BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Online tool for the discrimination of equi-distributions

Thorsten Pöschel*, Cornelius Frömmel and Christoph Gille

Address: Charité, Institut für Biochemie, Monbijoustraße 2, D-10117, Berlin, Germany

Email: Thorsten Pöschel* - thorsten.poeschel@charite.de; Cornelius Frömmel - cornelius.froemmel@charite.de; Christoph Gille - christoph.gille@charite.de

* Corresponding author

Published: 21 November 2003

BMC Bioinformatics 2003, **4**:58

Received: 19 June 2003

Accepted: 21 November 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/58

Abstract

Background: For many applications one wishes to decide whether a certain set of numbers originates from an equiprobability distribution or whether they are unequally distributed. Distributions of relative frequencies may deviate significantly from the corresponding probability distributions due to finite sample effects. Hence, it is not trivial to discriminate between an equiprobability distribution and non-equally distributed probabilities when knowing only frequencies.

Results: Based on analytical results we provide a software tool which allows to decide whether data correspond to an equiprobability distribution. The tool is available at http://bioinf.charite.de/equifreq/.

Conclusions: Its application is demonstrated for the distribution of point mutations in coding genes.

Background

Assume a set of certain events occur with frequencies M_i, $i = 1 \ldots N$, with $\sum_{i=1}^{N} M_i = M$, e.g., M_i = {4, 5, 2, 3, 2, 9,3,3,5,12,4, 6,4,... }. We ask the question whether the events obey an equiprobability distribution $p_i \equiv 1/N$. According to the general definition of probabilities

$$p_i = \lim_{M\to\infty} \frac{M_i}{M}, \qquad (1)$$

for an equiprobability distribution and for large sample size M it is expected to find each of the events approximately $M_i \equiv M/N$ times. For finite sample size, however, the frequencies M_i may deviate considerably from this value (Fig. 1).

The deviation from the equidistribution becomes particularly obvious if we order the events according to their rank, i.e., the most frequently occurring event appears left at the abscissa, then the next frequent, etc. (Fig. 2).

If we conclude naïvely from the observed frequencies to the probabilities, i.e., if we assume $p_i/p_j = M_i/M_j$, in the extreme case $M_{100} = 3$ we end up with a relative error of 70%. In other words, from the frequencies measured in an experiment as shown in Figs. 1 and 2, it might be erroneously concluded that the events are strongly non-equally distributed.

Using the methods of statistics we can generate (predict) the rank ordered frequency distribution for given N and M under the precondition that the events are equidistributed [1]. The predicted frequency distribution can then be compared with the distribution as measured in an

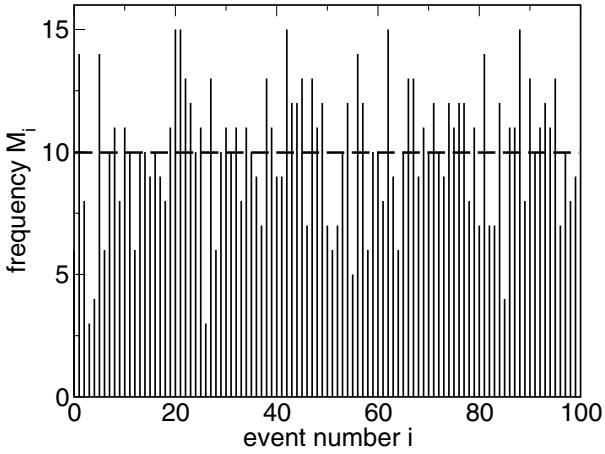


Figure 1
Histogram of frequencies of $M = 1000$ events according to an equiprobability distribution $p_i = 1/N = 1/100$. The dashed line displays the expectation value.

