SOME WAYS TO INFER A DNA FUNCTION FROM THE SEQUENCE INFORMATION

Abnizova I.*, te Boekhorts R., Gilks W.

MRC-BSU Cambridge, University of Hertfordshire Hatfield, UK
* Corresponding author: irina.abnizova@mrc-bsu.cam.ac.uk

Keywords: *regulatory regions, coding DNA, heterogeneity, computational methods, information entropy, long range correlations*

Summary

We present a computational approach to infer DNA function based on eukaryotic DNA sequence information. Namely, we utilise an observation that exons, regulatory regions and non coding non regulatory DNA exhibit different statistical information patterns. We suggest to capture and measure these patterns with several independent mathematical tools such as rescaled analysis, information entropy, density of low-entropy patches and similarity test.

Results: We introduce here a new optimization technique of which the outcome is independent of the size of the sliding window and hence avoids averaging. This technique, which takes account of the heterogeneity in the DNA sequence, performs an unsupervised search, without using reference sets or cross genome comparison. We also introduce new way of measuring DNA local heterogeneity, and a statistical test for abundance of similar words.

A preliminary application of our methods to the set of genes from six different species, namely human, mouse, fugu, sea urchin, drosophila and yeast, reliably identifies the borders of the regions of interest and thus reveals the potential of our approach.

We propose that this combination of established computational statistical methods, augmented with our sliding window optimization technique, and our new statistical tests, might create a powerful DNA sequence characterization and annotation tool that optimises the search for differences in statistical properties between coding, non-coding and regulatory DNA.

Availability: The software is available from the authors on request.

Introduction

The unsupervised search methods for analysing the structure of the genome fall, broadly speaking, in one of two categories: (i) methods for characterizing the composition of the genome and (ii) those used for detecting sequential or serial dependency (i.e. that focus on the actual ordering of nucleotides).

Nucleotide composition is commonly investigated with tools from information theory (i.e. various ways to estimate the entropy of parts of the genome), self-organising maps (Abe *et al*., 2003), complexity analysis (Wan *et al*., 2003; Li , 1997) and statistical linguistics (Mantegna *et al*., 1994).

Statistical dependencies between nucleotides/amino acids have been analysed using mutual information functions (Azbel, 1995), Markov models (Krogh *et al*., 1994), spectra (Voss, 1992) and methods derived from random walk dynamics such as detrended fluctuation- and rescaled range analysis (Peng *et al*., 1994). In particular the detection of long range correlations (LRCs) has attracted much attention (Li, 1997; Mantegna *et al*., 1994; Azbel, 1995; Peng *et al*., 1994; Herzel *et al*., 1997) and correlations ranging from a few base pairs up to 1000 bp have been found.

Results from the application of these methods seem to indicate that non-coding and coding DNA have distinguishable statistical properties: long range correlations have been reported for non-coding, but not for coding regions (Peng *et al*., 1994; Herzel *et al*., 1997; but see Voss, 1992).

Unfortunately, when these techniques are used in the conventional way, their results will be highly dependent on the size of the sliding window.

Methods and Result

Random walks and rescaled range analysis. One way to characterize the succession of nucleotides is to imagine a “walk” along the DNA string by moving up each time a pyrimidine (a T or C) occurs and by moving down whenever a purine (an A or G) is encountered. As a result, one may obtain a fractal-looking “landscape”, in which probably long stretches of mainly purine alternate with stretches that contain mostly pyrimidine. If the probability of the occurrence of a pyrimidine equals that of a purine and is independent of the position in the string (i.e. P (pyrimidine) = P (purine)), then such a “DNA walk” would actually be a “random walk”. The *increments* of the walk would form a series of independent and identically distributed events with constant mean and finite variance that is therefore stationary and furthermore characterised by a flat spectrum and the absence of autocorrelations (“white noise”). The random walk itself, being integrated white noise, is typically non-stationary. This is manifest in the largely non-vanishing spikes in the autocorrelation function and the predominance of low frequency components in the power spectrum (Voss, 1992).

Another way to do find long range correlations in DNA sequence is to calculate the Hurst exponent (H) by means of a “rescaled range analysis”. The Hurst exponent estimates the degree of “persistence” of the system. Rescaled range analysis can be applied to any time or space series. Setting $x_k = +1$ for k = T, C and $x_k = -1$ for k = A, G, the sequence $\{x_k\}$ can be characterised by:

$$\langle x \rangle_n = \frac{1}{n}\sum_{i=1}^{n} x_i \ , X(i,n) = \sum_{m=1}^{i}\left[x_m - \langle x \rangle_n\right]$$

$$R(n) = \max_{i \le n} X(i,n) - \min_{i \le n} X(i,n) \ , S(n) = [\frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x \rangle_n)^2]^{1/2}$$

for any $2 \le n \le N$.

