Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Mutual information content of homologous DNA sequences

Helena Cristina G. Leitão[1], Luciana S. Pessôa[1] and Jorge Stolfi[2]

[1]Instituto de Computação, Universidade Federal Fluminense (UFF), Niterói, RJ, Brasil
[2]Instituto de Computação, Universidade Estadual de Campinas (UNICAMP), Niterói, RJ, Brasil
Corresponding author: H.C.G. Leitão
E-mail: hcgl@ic.uff.br

Genet. Mol. Res. 4 (3): 553-562 (2005)
Received May 20, 2005
Accepted July 8, 2005
Published September 30, 2005

ABSTRACT. The necessary information to reproduce and keep an organism is codified in acid nucleic molecules. Deepening the knowledge about how the information is stored in these bio-sequences can lead to more efficient methods of comparing genomic sequences. In the present study, we analyzed the quantity of information contained in a DNA sequence that can be useful to identify sequences homologous to it. To reach it, we used signal processing techniques, specially spectral analysis and information theory.

Key words: DNA sequence analysis, Homologous DNA sequences, Information theory, Fourier analysis

 ©FUNPEC-RP www.funpecrp.com.br

INTRODUCTION

The search for homologous substrings in a DNA or protein database is arguably the most important problem for bio-informatics (Meidanis and Setubal, 1997). It is a necessary step in the reconstruction of large genomes from the raw data generated by DNA sequencing equipment, and it is the main tool for identifying the function and evolutionary history of genes.

Homology search can be formally modeled as follows: the databank is a set of sequences over some alphabet Σ, and the goal is to find the substrings of those sequences that match a given string, within a given similarity criterion. The expected number of substrings that match by a random query *s* in a random databank is proportional to the number, N, of databank substrings times the probability p that the query *s* matches one of those sequences. The value of p depends on the statistical distributions of the queries, of the databank strings, and of the similarity criterion.

Useful information content

The match probability p can be expressed by the Equation $I = -\log_2 p$, which may be called the useful information content of the query string; it is expressed as a number of binary digits or bits. To justify this nomenclature, consider the abstract examples below. In all cases, both the query *s* and the databank entries are random strings of n = 20 binary digits, with the same distribution.

A. The string bits are uniformly and independently drawn from {0,1}, and two strings are considered similar if and only if their last 15 bits are identical. Clearly, the match probability p is 2^{-15}, which implies I = 15; this is precisely the number of bits in the query that are useful for the search.

B. The string bits are uniformly and independently drawn, and two strings are considered similar if they differ by at most a single bit at any position. Then the match probability is $p = 2^{-n} + n\,2^{-(n-1)} = 41 \cdot 2^{-20}$, which means $I = 20 - \log_2 41 \approx 14.64$. That is, because of the tolerated errors, the 20-bit query string only determines about 14.64 bits of the matching entries.

C. The string bits are independently drawn, with probabilities $p_0 = 1/3$ for 0 and $p_1 = 2/3$ for 1, and the match tolerance is at most one bit in any position, as in the previous example. Then the probability of a match is $p = (p_0^2 + p_1^2)^n + 20\,(2\,p_0\,p_1)\,(p_0^2 + p_1^2)^{n-1} = (5/9)^{19}(5/9 + 80/9) = 17 \cdot (5/9)^{20}$, which means $I = 20\log_2(9/5) - \log_2(17) \approx 12.87$. That is, because of the biased bit probabilities, the 20 bits of the query contain less than 13 bits of useful information about the desired sequences.

D. The first bit of each string is 0 or 1 with equal probability, but each succeeding bit is equal to the previous one with probability $p_0 = 3/4$, and it is different from it with probability $p_1 = 1/4$, independently of all other bits. As above, similarity means at most one bit difference. Note that each bit, considered by itself, is 0 or 1 with equal probability; but we cannot apply the estimate of example B because the bits are not independent. To overcome this obstacle, we apply the following transformation to the strings: keep the first bit, and replace every subsequent bit by its absolute difference with the previous one. This transformation makes the bits independent of each other: the first bit is uniformly distributed, and the others are 0 or 1 with probabilities p_0 and p_1, respectively. Note that two original strings are similar if and only if the first k bits of their transforms are identical, and the remaining n-k bits are complementary, for some k

