Distribution of Human Genes Observes Zipf's Law.

Jean-Christophe Nebel and Sergio Pezzulli
Faculty of Science, Engineering and Computing
Kingston University, London
KT1 2EE

Keywords
Human genome, gene distributions, chromosomes, mathematical models, Benford's law, Zipf's law, gene detection, gene annotation, bioinformatics, data mining.

Abstract
Recent research suggests that gene distribution on chromosomes can be informative about their nature. Consequently, gene distribution analysis may contribute not only to better gene detection, but also to better gene annotation, which is particularly important to high-throughput genome projects. This paper investigates possible mathematical models, namely Benford's and Zipf's law, to describe gene's position distributions on human chromosomes. After a review of phenomena following either of these laws, it is shown that observance of Benford's law has to be rejected. However, most human chromosomes display gene distributions which can be accurately modelled by Zipf's law. This discovery may impact the analysis of genome sequence data since the proposed gene distribution model could be integrated in software involved in gene detection.

Introduction

Recent research suggests that not only gene distribution on chromosomes is not random (Rafiee *et al.*, 2008), but their location can be informative about their nature. A study of lineage-specific genes in *Plasmodium* revealed that species-specific genes are located near chromosome ends (Kuo & Kissinger, 2008). Moreover, experiment conducted on *C elegans* indicates that gene positions on chromosomes impact on physical trait variability (Rockman *et al.*, 2010). These findings suggest the analysis of gene distribution on chromosomes may contribute not only to better gene detection, but also to better gene annotation. This is particularly relevant to high-throughput genome projects where better automatic annotation methods are required (Yang *et al.*, 2010). This paper intends to contribute to this field by providing a mathematical model of gene's position distributions on human chromosomes.

Independently, Newcomb (1881) and Benford (1938) observed that the usage of logarithm books followed a very specific distribution, now called Benford's law, where numbers starting with a digit *d* are more frequent than those starting with the digit *d+1*. More specifically, this is expressed by the following equation where *P(d)* is the probability of observing a number starting with the digit *d*:

$$P(d) = log_{10}\left(1 + \frac{1}{d}\right), \text{where } d \in \{1,2,\ldots,9\}$$

The distribution of many phenomena has been shown to follow this law. It can be found not only in data derived from human activity (baseball statistics, numbers found in a newspaper, street addresses (Benford, 1938), computer file sizes (Torres *et al.*, 2007)), but also in various academic fields including physics (physical constants (Benford, 1938; Burke & Kincanon, 1991), molecular weights (Benford, 1938), nuclear physics (Dong-Dong *et al.*, 2009; Shao & Ma, 2009), pulsar quantities (Shao & Ma, 2010)), biology (microarray data (Hoyle *et al.*, 2002), biological pathway kinetic rates (Grandison & Morris, 2008), number of cells in colonies (Costas *et al.*, 2008), genome sizes (Friar *et al.*, 2012)), social sciences (hydrology data (Benford, 1938; Nigrini & Miller, 2007), populations and death rates (Benford, 1938), stock market indices (Ley, 1996)), and mathematics (survival distributions (Leemis *et al.*, 2000), differential equations (Berger *et al.*, 2005; Berger, 2005), prime numbers (Caldwell, 2009), Fibonacci sequence (Trono, 2009)).

Initially treated as a mathematical curiosity, Benford's law has now been rigorously explained and analysed through theoretical studies. It is not only scale-invariant (Berger *et al.*, 2008), but base-invariant (Hill, 1995a). Moreover, data distributions showing such invariance must follow Benford's law (Hill, 1995b).

In addition to this theoretical work, the practical usage of this law has also been investigated. Applications can be classified in two categories: data quality control and novel data processing techniques. In the financial sector, deviation from Benford's law is exploited to detect potential cases of either irregularities or fraud (Nigrini, 1996; Nigrini & Mittermaier, 1997; Busta & Weinberg, 1998; Rose & Rose, 2003; Geyer & Williamson, 2004; Hales *et al.*, 2009; Bhattacharya *et al.*, 2011). Similarly, Benford's law is used in other fields, such as drug discovery (Orita *et al.*, 2010), national elections (Taylor, 2005; Mebane, 2008; Roukema, 2009), marketing surveys (Judge & Schechter, 2009) and pollution self-reporting (De Marchi & Hamilton, 2006; Auffhammer & Carson, 2008), to highlight suspicious data in terms of either source or quality.