For scale free data $R(n)/S(n) \sim (n/2)^H$. Hence, the Hurst exponent (H) can be computed from the least squares fit of the regression of log[R(n)/S(n)] on log[n]. In the above formulas, the test-statistic (H) should be compared with a Hurst exponent obtained under the null-hypothesis that the *cumulative* data (i.e. the time series *after* integration) are from a random walk and therefore that the *original* data are white noise. In other words, $H_{observed}$ should be contrasted with $H_{expected} = 0.5$ and *not* with $H_{expected} = -0.5$ (which would be the Hurst exponent of the original data).

Obviously, changes in persistence are not accounted for by the conventional application of rescaled range analysis, i.e. by using a too large, fixed window. This might be detected as changes in the slope of the log-log plot. We therefore suggest a window that increases until the Hurst exponent reaches a minimum ($H \le 0.5$): this window detects the most probable boards of exons.

Entropy-based DNA segmentation with optimal window method. The conventional procedure for measuring DNA entropy (Li, 1997) typically consists of calculating a frequency vector describing the area’s nucleotide composition for a sufficiently large, but subjectively defined area, and then subjecting it to the well-known Shannon function:

$$H_{seq,M} = -\sum_{i=1}^{M} P_i \, log(P_i) \ ,$$

where M is the length of the frequency vector, which in case of single nucleotides is 4, in case of di-nucleotides is 16, etc. The entropy of a frequency vector is maximal when all elements occur with equal probability, in which case $H_{seq, M} = \log(1/M)$, and hence measures the “eveness” or the "diversity" of the composition. In contrast, low entropy indicates the "dominance" of a few of the elements. Clearly, when increasingly larger fragments of the same stretch of DNA are used, the entropy of the fragments will asymptotically approach the entropy of the entire stretch. Such an overall estimate does not capture the possibly deviating entropy of small but functionally important subparts. Too large and fixed windows therefore overlook local differences in nucleotide composition.

For a more powerful method, we therefore must optimise the length of the local windows. To this aim, we move sliding window of varying length along the DNA sequence, optimising the length of the window due to the local maximum of Entropy before it reaches asymptotic value. The most high entropy regions are the most likely locations of exons.

Low entropy patches density as heterogeneity measure. There exit two visions of DNA heterogeneity: 1) Heterogeneity is caused by gradual change (bias) of nucleotide composition in different parts of the sequence; 2) Heterogeneity is caused by short runs of repetitive (low-entropy) patterns. If you remove them, DNA will be homogeneous again. It is local heterogeneity.

To calculate the density of low entropy patches, we first developed an entropy-based algorithm able to fish for these patches; this algorithm occurred to detect very weird patterns, which no repeat masker could find. Thus, we used the density of these low entropy patches as a measure of local heterogeneity of DNA.

As the result, we present an example of splitting fugu DNA into functional parts due to low-entropy patches: exons typically have minimum density, non coding non regulatory DNA has maximum density, and regulatory DNA has intermediate density, as one can see at the table below.

In the table, "**diverged"** denotes non coding non conserved DNA: they were picked up randomly through out the Mayfolds fugu whole genome shotgun assembly v.3.0 (August 26, 2002); "**cne"** (conserved non coding elements) are considered as putative regulatory regions: these elements are collected due to ClustalW multiple alignment between fugu, mouse, rat and human in the group of Greg Elgar (Woolfe *et al.*, submitted 2004), short elements were concatenated together; exons are randomly picked up in Scaffolds1 and 21 fugu whole genome shotgun assembly v.3.0 (August 26, 2002), www.ensembl.org (Table).

Measure of similar word abundance: similarity-tail test. There is well known observation that there are unusually large (unexpectedly due to random multinomial model) number of some similar words in the regulatory regions. To quantify this fact, we got the collection of known functional regulatory regions from Drosophila genome. For each region, we computed the distribution of clusters of similar words inside it, so called **similarity-tail test.** We calculated the number of similar words of length m with few mismatches exhaustively for each m-word in the given regulatory region. Due to the presence of unusually high number of over-represented words, we expect to have more "fat" (containing a lot of similar words) clusters in comparison with random chance.

To sample this random chance, we shuffled our given sequence 200 times keeping its original di-nucleotide content. For each shuffled sequence we plotted the histograms of similar words distribution. We could easily observe that the original sequence "tails" were significantly longer than all randomised ones.

Almost all regulatory regions analysed (currently, 35) exhibit "fluffy" tails, in contrast with exon regions where there is no significant tail exist. This observation gave raise to the algorithm, which distinguish between exons and regulatory regions due to the presence of statistically significant tails in the original distribution. The algorithm currently is applied for Drosophila genome.

Conclusion: Our collection of tools: **• Information entropy • Rescaled Range Analysis: Hurst exponent• Density of low entropy patterns• Tests for frequency of similar words** might help to infer the function of a given not yet annotated DNA stretch, and thus to distinguish between different DNA functional parts when all the tools are combined.