in {0,…, n}. It follows that the match probability is

$$p = (1/4)(2p_0p_1)^{n-1} + \sum_{k=1}^{n}(1/4)(p_0^2 + p_1^2)^{k-1}(2p_0p_1)^{n-k} \qquad \text{(Equation 1)}$$

$$= (1/4)(2p_0p_1)^{n-1}\left(1+\sum_{i=0}^{n-1}\left(\frac{p_0^2+p_1^2}{2p_0p_1}\right)^i\right) = (1/4)(2p_0p_1)^{n-1}\left(1+\frac{r^n-1}{r-1}\right)$$

where $r = (p_0^2 + p_1^2)/(2p_0\ p_1) = 5/3$. Therefore, $p = (1/4)(3/8)^{19}(1 + 3((5/3)^{20}-1)/2)$ and $I \approx 13.56$.

Examples A-D (above) illustrate the difficulties we must deal when computing the match probability p for a DNA similarity search. As in example C, in a "random" DNA sequence (as drawn from a gene bank or produced by a sequencing machine) the bases {A, T, C, G} are known to occur with different probabilities. Moreover, because of the way that the genetic code works, we know that there is some dependency between adjacent bases, and between bases whose positions differ by a multiple of 3, especially within active genes. Long-range dependencies have also been suspected; for example, the frequencies of the four bases seem to vary along each chromosome in a fractal-like pattern, possibly as a consequence of gene duplication and/or physico-chemical constraints in the encoded proteins. For the same reasons, short- and long-range dependencies may exist between the mutations that distinguish the response to query *s* from the matching bank entries.

All these complications prevent us from computing p (or I) directly. To overcome these difficulties, we followed the approach used in example D; namely we transformed the sequences in such a way that its components become independent, and express the similarity criterion in terms of the transformed sequences.

The transformation we use is an encoding of the bases as complex numbers, followed by a Fourier transformation. Although the Fourier coefficients cannot be proven to be independent, this hypothesis is supported by both intuition and by statistical tests. An additional advantage of this approach is that it yields the amount of information contributed by each Fourier component of the sequence. This analysis is important for predicting the performance of the fast multiscale signal matching technique, proposed by Leitão and Stolfi (2002), which may be much faster than the matching algorithms that are currently in use.

DISCRETE FOURIER TRANSFORM AND POWER SPECTRUM

Let us consider a sequence x(n) represented in a time domain, where n assumes a range of N integer values. The discrete Fourier transform applied to this signal permits us to represent it in the frequency domain, as follows:

$$X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\exp\left[-i\frac{2\pi}{N}nk\right] \qquad \text{(Equation 2)}$$

The spectral components of a power spectrum are defined as $|X(k)|^2$, where the k values, ranging between 0 and N-1, are the frequencies that represent x(n).

INTERPRETING DNA AS SIGNALS

Before we can apply the tools of information theory to this problem, we must convert the DNA strings into signals - sequences of numerical elements. So, we opted for the compromise encoding used by Cheever et al. (1989), where each base is represented by a complex number. Specifically, A, T, C, and G are mapped to +1, -1, +i, -i, where $I = \sqrt{(-1)}$ is the imaginary unit. Note that each purine base and the corresponding pyrimidine base are mapped to complementary values. This encoding has the advantage that it produces a complex signal, which is a well-studied object in signal processing theory, and it is the canonical input for the Fourier transform.

Power spectrum of DNA

Figure 1 shows the spectrum of two DNA sequences, encoded as described in the "Interpreting DNA as signals"section.