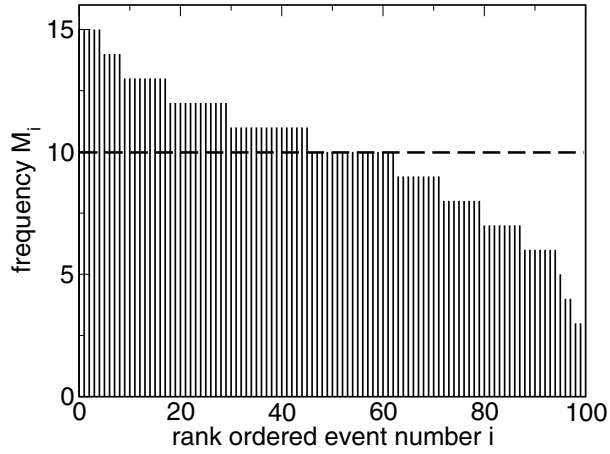


Figure 2
Same data as in Figure 1 but in rank order. From the figure it might be erroneously concluded that the events do not obey an equidistribution. The distribution is deformed, however, exclusively due to finite sample effects.

experiment with the same values of M and N. From the comparison it can be judged whether the events in the experiment obey an equidistribution.

Following this procedure we describe a tool which helps to decide whether a given set of frequencies complies with an equidistribution. For demonstration the tool is applied to the distribution of point mutations in human genes.

Implementation

The numerical tool is available via the web address http://bioinf.charite.de/equifreq/. The underlying kernel program which computes the most probable frequency distribution is implemented in *C++* and the user interface is written in PHP. The program source is available at this address.

Results and Discussion

Mathematical method

We want to sketch briefly the derivation of the basic formula: Assume we distribute M balls over N urns according to an equidistribution. The probability $p(k_i, i)$ to find k_i urns filled each with exactly i balls is given by

$$p(k_i, i) = \frac{M!}{N^M} \sum_{j=k_i}^{\lfloor M/i \rfloor} (-1)^{(j-k_i)} \binom{j}{k_i} \frac{(N-j)^{(M-ji)}}{(i!)^j (M-ji)!}, \qquad (2)$$

where $\lfloor x \rfloor$ denotes the integer of x.

Note that the probability to find a number of k_i urns which contain *exactly i* balls is different form the probability to find the number of urns which contain *at least i* balls which is a simple textbook problem, whereas the derivation of Eq. (2) requires quite involved algebra. The relation between both probabilities is provided by the exclusion-inclusion principle [2,3]. For our purpose we need the number $\langle K_i \rangle$ of urns filled with i balls which are found on average, i.e., we need the first moments of the probabilities Eq. (2). These values can be found in closed form applying the method of generating functions for the descending factorial moments. The averages $\langle K_i \rangle$ have been derived in a different context earlier, the details of the derivation can be found in [4,1]:

$$\langle K_i \rangle = N \binom{M}{i} \frac{1}{N^i} \left(1 - \frac{1}{N}\right)^{(M-i)}. \qquad (3)$$

As an interesting detail of the solution, the average number of filled urns is given by the total number of urns minus the number of empty ones, $N^* = N - \langle K_0 \rangle$, i.e. [1],

$$\frac{N^*}{N} = 1 - \left(1 - \frac{1}{N}\right)^M \approx 1 - \exp\left(-\frac{M}{N}\right). \qquad (4)$$

Obviously, for small M (numbers of balls) there is a significant number of urns which, on average, stay empty. Translating back to the language of biology we come to a surprising result: given a population of $N = 1000$ species. If we investigate a number of $M = 5000$ individuals, from Eq. (4) we obtain $N^* \approx 993.3$, i.e., about 7 species are

never found, although from naîve reasoning one expects each species occurring about 5 times.

The moments $\langle K_i \rangle$ given in Eq. (3) allow to reconstruct the rank ordered frequency distribution since they describe how many, on average, events do not occur (zero times), how many occur once, twice, etc. Hence, the desired rank ordered frequency distribution reads finally

$$M_i^{\text{theo}} = \begin{cases} 0 & \text{for} \quad N \geq i > N - \langle K_0 \rangle \\ 1 & \text{for} \quad N - \langle K_0 \rangle \geq i > N - \langle K_0 \rangle - \langle K_1 \rangle \\ & \ldots \\ j & \text{for} \quad N - \sum_{k=0}^{j-1} \langle K_k \rangle \geq i > N - \sum_{k=0}^{j} \langle K_k \rangle . \end{cases} \quad (5)$$

We apply Eq. (5) to predict the frequency distribution which arises from an equidistribution for different sample sizes *M* and compare with direct numerical simulations, s. Figs. 3, 4. The predictions due to Eq. (5) agrees well with the numerical experiment.