In this work we propose an adaptive optimal windowing technique applied for two popular methods which search for short and long range correlations in DNA: rescaled analysis to estimate long distance dependency by the Hurst exponent, and entropy measurement. Both improved techniques are capable to detect putative functional regions in DNA. In addition, we capture DNA homogeneity and presence of repetitive words in other two independent ways: with density of low-entropy patches, and similar words abundance in the DNA stretch. When we combine results of these tests with adaptive window methods, we therefore increase likelihood of detecting functional DNA regions.

Table.

Type of sequence	Density of low-entropy patterns	A + C + G + T	**Pcg** - probability to get **cg**	**Pcg/PcPg** **Pc** - probability to get **c**	Length, bp
Diverged					
Diverged1	0.34	0.29+0.25+0.18+0.26	0.028	0.60	5156
Diverged2	0.44	0.28+0.20+0.21+0.28	0.018	0.41	15221
Diverged3	1.0	0.22+0.36+0.18+0.21	0.03	0.44	755
Diverged4	0.66	0.26+0.18+0.17+0.37	0.02	0.65	1000
Diverged5	0.56	0.30+0.20+0.19+0.30	0.014	0.38	4819
Diverged6	0.62	0.26+0.20+0.20+0.32	0.02	0.48	4750
Diverged7	0.37	0.28+0.20+0.20+0.30	0.02	0.53	8501
Means	**0.58**	**0.28+0.22+0.20+0.29**	**0.02**	0.46	
Cnes					
Cne2	0.032	0.30+0.21+0.18+0.29	0.018	0.44	14782
Cne3	0.0	0.31+0.19+0.19+0.29	0.012	0.34	3423
Cne5	0.018	0.30+0.18+0.19+0.29	0.012	0.33	18876
Cne6	0.078	0.31+0.18+0.20+0.30	0.01	0.29	14500
Cne7	0.02	0.28+0.20+0.21+0.29	0.017	0.39	23258
Cne8	0.05	0.30+0.18+0.21+0.30	0.013	0.34	10378
Means	**0.07**	**0.30+0.19+0.20+0.30**	**0.014**	**0.37**	
Exons					
Ex1_fugu	0.0	0.24+0.27+0.27+0.20	0.048	0.64	4295
Ex2_fugu	0.0	0.27+0.25+0.26+0.22	0.036	0.54	22373
Ex3_fugu	0.03	0.27+0.26+0.26+0.19	0.04	0.57	19940
Ex4_fugu	0	0.25+0.27+0.28+0.20	0.058	0.73	1224
Ex5_fugu	0	0.27+0.26+0.26+0.19	0.045	0.63	3489
Ex6_fugu	0	0.27+0.27+0.26+0.20	0.042	0.58	2835
Ex7_fugu	0.0	0.26+0.27+0.26+0.19	0.043	0.60	3714
Ex8_fugu	0.05	0.27+0.25+0.27+0.20	0.035	0.51	5740
Means	**0.007**	**0.27+0.26+0.27+0.20**	**0.0412**	**0.58**	

References

Abe T., Kanaya S., Kinouchi M, Ichiba Y., Kozuki T., Ikemura T. Informatics for unveiling hidden genome signatures // Genome Res. 2003. V. 13(4). P. 693–702.

Abnizova I., Schilstra M., te Boekhorst R., Nehaniv C.L. A statistical approach to distinguish between different DNA functional parts // WSEAS Transactions on Computational Methods. 2003. V. 2. Issue 4. P. 1188–1196.

Azbel Y.M. Universality in a DNA statistical structure // Physical Review Letters. 1995. V. 75. P. 68–171.

Herzel H., GroЯe I. Correlations in DNA sequences: the role of protein coding segments // Physical Review E. 1997. V. 55. P. 800–810.

Krogh A., Mian S., Haussler D. A hidden markov model that finds genes in *E. coli* DNA // Nucleic Acids Res. 1994. V. 22. P. 4768–4778.

Li W. The complexity of DNA // Complexity. 1997. V. 3. P. 33–37.

Mantegna R.N., Buldyrev S.V., Goldberger A.L., Havlin S., Peng C.K., Simons M., Stanley H.E. Linguistic features of noncoding DNA sequences // Physical Review Letters. 1994. V. 73. P. 3169–3172.

Peng C.K., Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger A. Mosaic Organization of Nucleotides // Physical Rev. E. 1994. V. 1. P. 1685–1689.

Voss R. Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences // Physical Review Letters. 1992. V. 68. P. 3805–3808.

Woofle A., Goodson M., Goode D., Snell P., Smith S., Vavouri T., McEwen G., Gilks W., Walter K., Edwards Y., Elgar G. Highly conserved non coding sequences are associated with developmental control genes in vertebrates, submitted 2004.

Wan H., Li L., Federhen S., Wootton J.C. Discovering simple regions in biological sequences associated with scoring schemes // J. Comput Biol. 2003. V. 10(2). P. 171–85.