The development of new data processing approaches has also contributed to very different disciplines. In computer science, Benford's law allowed the conception of novel algorithms to optimise processing time and storage space (Barlow & Bareiss, 1985; Schatte, 1988; Berger & Hill, 2007; Osmond, 2009). In bioinformatics, Benford's law led to the design of a new normalisation technique for microarray data which is particularly suitable for between-array gene intensity comparisons (Lu *et al.*, 2005). Finally, in medicine, Benford's law can separate states of consciousness and unconsciousness through digit distribution analysis of electroencephalographic signals (EEG) (Horn *et al.*, 2006).

As a whole, the discovery of phenomena following Benford's law, their potential application and theoretical analysis have been growing fields of interest which have generated more than 600 entries in the 'Benford Online Bibliography' (Berger & Hill, 2011).

Related to Benford's law (Pietronero *et al.*, 2004), Zipf's law is another statistical law (Zipf, 1949) which expresses power law relationships between the frequency of an event, *P*, and its rank, *i*:

$$P_i \sim i^{\alpha}$$, where α is close to *1*

Zipf's law was first observed in linguistics between the number of times an English word occurs in a text and its ranking in the list of the most common words (Zipf, 1949). Due to its origin, Zipf's law has been mainly used in bibliometrics showing that the law holds even in non-European languages (Rousseau & Zhang, 1992). However, Zipf's law has also been detected in topics as varied as city populations (Zipf, 1949) and protein families, folds and functions (Luscombe *et al.*, 2002). Although there is no definite criterion which allows predicting that a dataset observes either Benford's or Zipf's law, some characteristics are shared by those which do: data must spread across several orders of magnitude and dataset size above 1000 has the best chance to produce good results (Hales *et al.*, 2008). Since positions of human genes usually display these features and their distribution is known to be non-random (Rafiee *et al.*, 2008), this paper investigates the hypotheses that positions of human genes may observe Benford's and Zipf's laws.

Methods

Positions of genes on human chromosomes were studied to evaluate if their distribution fits Benford's and Zipf's laws. The transcription start positions of all human protein coding genes measured from the start of their associated chromosome were collected from the Homo Sapiens Ensembl database release 60, 8 November 2010 (Hubbard *et al.*, 2009). Using the Ensembl Perl API, 20, 593 known and novel protein-coding genes were retrieved from the twenty two autosomes, the X and Y sex chromosomes and the mitochondrial genome (MT).

The chi-square ($\chi 2$) test is the most popular goodness of fit test because it is a nonparametric asymptotic test. In other words, there are no distributional assumptions and the only requirement regards the number of observations. The standard rule of thumb for results' accuracy is that the expected counts in each cell should be at least 5. This is fulfilled with large margin on all those data, except when dealing with MT and Y chromosomes. Whereas MT could not be processed (it contains only 13 genes), Y required merging the 3 last classes, i.e. distributions of digits 7, 8 and 9.

Since the chi-square statistic is suitable for gene data and has already been applied in similar studies (Hoyle *et al.*, 2002; Dong-Dong *et al.*, 2009; Hales *et al.*, 2009; Orita *et al.*, 2010; Shao & Ma, 2010), it was used to calculate and compare Benford's distribution against the observed frequencies of the first digit of gene positions for each chromosome individually.

Finally, using the collected positions, genes were ranked on each chromosome. The relationship between rank, i, and position was studied by least-square fitting a power law for each chromosome. In addition to estimating power law exponent terms, α, the coefficients of determination (R^2) were calculated as a measure of fit between data and power law (the closer the value is to *1* the better the fit is).

Results

Observance to Benford's law

Using the $\chi 2$ test, *p*-values were calculated for each human chromosome (except MT). They are shown on Table 1. As it is generally accepted, *p*-values below *0.05* are used to reject the "null" hypothesis, *i.e.* to reject Benford's law. There is no ambiguity in these results: Benford's law cannot be accepted for any human chromosome.

Chromosome	1	2	3	4	5	6	7	8
Number of genes	2034	1275	1075	779	897	1050	944	782
p-value	4E-108	6E-78	3E-116	6E-52	8E-115	8E-37	6E-62	1E-23

Chromosome	9	10	11	12	13	14	15	16
Number of genes	811	786	1355	1056	335	633	685	913
p-value	4E-77	8E-61	2E-184	3E-96	2E-11	1E-69	1E-105	5E-63

Chromosome	17	18	19	20	21	22	X	Y
Number of genes	1218	292	1462	555	239	464	857	83
p-value	4E-194	2E-17	4E-219	4E-77	1E-118	4E-82	4E-66	2E-11