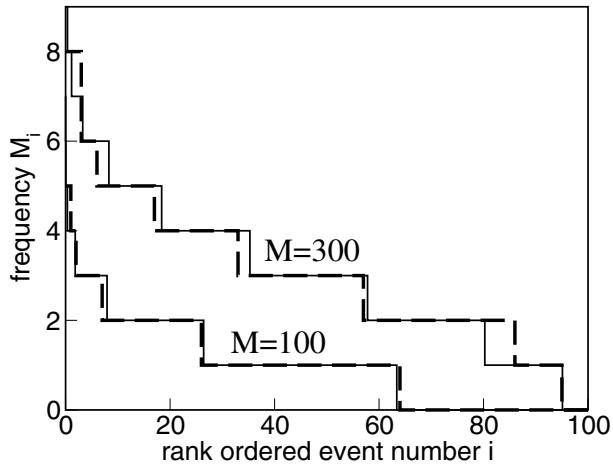


Figure 3
Rank ordered frequency distribution for *N* = 100 equally distributed events for different sample size *M*. The solid line shows the distribution as predicted by Eq. (5), the dashed line shows the distribution of independently drawn equidistributed random numbers from the interval [1, *N*].

Exploration of experimental data

The theoretical distribution of frequencies due to Eq. (5) can be compared with experimentally obtained frequencies. From the distance between both (rank ordered) frequency distributions we can conclude whether the experimental data obey an equidistribution. To this end we have elaborated a web based tool http://bioinf.charite.de/equifreq/. The user interface offers four alternative input masks which differ in the way the input file is generated:

(1) The measured frequencies of each species M_i are given directly.

(2) The number of species *N* and the total number of individuals *M* are specified. Each individual is assigned a species by chance.

(3) As for (2) the rank ordered frequencies are computed but with the generalization that each species is assigned an individual probability. The theoretical basis for this computation is not given here but will be published elsewhere [5].

(4) The last input mask is intended for the investigation of the spatial distribution of point mutation in genes which is presently the most specialized application of the described program.

The program computes the expected frequency distribution due to Eq. (5) with the assumption that the species obey an equiprobability distribution. Three output files are generated: *freq, ktheo* and *kexp*. The file *freq* contains the rank ordered frequencies as generated from the input data set (cases (1) and (4)) or randomly due to an equiprobability distribution (case (2)) or a general distribution (case (3)). *ktheo* contains the moments $\langle K_i \rangle$ for each rank *i*, i.e. the expected number of individuals occurring *i* times, due to Eq. (3) for given numbers *N* of species and *M* of individuals. For cases (1) and (4) the values of *M* and *N* are extracted from the input data, for (2) and (3) they are provided by the user. (Note that these expectation values are real numbers in general.) The third column of line *i* contains the value $M - \sum_{j=0}^{i} \langle K_i \rangle$. The last file, *kexp* contains the same data as *ktheo*, but based on the input data (cases (1) and (4)) or on the randomly generated data (cases (2) and (3)), respectively. Besides the pure output files the program generates a number of visualizations (see section *Example: Distribution of point mutation in genes*). In order to compare the experimental data with the mathematical prediction both, the experimental data and the theoretical data, are plotted in the same chart. Congruence of both curves indicates that the experimental data obey an equidistribution (case (2)) or the specified distribution (case (3)), respectively.

