novel multicomponent reactions [1], where the product yield is not optimized, or, in the case of combinatorial synthesis, where novel molecules are synthesized with unknown properties [2,3]. Frequently, enzymes demonstrate functional heterogeneity, which cannot be represented by employing classical bulk reactions and classical statistics [4]. For example, single-molecule experiments have revealed a dynamic disorder in the subsequent catalytic cycles of cholesterol oxidase [5], horseradish peroxidase [6], and lipase B [7]. From this point of view, the statistically unreliable experimental data may include valuable scientific information that can help improve the experiment or understanding of those processes. The generally accepted opinion is that the scientist has to improve his experiment to get better statistics before presenting the results to the scientific community, but this very step remains a bottleneck in comparison with the modern communication and possibilities of generating a large amount of experimental data. Dealing with their own imperfect experiments, the scientist is limited in communicating about them because he will not publish statistically failed results.

In this study, we presented a novel synthesis of substitutional peptide arrays and, using this example, proposed a novel statistical distribution to interpret the data measured. The distribution proposed in paper is called "Renaissance," by analogy with the period in European history which is related to several Italian cities, but that would not exhibit any Renaissance features in the case where the development level over all European cities of that time is averaged. The Renaissance distribution is a novel statistical method for cases when the standard deviation from the mean value is higher than 30%. In general, the Renaissance distribution is suitable for automatically searching for artifacts and providing a reliable comparison of the signals by varying parameters of a system with a high noise.

2. Results and Discussion

The principle of the stochastic array synthesized in this study is shown in Figure 1. The stochastic array delivers all possible substitutions of a specific amino acid sequence with amino acid alanine. The synthesis of all substitutional sequences occurs by stochastically assembling of the amino acid particles (Figure 1a) on the cavities (Figure 1b) functionalized with amino groups.

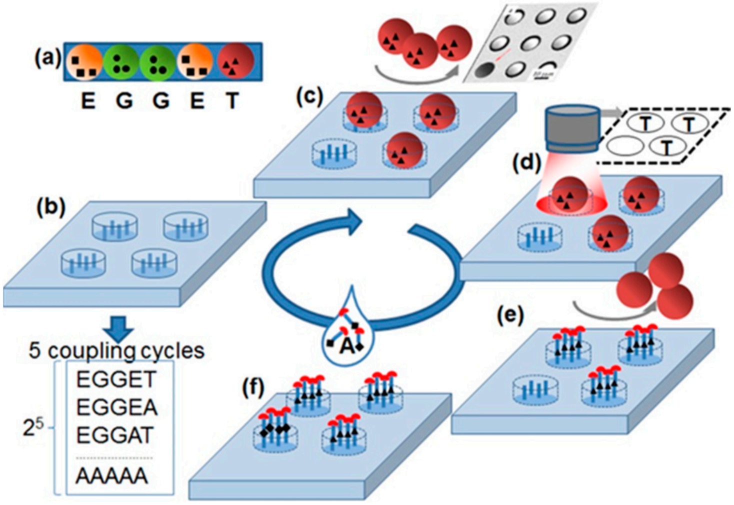


Figure 1. Stochastic substitutional arrays with amino acids particles. (**a**) Each amino acid particle carries a special amino acid with protecting groups. (**b**) Functionalized substrate is exposed to (**c**) one kind amino acid particles. The particles fill in the cavities stochastically. (**d**) The particle pattern is detected with a camera. (**e**) The amino acids are extracted from the particles and coupled inside the cavities via peptide bonds. The particles are removed from cavities. (**f**) The slide is exposed to a solution containing the amino acid alanine. After the removal of protecting groups (red half circles), a new cycle can be started. The repetition of 5 cycles results in the synthesis of peptides with all possible Ala substitutions in original sequences TEGGE.

The shape of the cavities was chosen so that only one particle fits inside. The particle patterns were detected with a fluorescent scanner equipped with a visible light filter (Figure 1c,d). Then, amino acids were extracted from the particles and coupled to the bottom of the corresponding cavities, building peptide bonds with amino acid groups on the surface. After coupling, the particles were removed from the cavities with an ultrasound (Figure 1e). Subsequently, in the wells that did not capture particles, the amino acid alanine was coupled out of the solution. After the coupling of the amino acids from the particles and solution, the protective groups were removed so that the cavities were ready for the next coupling step. Using this principle, a stochastic alanine substitution array was synthesized for the amino acid sequence KEVPALTAVETGAT. This peptide was reported to be reactive for 90% of the European blood sera [8].