Table 1: *P-value for each human chromosome*

Observance to Zipf's law

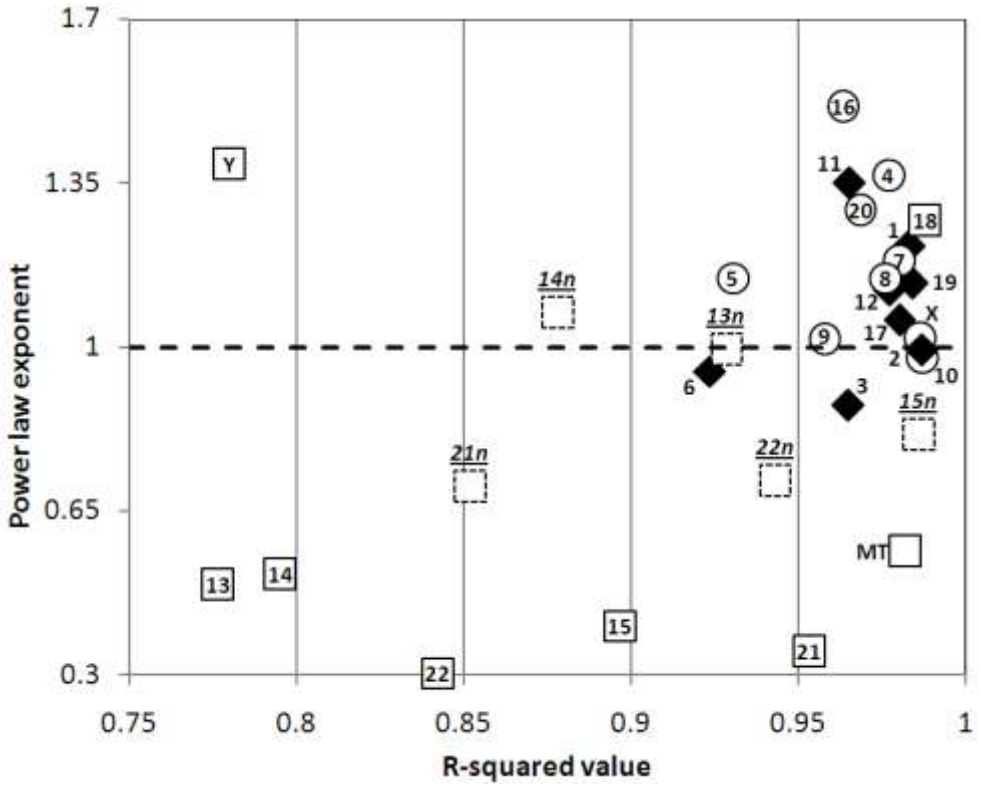
As measured by *R*-squared values, Fig. 1 reveals that the distributions of all chromosomes are well represented by power laws ($0.77<R^2<0.99$) and their power law exponents tend to cluster around a value of *1*. Consequently, since Zipf's law requires good fit of a power law with an exponent close to *1*, these results advocate that the formulated hypothesis regarding human genes' positions observing Zipf's law is likely to be correct.

More specifically, Fig. 1 shows that chromosomes with the largest number of genes, here more than 1000, have distributions which fit more closely Zipf's law: $0.92<R^2<0.99$ and $0.87<\alpha<1.36$. Conversely, chromosomes with fewer than 700 genes tend to display the lowest R^2 values and the power law exponents which are the furthest away from a value of *1*. Analysis of this set of chromosomes reveals that, in addition to include MT and Y, which contain very few genes, it comprises all acrocentric chromosomes. Among them, chromosomes 13, 14, 15 and 22 do not have a single gene on their 'p' arm. In order to address this specificity, further experiments were conducted by only considering the 'n' arm of acrocentric chromosomes, *i.e.* power laws were fitted on gene positions indexed from their centromere. Results, also shown in Fig. 1, uncover that gene distributions on the 'n' arm of those chromosomes display a close fit to Zipf's law: $0.85<R^2<0.99$ and $0.70<\alpha<1.08$.

Experiments confirms that the distribution of genes on all human chromosomes, possibly with the exceptions of MT and Y, can be satisfactorily, *i.e.* $0.85<R^2$, modelled by Zipf's law. Moreover, the need of modelling acrocentric chromosomes

according to the position of their centromere could inform current theories regarding chromosome evolution (Schubert, 2007).

Figure 1: *Fit of Zipf's law versus power law exponent for all human chromosomes. Square, circle and filled-diamond markers represent chromosomes with, respectively, fewer than 700, between 700 and 1000, and more than 1000 genes. Dashed square markers (with italic and underlined labels) correspond to acrocentric chromosomes where only the 'n' arm is considered.*



Conclusions

This paper investigates possible mathematical models, namely Benford's and Zipf's law, to describe the distribution of gene positions on human chromosomes. None of them follows the Benford law, as the observed departures from theoretical frequencies are highly significant. On the other hand, most human chromosomes display gene distributions which can be accurately modelled by Zipf's law. Preliminary results (not shown) obtained on other genomes, *i.e.* mouse, chicken and yeast, suggest gene distribution properties discovered in the human genome are also valid for other eukaryotes.