It may occur that the curve of the rank ordered experimental data decays significantly slower than the corresponding theoretical curve due to Eq. (5). Since there is no distribution more homogeneous than the equidistribution this situation may occur either as a rare fluctuation

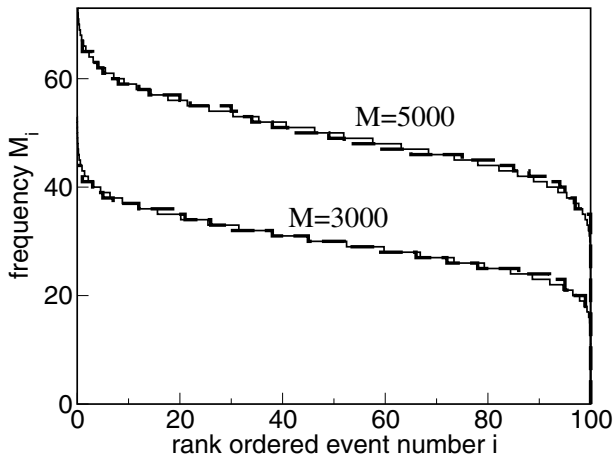


Figure 4
Same as Fig. 3 but for larger sample size *M*.

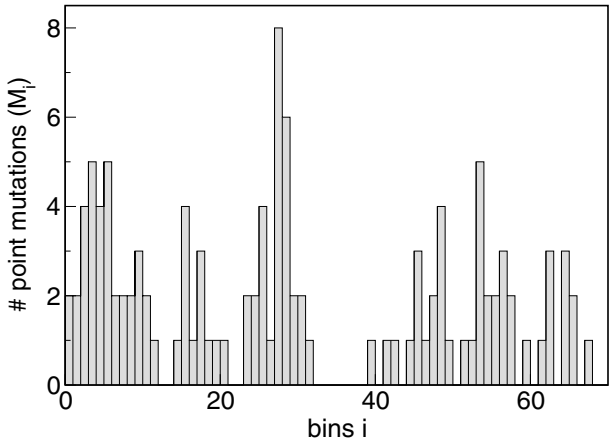


Figure 5
Observed numbers of point mutations. The sequence has been subdivided into 74 parts of length 20.

(recall that the theoretical curve was generated according to the *averaged* occupation numbers, Eq. (3)). In such cases there is no probability distribution $\{p_i\}$ which reproduces the experiment *on average*. This case can be artificially evoked when the species in the input file occur with almost identical frequencies.

The difference between the experimental rank ordered frequency distribution and the corresponding theoretical distribution (Eq. (5)) evaluates the degree of coincidence of the input data with an equidistribution (case (1)) or with a specified distribution (case (3)). We define the score by

$$S = \sum_{i=0}^{M} \left| M_i - M_i^{\text{theo}} \right|. \qquad (6)$$

The significance of a particular difference score can be assessed by relating it to the distribution of difference scores. This distribution depends on *M* and *N*.

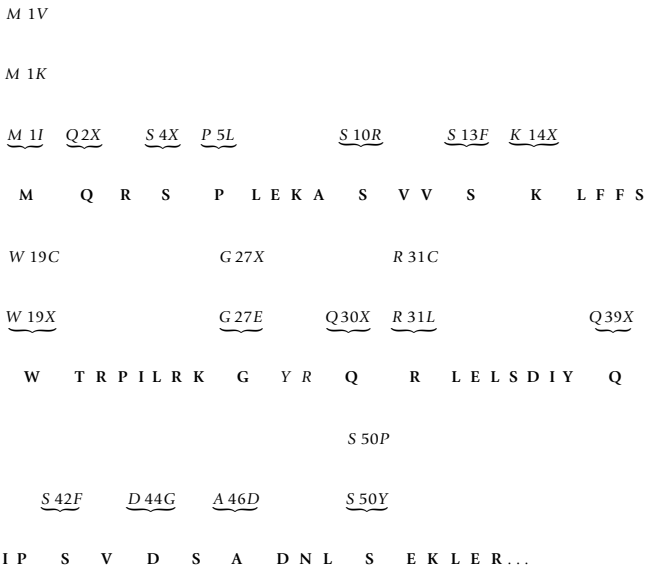
Example: Distribution of point mutations in genes

The increasing number of known point mutations and polymorphisms in many genes coding for pathogenetically important proteins offers the opportunity to apply statistical tests to correlate their type and location to evolutionary, biological and clinical features.