Theoretically, 2^{11} = 2048 different peptides should be generated according to all possible substitutions of the target peptide KEVPALTAVETGAT with alanine. By analyzing the particle deposition, only 1142 different peptides were displayed on the chip. For the target sequence, 2672 copies were identified.

Figure 2 shows the distribution of the fluorescent signal intensity of the peptide spots across these copies, whereby the peptides are arranged according to the descending signal. This curve was approximated with the logarithmic function:

$$y(\mathrm{x}) = -a\cdot\ln(\mathrm{x}) + b, \tag{1}$$

where $a = 92$, $b = 834$, and the coefficient of determination is $R^2 = 0.895$. Formula (1) represents a general function that is defined as the Renaissance distribution. In an ideal case, all the signals would have the same fluorescent intensity. However, because of the statistically non-ideal amino acid extraction from the particles, the majority of signals are close to the chip surface background value of approx. 200 a.u., and only few high-intensity signals are demonstrated by the first 50 peptides. Formula (1) fit the signal distributions for all substitutions of the target sequence very well, with coefficients of determination > 0.8. Interesting to note is that removal of the signals with maximal intensity significantly increased the coefficient of determination to $R^2 = 0.97$. Such deleted points will be considered as artifacts.

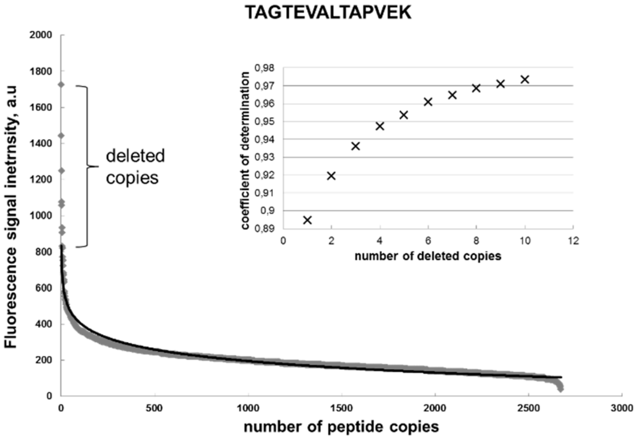


Figure 2. Distribution of the fluorescent signal intensity of the identical peptides (gray diamonds) and its logarithmic approximation (black) on a chip. The curves represent the Renaissance distribution. $y = -92\ln(x) + 834$; $R^2 = 0.895$. The inlet shows an increase of the R^2 by deleting the copies from the original distribution of the fluorescent intensity.

Apparently, the estimation of the mean value for the signal intensity over all peptide copies in Figure 2 will give a false result, which is close to the background value. Instead of using mean value, the coefficient b in the Formula (1) was considered to represent the true intensity signal. In this case, the term $-a\cdot\ln(\mathrm{x})$ is a negative entropy responsible for inefficient coupling. The crucial feature of the Renaissance distribution (1) is that the true value, parameter b, is independent of the number of the trials. If a set of x copies will be reduced, that would correspond to the set of x^{β}. Thus, the parameter β will only modify the negative information entropy a to $a\cdot\beta$. The smaller the number of trails, the higher the coefficient *a will be*. The differential presentation of Formula (1)

$$\delta y = -\delta x/x \tag{2}$$

reveals the meaning of the Renaissance distribution. Small changes of the feature parameter y of a system is inversely proportional to the number of the non-ideal system states. In particularly, this law means that the appearance of a small number of the non-ideal states in the system with the zero entropy has a significant influence on the feature parameters. Vice versa, if the number of non-ideal states is already large, the appearance of the additional non-ideal states will have only marginal influence on the feature parameter y. Thus, the Renaissance distribution appears both as a statistical distribution and as a law that describes the decay of a state with the zero entropy.

To prove the Renaissance distribution, we applied an algorithm to all the substitutional peptides to eliminate the copies (artifacts) until the remaining points could be fitted using the Renaissance distribution, with $R^2 > 0.95$ for each specific sequence (Figure 3).