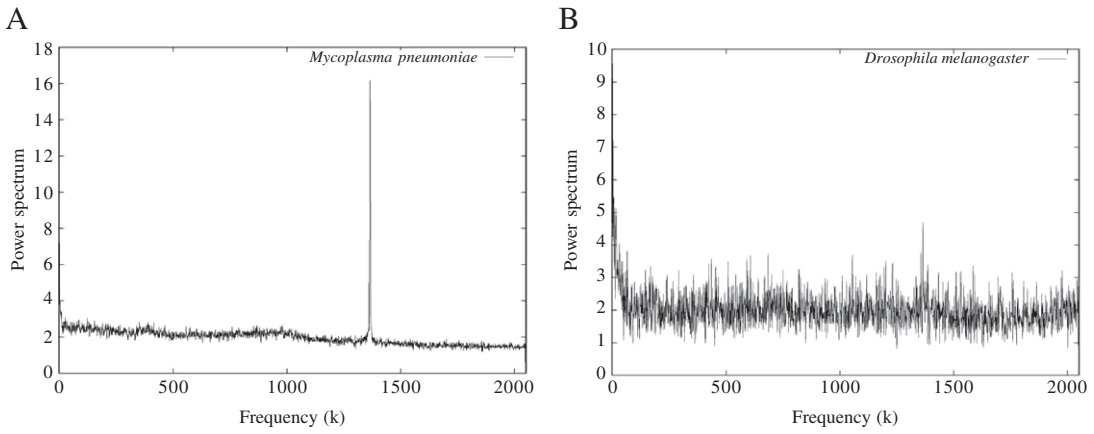

Figure 1. Power spectra of two DNA genomes taken from GenBank: **A.** the prokaryote (*Mycoplasma pneumoniae*, NC-000912) and **B.** the eukaryote (*Drosophila melanogaster*).

Note that the power spectrum of prokaryote DNA has a strong peak centered at frequency k = n/3, which corresponds to a spectral component with period 3. This peak is due to the asymmetry and non-uniform probability distribution of the codons (the three-base codes for amino acids). Anasstassiou (2002) investigated the use of this phenomenon to detect protein-coding regions in genomes. The peak spreads to the neighboring frequencies, presumably because occasional point deletions and insertions cause the period to fluctuate a little around 3.

The peak at k = n/3 is much weaker in the second spectrum, because in eukaryote DNA the coding regions are interrupted by numerous non-coding segments (introns). These are almost random sequences, which tend to have a flat power spectrum. Besides being diluted by the introns, the coding regions are also displaced by random amounts, not necessarily multiples of 3; these random shifts have the effect of broadening the peak.

Independence of the Fourier coefficients to the DNA signal

As explained in the "Useful information content" section, the main obstacle to computing the useful information contents of DNA for a similarity search is the definite but largely unknown dependence between nearby bases. We argue that the Fourier transform allows us to overcome that obstacle, because the real and imaginary parts of the Fourier coefficients are all independent random variables. The independence of the Fourier coefficients seems to be confirmed by experiments with DNA strings extracted from the GenBank database (see Figure 2).

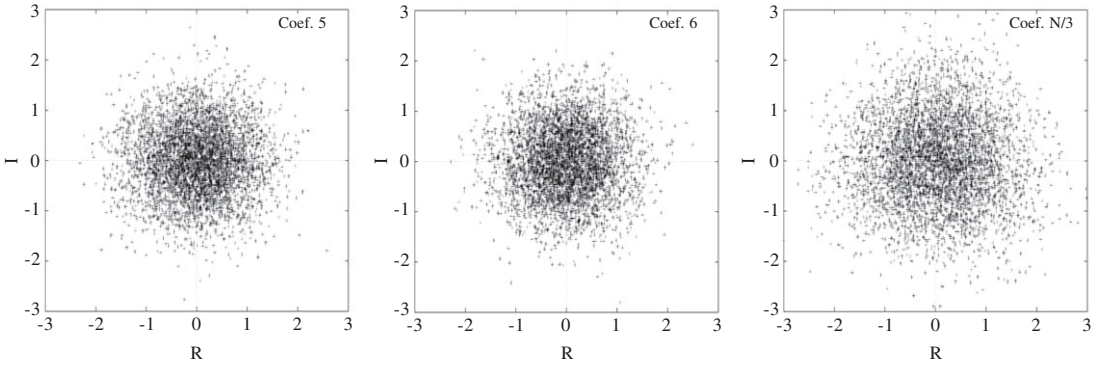

Figure 2. Each graph represents the real and imaginary values that compose the coefficients of frequencies 5, 6 and (N/3) of 5000 random DNA sequences.