In each replication generation there occur mutations of the genome but frequently they remain unnoticed since they do not cause diseases. These so-called polymorphisms or variants may occur either in regions of the genome which are coding for amino acid sequences or in non-coding segments. Those changes of the DNA sequence that alter the amino acid sequence are frequently associated with diseases because the respective proteins cannot operate properly. Screenings for mutations using DNA of patients have been performed for many human diseases and the identified mutations are accessible in mutation databases [6].

The detection of so-called mutation hot spots, i.e. sequence regions with many mutation positions, is important for the identification of the functional and genetical properties of the genetic code [7]. These hot spots must be distinguished from statistical fluctuations that occur even when the probabilities for mutations are identical for each residue position. Moreover, the spatial distribution of point mutations in genes is of importance for the localization of coding and non-coding parts in the genome.

We wish to apply the described method to the investigation of the amino acid sequence of the cystic fibrosis transmembrane conductance regulator. The unperturbed gene (wild type) is given as a sequence of 1480 letters: MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIY-QIPSVDSADNLSEKLER..., each standing for one amino acid [8]. In experiments there has been observed a large number of mutations, i.e., deviations from this sequence. Such mutations are available from data bases, e.g. [6].



The codes on top of the underbraces stand for the found mutations, e.g., *P5L* means that at position 5 it has been found that the amino acid proline (P) was replaced by leucine (L).

We subdivided the sequence into 74 parts of equal length 20 and counted the number of point mutations in each part. This way we obtain the *measured frequencies* M_i = {2,2,4,5,4, 5, 2, 2, 2, 3, 2,1,0,0,1,4,... } which serve as input data. (The subdivision into parts may be repeated with a different starting point which yields similar results.) Certainly, measured frequencies as small as given above do not allow for the application of the χ^2-test. The measured frequencies are shown in Fig. 5. Obviously, based on this data it is not possible to decide a priori whether the frequencies are equidistributed.

After processing the data as described above we obtain the rank ordered measured distribution (bars in Fig. 6). The full line shows the expected (theoretical) frequency distribution due to Eq. (5) which has been generated with the hypothesis that the positions of the point mutations are equidistributed. Both curves deviate significantly from each other, therefore, we conclude that the mutations are not equidistributed. This conclusion agrees with the hypothesis in ref. [9].

Since the investigation of point mutation is an interesting field of application of the program we developed a separate input mask for this purpose (case (4) of the list in the previous section). The input syntax for this mode is described in detail in the online help file of the program.

Recently, it has been shown for point mutations in the human androgen receptor (AR) that the severity of the disease correlates with the local sequence conservation [11]. Germline mutations in the gene of the androgen receptor lead to the androgen insensitivity syndrome (AIS). In addition it was found that somatic point mutations associated with prostate cancer are more frequently found at locations with higher sequence variation compared to germline mutations leading to complete AIS. The related prediction method SIFT [10] has been proposed recently. Both methods, SIFT and the method used in [11] are based on the alignment of a large number of related proteins. Inspired by their observation we asked the question whether mutations in the androgen receptor are distributed randomly over the sequence depending on the association with AIS or prostate cancer. The disease-associated mutations in the AR were obtained from the AR gene mutation database [12]. Multiple mutations at identical positions were counted only once. Those mutations resulting in single amino acid substitutions were included in the analysis. The test was performed for 61 mutations associated with prostate cancer and 86 mutations found in patients with complete AIS. To perform the analysis we divided the sequence of 919 amino acids into 46 intervals of length 20 and counted the number of mutations in each interval. As expected, the results for the two datasets were different: Cancer associated mutations are more disseminated than congenital mutations found in patients with AIS. For mutations associated with prostate cancer the bar chart of the rank ordered frequencies nearly follows the theoretical curve for equal probabilities (Figs. 7,