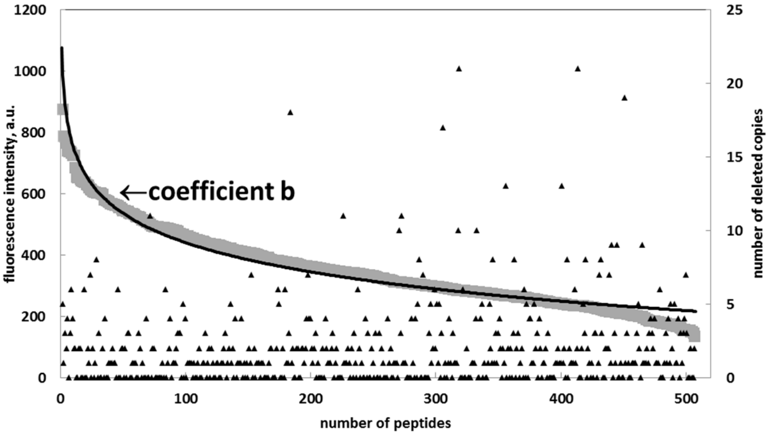
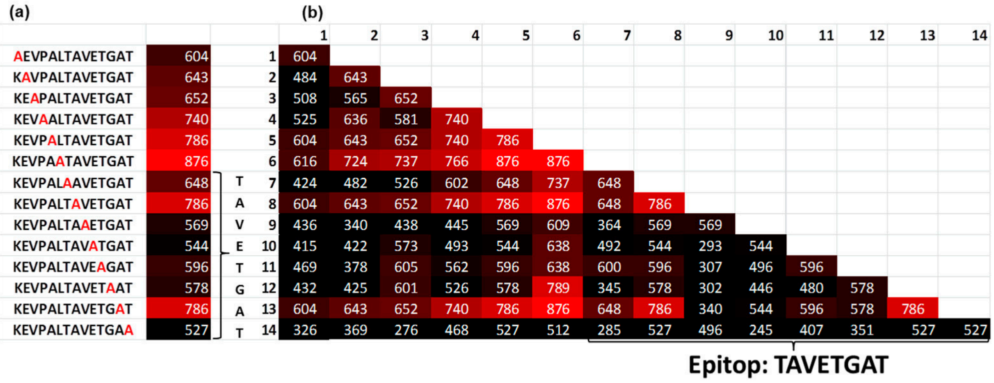


Figure 3. Distribution of coefficient b for peptides representing multiple alanine substitutions of the target sequence TAGTEVATLAPVEK (diamonds) and its approximation (black) with a Renaissance distribution: a = −133, b = 1051, and $R^2 = 0.96$.

Using the Renaissance distribution (1), the coefficient *b* (true value) was estimated for the single alanine and double alanine substitutions of the target sequence to obtain the binding motif of the serum antibodies (Figure 4). The alanine substitutions revealed the binding motif **KXXXXXXVETGXT**. This result corresponds to the known epitope **LTAVETG**, which is part of the poliovirus and can also be found in the human epitope database (iedb.org) [9]. The location of the target peptide in the capsid protein VP1 of the poliovirus [10,11] is shown in the Supplementary Figure S5 online. By contrast, the estimation of the fluorescent signals using classical distribution, which is based on the mean values over all peptide copies, delivers significantly different results in which the fluorescent signal of the target peptide was not distinguished from the background.


(a)
(b)
Epitop: TAVETGAT

				1	2	3	4	5	6	7	8	9	10	11	12	13	14
AEVPALTAVETGAT	604		1	604													
KAVPALTAVETGAT	643		2	484	643												
KEAPALTAVETGAT	652		3	508	565	652											
KEVAALTAVETGAT	740		4	525	636	581	740										
KEVPALTAVETGAT	786		5	604	643	652	740	786									
KEVPAATAVETGAT	876		6	616	724	737	766	876	876								
KEVPALAAVETGAT	648	T	7	424	482	526	602	648	737	648							
KEVPALTAVETGAT	786	A	8	604	643	652	740	786	876	648	786						
KEVPALTAAETGAT	569	V	9	436	340	438	445	569	609	364	569	569					
KEVPALTAVATGAT	544	E	10	415	422	573	493	544	638	492	544	293	544				
KEVPALTAVEAGAT	596	T	11	469	378	605	562	596	638	600	596	307	496	596			
KEVPALTAVETAAT	578	G	12	432	425	601	526	578	789	345	578	302	446	480	578		
KEVPALTAVETGAT	786	A	13	604	643	652	740	786	876	648	786	340	544	596	578	786	
KEVPALTAVETGAA	527	T	14	326	369	276	468	527	512	285	527	496	245	407	351	527	527