This assertion cannot be proven mathematically, since one can construct distributions of DNA strings that violate it. However, it is supported by the observation that the real and imaginary parts of the Fourier basis functions are pairwise orthogonally. In fact, for every pair of distinct samples s_i and s_j that contributes to distinct coefficients S_l and S_k, there is another pair s_{i+r}, s_{j+r} that contributes to S_l with the same weights, and to S_k with opposite weights. Therefore, if the suspected correlation between s_i and s_j depends only on the distance i-j (as should be the case for randomly clipped substrings of a genome), then these correlations cancel out and do not give rise to correlations between the Fourier coefficients.

INFORMATION THEORY BASICS

Entropy

Let X be a random variable that can assume values $x_1,\ldots, x_m$ with probabilities $p_1,\ldots, p_m$. The entropy or expected information contents of X is

$$H(X) = -\sum_i p_i \log_2 p_i \qquad \text{(Equation 3)}$$

Conditional information

Suppose A and B are two random variables, with values $\{a_1,\ldots, a_p\}$ and $\{b_1,\ldots, b_q\}$, respectively. When we know that A has a particular value a_i, the information that we gain about B is:

$$I(B \mid A = a_i) = H(B) - H(B \mid A = a_i) \qquad \text{(Equation 4)}$$

where $H(B|A = a_i)$ is the result of applying Equation 3 to the conditional probability distribution $Pr(B = b_j|A = a_i)$.

The average information carried by A about B is the expected value of $I(B|A = a_i)$ averaged over all a_i, that is

$$I(B \mid A) = H(B) - H(B \mid A) \qquad \text{(Equation 5)}$$

where

$$H(B \mid A) = \sum_{i=1}^{p} H(B \mid A = a_i) \Pr(A = a_i) \qquad \text{(Equation 6)}$$

Mutual information of a Gaussian variable

An important special case that matters to us is when A = S + N and B = S + P, where S, N and P are independent complex variables with symmetric Gaussian distributions and variances $\hat{S}$, $\hat{N}$, and $\hat{P}$. We may think of S as a "message" from which one makes two independent copies, which get corrupted by "noises" N and P. Our goal is determine how much information A gives about B, on average. In this case, it turns out that A and B also have symmetric Gaussian distributions, with variances $\hat{A} = \hat{S} + \hat{N}$ and $\hat{B} = \hat{S} + \hat{P}$, respectively. So the first term of Equation 5 is simply

$$H(B) = \log_2(\pi e \hat{B}) = \log_2\left[\pi e(\hat{S} + \hat{P})\right] \qquad \text{(Equation 7)}$$

Moreover, the conditional distribution $Pr(S \approx x|A = y)$ turns out to be another Gaussian distribution, with mean $y\hat{S}/\hat{A}$ and variance $\hat{S}\hat{N}/\hat{A}$. Since P is independent of N and S, the conditional distribution of B = S + P, given A = y is also Gaussian, with the same mean $y\hat{S}/\hat{A}$ and variance $\hat{S}/\hat{N}/\hat{A} + \hat{P}$. Note that this value does not depend on y; so Equation 6 reduces to

$$H(B \mid A) = \log_2\left[\pi e\left(\frac{\hat{S}\hat{N}}{\hat{A}} + \hat{P}\right)\right] \qquad \text{(Equation 8)}$$

According to Equations 5, 7 and 8, the information given by A about B is then

$$I(B \mid A) = \log_2 \left[\frac{\hat{A}\hat{B}}{\hat{S}\hat{N} + \hat{A}\hat{P}} \right] \qquad \text{(Equation 9)}$$

The Shannon-Hartley formula

In particular, if we set N = 0 (that is, B = S) we get the Shannon-Hartley formula

$$I(S \mid A) = \log_2 \left[\frac{\hat{A}}{\hat{N}} \right] = \log_2 \left[\frac{\hat{S} + \hat{N}}{\hat{N}} \right] \qquad \text{(Equation 10)}$$

which gives the amount of information carried by the corrupted message A about the original message S (Lathi, 1968).

Gaussian signals

The formula for mutual information is greatly simplified in the case of Gaussian signals, signals with Fourier coefficients that are random independent variables with symmetric Gaussian distributions on the complex plane. Many natural signals fit this model.