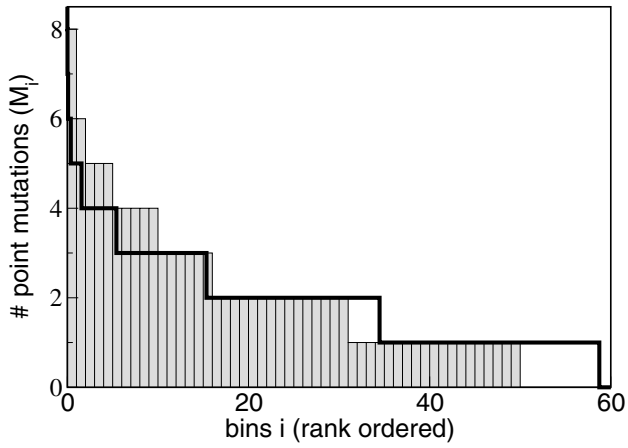


Figure 6
Results of the computation. The bars show the rank ordered frequencies, the line displays the expected frequency distribution which would be obtained if the point mutations were (equally) randomly distributed. The curves differ significantly, therefore, we conclude that the point mutations are not random.

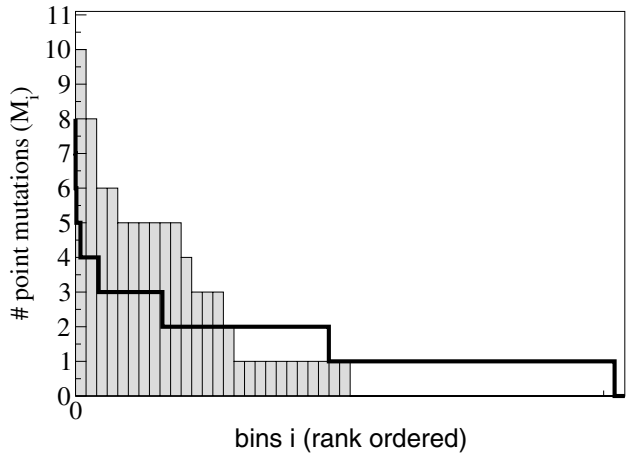


Figure 7
Distribution of missense mutations in the androgen receptor for germline mutations leading to AIS. The heights of the bars reflect the rank ordered frequencies of mutations in sequence intervals of length 20. The thick line displays the expected frequencies which would be obtained if point mutations were randomly distributed.

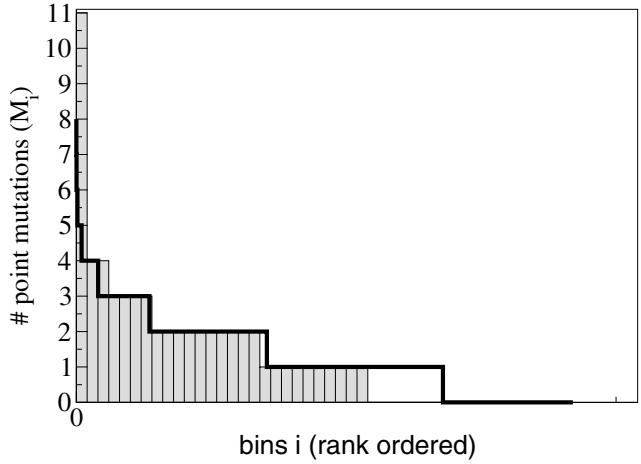


Figure 8
Distribution of missense mutations in the androgen receptor for somatic mutations associated with prostate cancer. Explanation see caption of Fig. 7.

8) whereas for AIS associated mutations the bar chart deviates markedly from the theoretical curve. Based on this finding we hypothesize that mutagenesis in the germline is followed by a selection process so that only a portion of the mutations are found in patients while others lead to early embryonal or fetal death. Conversely, mutations associated with prostate cancer may persist and are recorded.

Conclusions

For small sample sizes the relative frequencies M_i/M of occurrence of individuals of a certain species i deviate significantly from the probabilities of occurrence p_i. With the assumption that the N species occur with equal probability $p_i = 1/N$ the expectation values $\langle K_j \rangle$ of the numbers of events which are contained j times ($j = 0,..., M$) in a sample of M individuals can be determined based on combinatorial algebra. These expectation values allow for a prediction of the rank ordered frequency distribution.