Figure 4. (**a**) Single and (**b**) double Ala substitutions of the target peptide KEVPALTAVETGAT. The number indicates the fluorescence intensity. In the table (**b**), the position (x,y) means that Ala was placed in position x and position y in the target peptide. The intensity of the signals reveals a binding motif **KXXXXXXVETGXT** (**c**), indicating the location of the target peptide in the capsid protein VP1 of the poliovirus.

The Renaissance distribution fit the fluorescent signals from different peptides very well, with a coefficient of determination of $R^2 = 0.96$. Interestingly, that deep alanine substitutions cause the decay of the epitope affinity, according to the Renaissance distribution. The deep alanine substitution reveals that the difference in the fluorescence signal between epitopes and its substitutions was not more than

40%. This fact could explain the off-target activity of antibodies which are simultaneously applied for specific labeling of the proteins. From this point of view, the Renaissance distribution can be used for the development of the novel specific drugs in silico. First, to increase the affinity of the initial scaffold, novel building blocks should be tested according to the minimization of the binding energy. Then, only those building blocks should be accepted, which increases the affinity of the scaffold along the Renaissance distribution. Such an approach would mimic the natural maturation of the epitopes.

3. Materials and Methods

3.1. Synthesis of Peptide Sequences

The full-size microstructured substrates (75 mm × 25 mm × 1 mm) were made of fused silica using photolithography and reactive ion etching (RIE). The array of cylindrical microwells, with diameter of 12 μm and the depth of 9 μm, has a pitch size of 20 μm.

The crosslinked poly(methyl methacrylate) (PMMA) microbeads (MICROBEADS®) were selected as solid carriers for the amino acid derivatives. They were manufactured by emulsification polymerization and crosslinking at 3%. The microspheres had a mean diameter of 10 μm with an extremely narrow size distribution. The coefficient of variation was declared to be less than 5%. The commercially available amino acid derivative (fluorenylmethoxycarbonylamino acid pentafluorophenyl ester) was used as monomer. The amino acid derivative was introduced into the polymer matrix of the microbeads. Crosslinked PMMA microspheres (1 g) were dispersed in 5 mL of the amino acid derivative solution in dichloromethane (DCM). The mass per volume fraction (%, (*w/v*)) of amino acid in the DCM solution was adjusted for each type of monomer. The dispersion was gently stirred until a paste-like medium was obtained. The paste underwent drying overnight at normal conditions in an open vessel in a fume hood. The resulting dry mass of microbeads was milled in a falcon tube with two metal spheres (Ø 5 mm) on a vortex shaker.

Prior to applying the amino acid particles, the functional surface underwent Fmoc (Fluorenylmethoxycarbonyl) deprotection in a solution of 20% (*v/v*) piperidine in Dimethylformamide (DMF) for 30 min at room temperature. The residues of the deprotection solution were removed by washing in DMF for 5 min twice, in methanol for 3 min twice, and in DCM for 1 min once. The slide was dried in a flow of argon.

The deposition of the microbeads was performed in two steps. First, the dry mixture of the microbeads was spread over the surface of the microstructured substrate with a soft lint-free tissue. Although the microstructures were completely filled with the microbeads, the top surface of the substrate still contained a monolayer of excessive microbeads. This excess was removed in the second step with the flow of compressed air applied tangentially to the surface of the substrate. The air flow totally removed the excessive microbeads from the top surface, whereas the microwells remained filled. The imaging of the microbeads' deposition pattern was performed using a fluorescence scanning with optical density filters.

The extraction of the amino acid derivatives from the microbeads took place in a saturated vapor of organic solvent inside a special chamber. The slide holders with the substrates were inserted into the chamber for 1 min each. After taking the slides out of the chamber, the process was repeated 5 times to enhance the extraction of the amino acid derivatives from the microbeads. The subsequent coupling step was performed in an oven under argon for 60 min at 90 °C.