This discovery may impact the analysis of genome sequence data since the proposed gene distribution model could be integrated in software involved in gene detection. This work also suggests that, although the production of genome assemblies aligned to a genome of reference is essential for inter species comparisons, *e.g.* human and chimpanzee, availability of the actual positions of genes on chromosomes is also indispensable to allow complete analysis of a specific genome.

References

Auffhammer, M. & Carson, R.T. (2008) Forecasting the path of China's CO2 emissions using province-level information. Journal of Environmental Economics and Management, 55(3), 229-247.

Barlow, J.L. & Bareiss, E.H. (1985) On round off error distributions in floating point and logarithmic arithmetic. Computing, 34, 325–347.

Bhattacharya, S.D., Xu, D. & Kumar, K. (2011) ANN-based auditor decision support system using Benford's law. Decision Support Systems, 50(3), 576-584.

Benford, F. (1938) The law of anomalous numbers. Proceedings of the American Philosophical Society, 78, 551–572.

Berger, A., Bunimovich, L.A. & Hill, T.P. (2005) One-dimensional dynamical systems and Benford's law. Transactions of the American Mathematical Society, 357, 197–219.

Berger, A. (2005) Multi-dimensional dynamical systems and Benford's law. Discrete and Continuous Dynamical Systems, 13, 219–237.

Berger, A. & Hill, T.P. (2007) Newton's method obeys Benford's law. American Mathematical Monthly, 114, 588–601.

Berger, A., Hill, T.P. & Morrison, K.E. (2008) Scale-distortion inequalities for mantissas of finite data sets. Journal of Theoretical Probability. 21, 97–117.

Berger, A. & Hill, T.P. (2011) Benford online bibliography. www.benfordonline.net [accessed 12/7/2011]

Burke, J. & Kincanon, E. (1991) Benford law and physical constants–the distribution of initial digits. American Journal of Physics, 59, 952.

Busta, B. & Weinberg, R. (1998) Using Benford's law and neural networks as a review procedure. Managerial Auditing Journal. 13(6), 356-366.

Caldwell, C.C. (2009) Does Benford's law apply to prime numbers? The Prime Pages. http://primes.utm.edu/notes/faq/BenfordsLaw.html [accessed 12/7/2011]

Costas, E., López-Rodas, V., Toro, F.J. & Flores-Moya, A. (2008) The number of cells in colonies of the cyanobacterium Microcystis aeruginosa satisfies Benford's law. Aquatic Botany, 89(3), 341-343.

Dong-Dong, N., Lai, W. & Zhong-Zhou, R. (2009) Benford's Law and Beta-Decay Half-Lives. Communications in Theoretical Physics, 51, 713.

Friar, J.L., Goldman, T. & Perez–Mercader, J. (2012) Genome sizes and the Benford distribution. PLoS ONE 7(5).

Geyer, C.L. & Williamson, P.P. (2004) Detecting fraud in data sets using Benford's law. Communications in Statistics: Simulation and Computation, 33, 229–246.

Grandison, S. & Morris, R.J. (2008) Biological pathway kinetic rate constants are scale-invariant. Bioinformatics, 24(6), 741-743.

Hales, D.N., Sridharan, V., Radhakrishnan, A., Chakravorty, S. & Siha, S. (2008) Testing the accuracy of employee-reported data: an inexpensive alternative approach to traditional methods. European Journal of Operational Research, 189 (3), 583–593.

Hales, D.N., Chakravorty, S.S. & Sridharan, V. (2009) Testing Benford's Law for improving supply chain decision-making: A field experiment. International Journal of Production Economics, 122(2), 606-618.

Hill, T.P. (1995a) The significant-digit phenomenon. American Mathematical Monthly, 102, 322–327.

Hill, T.P. (1995b) Base-invariance implies Benford's law, Proceedings of the American Mathematical Society, 123, 887–895.

Hoyle, D.C., Rattray, M. & Jupp, R. (2002) Making sense of microarray data distributions. Bioinformatics, 18(4), 576-584.

Horn, B., Kreuzer, M., Kochs, E.F. & Schneider, G. (2006) Different states of anesthesia can be detected by Benford's Law. Journal of Neurosurgical Anesthesiology, 18(4), 328-329.

Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. & Flicek, P. (2009) Ensembl 2009. Nucleic Acids Research, 37, D690-D697.