For many practical problems the amount of available data is insufficient to employ standard tests, such as χ^2, to discriminate whether or not a certain set of events complies with an equiprobability distribution. For such situations which occur frequently in the biological sciences we have developed an online tool which is available at http://bioinf.charite.de/equifreq/. As demonstrated for the case of point mutations in the sequence of amino acids of the cystic fibrosis transmembrane conductance regulator and the androgen receptor, even for sample set sizes which are certainly not sufficient to decide this question directly from the observed frequencies (see Figs. 3, 4) this tool helps to make a reliable statement.

The proposed method may be generalized to arbitrary probability distributions provided there exists a hypothesis on the functional form of the distribution [13]. For mathematical reasons, however, (see [5]) it is more difficult to derive an equivalent to Eq. (5) formula for non-equiprobability distributions, which is subject of current research.

Avalability and requirements

- **Project name:** equifreq
- **Project home page:** http://bioinf.charite.de/equifreq/
- **Operating systems:** platform independent
- **Programming language:** C++
- **Other requirements:** none
- **License:** GNU GPL
- **Any restrictions to use by non-academics:** none

Authors' contributions

TP worked out the statistical and combinatorial background, wrote the kernel C++-program and drafted the

manuscript. CF and CG provided the biological expertise, collected relevant biological data and organized the biological relevant applications. CG wrote the PHP user interface. All authors contributed in writing the manuscript.

Acknowledgments

The authors are grateful to W. Ebeling, J. Freund and R. Mrowka for helpful discussion. We thank the reviewers for their helpful remarks and recommendations. Particularly fruitful was the analysis of mutations in the androgen receptor.

References

1. Pöschel T, Freund JA: **Finite-sample frequency distributions originating from an equiprobability distribution.** *Physical Review E* 2002, **66:**026103.
2. von Mises R: **Über Aufteilungs- und Besetzungswahrscheinlichkeiten.** *Revue de la Faculté de Sciences de l'Université d'Istanbul* 1939, **4:**145-163.
3. Johnson JN, Kotz S: *Urn Models and Their Application New York: Wiley*; 1977.
4. Freund JA, Pöschel T: **A statistical approach to vehicular traffic.** *Physica A* 1995, **219:**95-114.
5. Pöschel T, Ebeling W, Frömmel C, Ramírez R: **Correction algorithm for finite sample statistics.** *European Physical Journal E* in press.
6. Cotton RG, Horaitis O: **The HUGO mutation database initiative. Human genome organization.** *Pharmacogenomics* 2002, **2:**16-19.
7. Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS, Greenblatt MS: **Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features.** *Oncogene* 1999, **7:**211-218.
8. Zielenski J, Rozmahel R, Bozon D, Kerem B, Grzelczak Z, Riordan JR, Rommens J, Tsui LC: **Genomic DNA sequence of the cystic fibrosis transmembrane conductance regulator (CFTR) gene.** *Genomics* 1991, **10:**214-228. (see entry CFTR_HUMAN in the SWISSPROT database).
9. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan JR, Tsui LC, Collins FS: **Identification of the cystic fibrosis gene: chromosome walking and jumping.** *Science* 1989, **245:**1059-1065.
10. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Research* 2003, **13:**3812-3814.
11. Mooney SD, Klein TE, Altman RB, Trifiro MA, Gottlieb B: **A functional analysis of disease-associated mutations in the androgen receptor gene.** *Nucleic Acids Research* 2003, **31:**e42.
12. Gottlieb B, Lehvaslaiho H, Beitel LK, Lumbroso R, Pinsky L, Trifiro M: **The Androgen Receptor Gene Mutations Database.** *Nucleic Acids Research* 1998, **26:**234-238.
13. Pöschel T, Ebeling W, Rosé H: **Guessing probability distributions from small samples.** *Journal of Statistical Physics* 1995, **80:**1443-1452.