Removal of the microbeads from the microwells was performed in several steps by placing the substrate in a mixture of 5% (*v/v*) MEA (2-aminoethanol) in acetone and exposing it to acoustic waves in an ultrasonic bath for 2 min twice. Alanine (Fmoc-protected, OPfp (pentafluorophenyl)-activated) was coupled to the remaining free amino groups of the substrate from 0.1 M solution in DMF for 60 min at room temperature.

The remaining free amino groups were capped in a solution of 10% (*v/v*) acetic anhydrate and 20% (*v/v*) DIPEA (N,N-Diisopropylethylamine) in DMF for 10 min at room temperature repeated twice. The residues of the capping solution were removed by washing in DMF for 5 min twice, in methanol for 3 min twice, and in DCM for 1 min once. The slide was dried in a flow of argon.

The steps of Fmoc deprotection, amino acid coupling, and capping were performed multiple times until the full-length peptides were synthesized.

3.2. Immunostaining

The array was incubated with a blood serum from a healthy donor diluted 1:1000 in PBS-T (Phosphate Buffered Saline with Tween) overnight at 4 °C and then stained with the secondary anti-human-IgG antibodies. The fluorescent signals from the spots stained with secondary antibodies were read out with a fluorescent scanner.

3.3. Statistical Analysis

The general function of the Renaissance distribution is presented by a logarithmic formula (1). To fit this formula, a Python program (Python 3.7) was used. The Python codes are available as a separate file in the Supplementary Materials online. The initial data included approximately 78,000 peptides and their corresponding immunostaining signals. The program (a) separates unique sequences; (b) generates the Renaissance distribution for each unique sequence and calculates coefficients a (negative entropy) and *b* (true signals) with a defined minimum coefficient of determination $R^2{}_{min}$. The increasing of the initial R^2 (if $R^2 < R^2{}_{min}$) is realized by searching for single peptide copies that are responsible for the low R^2 and their deletion. The program also shows how many copies were deleted to reach $R^2{}_{min}$ and the resulting R^2. Python is an open source programming language and was download at https://www.anaconda.com/distribution/.

In the scope of this program, a common method of logarithmic fitting was used: all initial signals x_i were replaced with $X_i = \ln(x_i)$ and a linear regression with the (X_i, y_i) was performed to get a logarithmic fitting. An additional simple example of fitting experimental data with a large standard deviation using the Renaissance distribution is demonstrated in the Supplementary Materials online. All data processing (calculation of mean value, standard deviation, and logarithmic approximation) was performed using an Excel program (Supplementary Figures S1–S4).

4. Conclusions

The stochastic high-throughput synthesis of deep substitutions for a target amino acid sequence was presented. Deep substitution peptide chips were used for studying interactions with antibodies in a high-density-array format. Renaissance distribution was developed to analyze the fluorescent signals from the peptide–serum antibody interactions. Renaissance distribution differs substantially from the statistics based on calculations of the mean values. This allows for identification of the artifact signals which deviate from the Renaissance distribution. In contrast to the other statistical distributions, the Renaissance distribution represents the law of the decay of states with zero entropy. Its application was demonstrated for the known polio epitope.

The Renaissance distribution may have an impact in life sciences and organic chemistry for the analysis of experimental data produced under poorly defined stochastic conditions with a large number of trials. The example with deep substitutions for the target epitope demonstrates the principle of Renaissance increasing of the binding affinity, which can be used in the development of novel specific drugs.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/13/3250/s1.

Author Contributions: R.P. developed the amino acid particles and protocols to particle extraction, performed data analysis. G.K.S. performed surface functionalization and the arrays, C.v.B.-K. performed the optical decoding, A.N.-M. performed data analysis, conceived the Renaissance distribution and wrote the paper.