Judge, G. & Schechter, L. (2009) Detecting problems in survey data using Benford's law. Journal of Human Resources, 44, 1-24.

Kuo, C.-H. & Kissinger, J.C. (2008) Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria, BMC Evolutionary Biology, 8, 108.

Ley, E. (1996) On the peculiar distribution of the US stock indexes' digits. The American Statistician, 50, 311–313.

Leemis, L.M., Schmeiser, B.W. & Evans, D.L. (2000) Survival distributions satisfying Benford's law. The American Statistician, 54, 236–241.

Luscombe, N.M., Qian, J., Zhang, Z., Johnson, T. & Gerstein, M. (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biology, 3(8).

Lu, T., Costello, C.M., Croucher, P.J.P., Häsler, R., Deuschl, G. & Schreiber, S. (2005) Can Zipf's law be adapted to normalize microarrays? BMC Bioinformatics, 6:37.

De Marchi, S. & Hamilton, J.T. (2006) Assessing the accuracy of self-reported data: An evaluation of the toxics release inventory. Journal of Risk and Uncertainty, 32(1), 57-76.

Mebane, W.R. Jr (2008) Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections. In Alvarez, R.M. et al. (eds.), Election Fraud: Detecting and Deterring Electoral Manipulation. Brookings Press, Washington DC, pp. 161-181.

Newcomb, S. (1881) Note on the frequency of use of the different digits in natural numbers, American Journal of Mathematics, 4, 39–40.

Nigrini, M.J. (1996) A taxpayer compliance application of Benford's law. Journal of the American Taxation Association, 18 (1996), pp. 72–91.

Nigrini, M.J. & Mittermaier, L.J. (1997) The use of Benford's Law as an aid in analytical procedures. Auditing, 16, 52–67.

Nigrini, M.J. & Miller, S.J. (2007) Benford's Law Applied to Hydrology Data—Results and Relevance to Other Geophysical Data. Mathematical Geology, 39(5), 469-490.

Orita, M., Moritomoa, A., Niimia, T. & Ohno, K. (2010) Use of Benford's law in drug discovery data. Drug Discovery Today, 15(9-10), 328-331.

Osmond, R.F. (2009) Method for improving the effectiveness of hash-based data structures. United States Patent and Trademark Office Application, #20100299333 - Class: 707747.

Pietronero, L., Tosatti, E., Tosatti, V. & Vespignani, A. (2004) Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. Journal of Physica A, 293, 297-304.

Rafiee, L., Mohsenzadeh, S. & Saadat, M. (2008) Nonrandom gene distribution on human chromosomes. EXCLI Journal, 7, 151-153.

Rousseau, R. & Zhang, Q. (1992) Zipf's data on the frequency of Chinese words revisited. Scientometrics, 24(2), 201-220.

Rose, A.M. & Rose, J.M. (2003) Turn Excel into a financial sleuth: an easy-to-use digital analysis tool can red-flag irregularities. Journal of Accountancy, 196, 58–60.

Roukema, B.F. (2009) Benford's Law anomalies in the 2009 Iranian presidential election. eprint arXiv, 0906.2789.

Rockman, M.V., Skrovanek, S.S. & Kruglyak, L. (2010) Selection at Linked Sites Shapes Heritable Phenotypic Variation in C. elegans. Science, 330, 372-376.

Schatte, P. (1988) On mantissa distributions in computing and Benford's law. Journal of Information Processing and Cybernetics, 24, 443–455.

Shao, L. & Ma, B.-Q. (2009) First digit distribution of hadron full width. Modern Physics Letters A, 24, 3275–3282.

Shao, L. & Ma, B.-Q. (2010) Empirical mantissa distributions of pulsars. Astroparticle Physics, 33(4), 255-262.

Schubert, I. (2007) Chromosome evolution, Current Opinion in Plant Biology, 10(2), 109-115.

Taylor, J. (2005) Too many ties? An empirical analysis of the Venezuelan recall referendum counts. unpublished manuscript, Stanford University, USA.

Torres, J., Fernández, S. & Gamero, A. (2007) How do numbers begin? (The first digit law). European Journal of Physics, 28, L17–L25.

Trono, J A. (2009) Discovering more properties of the Fibonacci sequence. Journal of Computing Sciences in Colleges, 24(5), 130-135.

Yang, Y., Gilbert, D. & Kim, S. (2010) Annotation confidence score for genome annotation: a genome comparison approach. Bioinformatics, 26(1): 22-29.

Zipf, G.K. (1949) Human Behavior and the Principle of Least Effort. Addison-Wesley Press Inc., Cambridge.