Funding: This work was supported by DFG (no. AOBJ 655892) and BMBF (no. 03INT502BB). We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ganem, B. Strategies for innovation in multicomponent reaction design. *Acc. Chem. Res.* **2009**, *42*, 463–472. [CrossRef] [PubMed]
2. Marsault, E.; Peterson, M.L. Macrocycles are great cycles: Applications, Opportunities, and challenges of synthetic macrocycles in drug discovery. *J. Med. Chem.* **2011**, *54*, 1961–2004. [CrossRef] [PubMed]
3. White, C.J.; Yudin, A.K. Contemporary strategies for peptide macrocyclization. *Nat. Chem.* **2011**, *3*, 509–524. [CrossRef] [PubMed]
4. Liebherr, R.B.; Renner, M.; Gorris, H.H. A Single Molecule Perspective on the Functional Diversity of in Vitro Evolved β-Glucuronidase. *J. Am. Chem. Soc.* **2014**, *136*, 5949. [CrossRef] [PubMed]
5. Lu, H.P.; Xun, L.; Xie, X.S. Single-molecule enzymatic dynamics. *Science* **1998**, *282*, 1877–1882. [CrossRef] [PubMed]
6. Edman, L.; Földes-Papp, Z.; Wennmalm, S.; Rigler, R. The fluctuating enzyme: A single molecule approach. *Chem. Phys.* **1999**, *247*, 11–22. [CrossRef]
7. Velonia, K.; Flomenbom, O.; Loos, D.; Masuo, S.; Cotlet, M.; Engelborghs, Y.; Hofkens, J.; Rowan, A.E.; Klafter, J.; Nolte, R.J.M.; et al. Single-enzyme kinetics of CALB-catalyzed hydrolysis. *Angew. Chem. Int. Ed.* **2005**, *44*, 560–564. [CrossRef] [PubMed]
8. Weber, L.K.; Palermo, A.; Kugler, J.; Armant, O.; Isse, A.; Rentschler, S.; Jaenisch, T.; Hubbuch, J.; Dubel, S.; Nesterov-Mueller, A.; et al. Single amino acid fingerprinting of the human antibody repertoire with high-density peptide arrays. *J. Immunol. Methods* **2017**, *443*, 45–54. [CrossRef] [PubMed]
9. Immune Epitope Database. Available online: https://www.iedb.org/ (accessed on 4 April 2019).
10. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics". *J. Mol. Graph.* **1996**, *14*, 33–38. [CrossRef]
11. Capsid Protein VP1 of the Poliovirus, Protein Data Bank. Available online: http://www.rcsb.org/structure/3J8F (accessed on 21 June 2019).

International Journal of *Molecular Sciences*

Supplementary Materials

Article

Renaissance Distribution for Statistically Failed Experiments

Roman Popov, Girish Karadka Shankara, Clemens von Bojničić-Kninski and Alexander Nesterov-Mueller *

Institute of Microstructure Technology, Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Germany; roman.popov@kit.edu (R.P.); girish.shankara@kit.edu (G.K.S.); clemens.bojnicic-kninski@kit.edu (C.B.-K.);

* Correspondence: Alexander.Nesterov-Mueller@kit.edu (A.N.-M.)

An example of Renaissance distribution with EXCEL

1. Open an a EXCEL sheet. Copy your data into a column;
2. Generate the graph (Fig. S1, the signals from 100 identical experiments have been taken);
3. Calculate the mean value M and its standard deviation STDV (For the data in Figure S1: M = 173.7; STDV = 106.2).
   The experiment has failed statistically because STDV is very high.
4. Apply to the column the command "Sort and Filter" and "from largest to smallest". The graph gets the form presented in Figure S2.

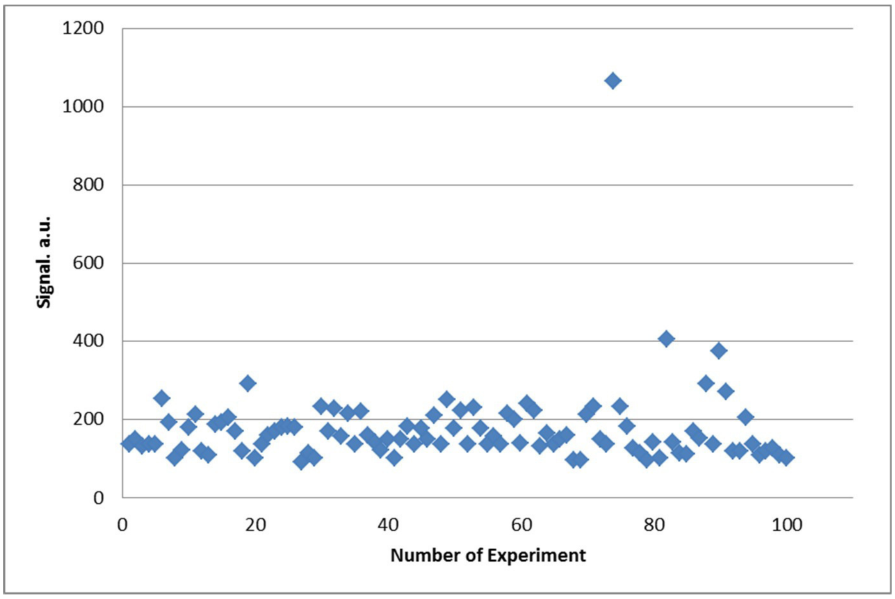


Figure S1: The distribution of the signal over 100 identical experiments.

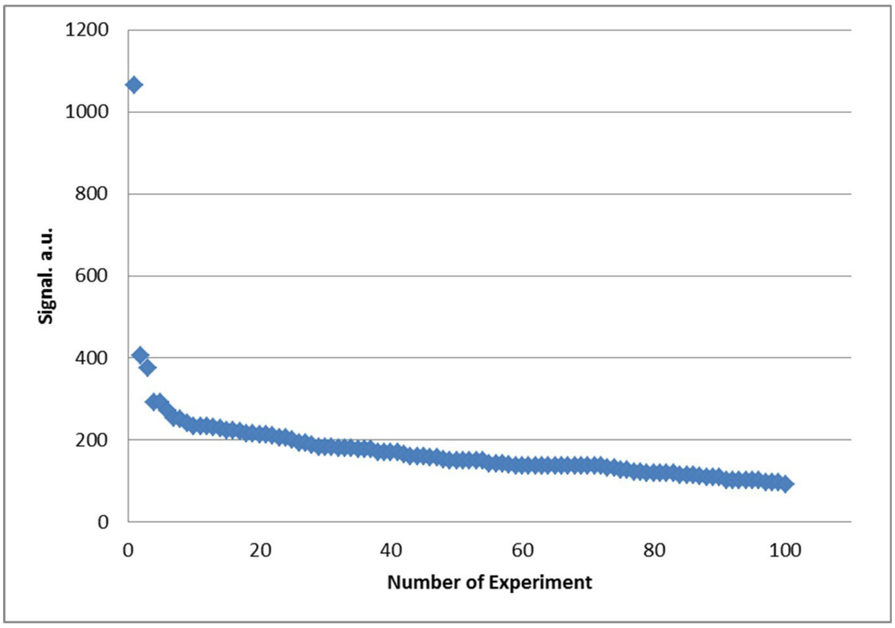


Figure S2: The distribution of the signal over 100 identical experiments after sorting the signal from Figure 2 "from largest to smallest".

5. Open the menu with the right mouse click on the dotted line and choose the button "Insert trend line". A window "Trend line options" appears. Choose "logarithmic " and the last two options "Show the formula in diagram" and "Show the approximation rate in diagram". After performing these options, Renaissance distribution is generated (Figure S3, black line).

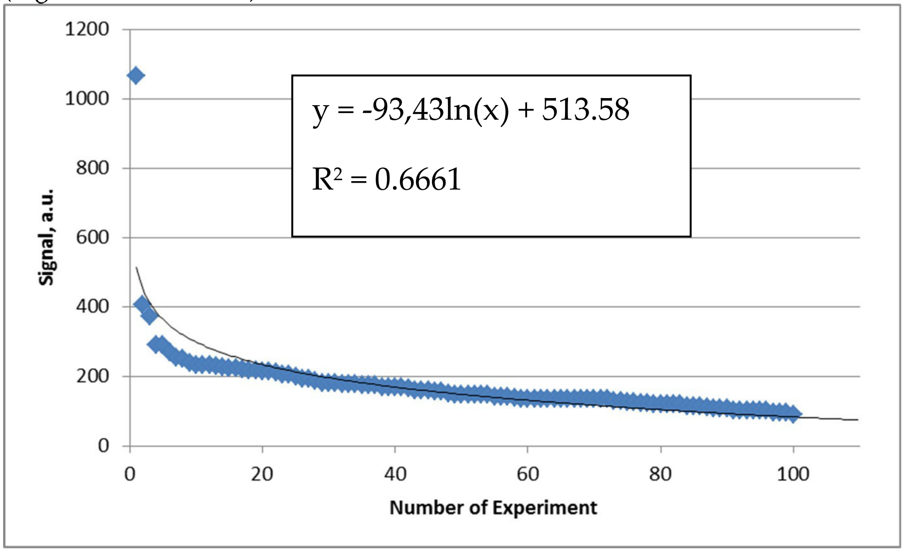


Figure S3: The distribution of the signal over 100 identical experiments and corresponding Renaissance distribution (black line) with $R^2 = 0.66$. ($R^2 \in (0:1)$).

6. Improve Renaissance approximation – R^2 by deleting critical points. One point S = 1065 is identified as artifact. Its deletion leads to significantly better approximation (Figure S4): $R^2 = 0.9719$.

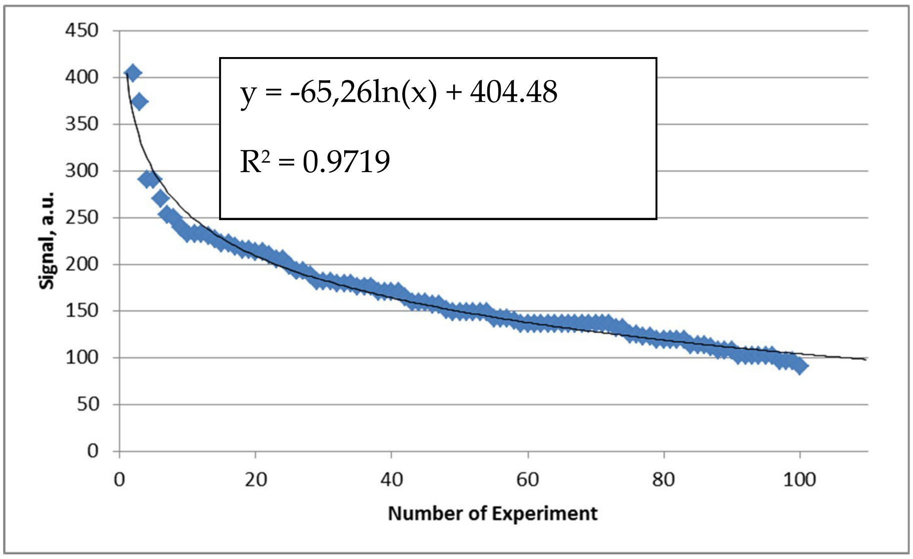


Figure S4: The distribution of the signal over 99 identical experiments and corresponding Renaissance distribution (black line) with $R^2 = 0.9719$.

7. Thus, the truth value is 404.48 with a approximation precision $R^2 \approx 0.97$. In contrast, the mean value of the original signals is 173.7 with standard deviation of 106.2. While the deletion of the point S = 1065 (artifact) from the original data is justified by reaching higher R^2 in the case of Renaissance distribution, its deletion would be a voluntary action in the case of the classical statistics based on the calculation of mean value and its deviations.

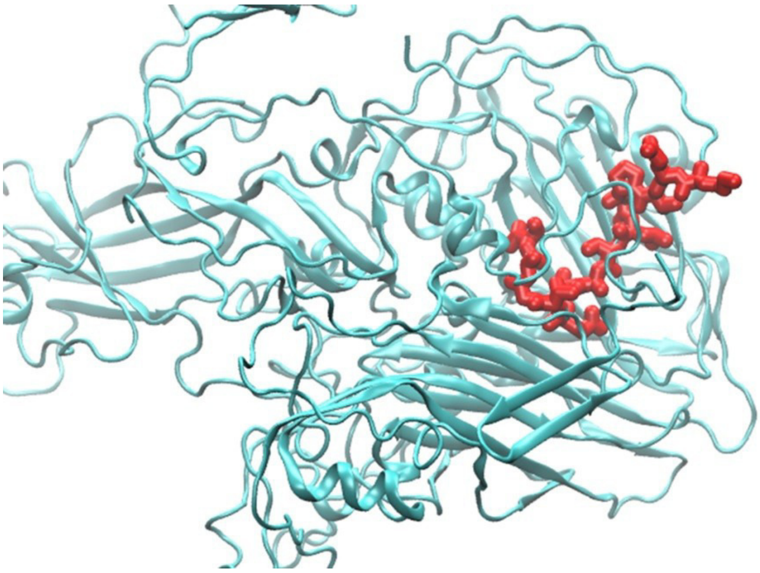


Figure S5: The location of the target peptide in the capsid protein VP1 of the poliovirus. The location was obtained using VMD: Visual Molecular Dynamics [10]. The 3D structure of the capsid protein VP1 of the poliovirus is available online in the Protein Data Bank [11].