Let A_k, B_k, S_k, N'_k, and N''_k be the Fourier coefficients of a, b, s, n', and n", respectively. Let us assume that the coefficients S_k, N'_k, and N''_k are independent random variables with symmetric, zero-mean Gaussian distributions over the complex plane. Let us assume also that N'_k and N''_k have the same variance $\hat{N}_k$. Then, the information given by each coefficient A_k about the corresponding coefficient B_k (Lathi, 1968) turns out to be

$$I_k = \log \left[\frac{\hat{A}_k \hat{B}_k}{\hat{S}_k \hat{N}_k + \hat{A}_k \hat{N}_k} \right] = \log \left[\frac{(\hat{S}_k + \hat{N}_k)^2}{(2\hat{S}_k + \hat{N}_k)\hat{N}_k} \right] \qquad \text{(Equation 11)}$$

The total information about b carried by a is then simply

$$I_{tot} = \sum_{-m}^{m-1} I_k$$

INFORMATION CONTENT OF DNA

We now apply this theory to estimate the useful information content of a DNA sequence for the purpose of finding homologs in a DNA bank. Note that two moderately long homologous DNA sequences are always descendants from a common ancestor.

We can view the DNA sequence abstractly as a signal corrupted by noise (differences between bases). Specifically, two homologous sequences can be written as a(t) = s(t) + n'(t) and b(t) = s(t) + n"(t), where s is the ancestral sequence, and n', n" are "noise" functions that represent mutations or lost bases.

Determining $\hat{S}_k$ and $\hat{N}_k$

Unfortunately, we have no direct information about the variance of the original signal $\hat{S}_k$ (the genomic sequence of the mutual ancestor) or of the noise $\hat{N}_k$ (the difference between the sequences caused by mutations or deletions). Let us then denote m as the average of the two signals a, b, and d, their difference. Then the Fourier coefficients M_k and D_k of signals m and d have variances

$$\hat{M}_k = \text{var}\left(\frac{S_k s + N_k^{'} + S_k + N_k^{''}}{2}\right) = \hat{S}_k + \frac{1}{2}\hat{N}_k \qquad \text{(Equation 12)}$$

$$\hat{D}_k = \text{var}((S_k + N_k^{'}) - (S_k + N_k^{''})) = 2\hat{N}_k \qquad \text{(Equation 13)}$$

Thus, given a sample of homologous DNA chains, we can compute the variances $\hat{M}_k$ and $\hat{D}_k$, and then estimate the variances $\hat{S}_k$ and $\hat{N}_k$ by the formulas

$$\hat{S}_k = \hat{M}_k - \frac{1}{4}\hat{D}_k \qquad \hat{N}_k = \frac{1}{2}\hat{D}_k \qquad \text{(Equation 14)}$$

Therefore, by Equation 11, the amount of information contained in the frequency-k component of sequence a about the same component of its partner b is

$$I_k = \log\left[\frac{(\hat{A}_k)^2}{\left(2\left(\hat{M}_k - \frac{1}{4}\hat{D}_k\right) + \frac{1}{2}\hat{D}_k\right)\left(\frac{1}{2}\hat{D}_k\right)}\right] = \log\left[\frac{(\hat{A}_k)^2}{\hat{M}_k\hat{D}_k}\right] \qquad \text{(Equation 15)}$$

Since $\hat{A}_k = \hat{S}_k + \hat{N}_k'$ (Equation 15) could be rewritten as

$$I_k = \log\left[\frac{\hat{M}_k}{\hat{D}_k} + \frac{1}{2} + \frac{1}{16}\frac{\hat{D}_k}{\hat{M}_k}\right] \qquad \text{(Equation 16)}$$

RESULTS

To test this theory, we used 48 pairs of homologous sequences of prokaryote DNA taken from GenBank, selected to have at least 90% identity within each pair. Figure 3 shows the estimated variances $\hat{M}_k$, and $\hat{D}_k$and the useful information contents I_k for each component frequency k, as computed by Equation 16. The peak obtained at frequency N/3 was expected from the corresponding peak in the mean power spectrum $\hat{M}_k$. Note that, apart from this peak, the useful information is distributed almost uniformly over all frequencies.

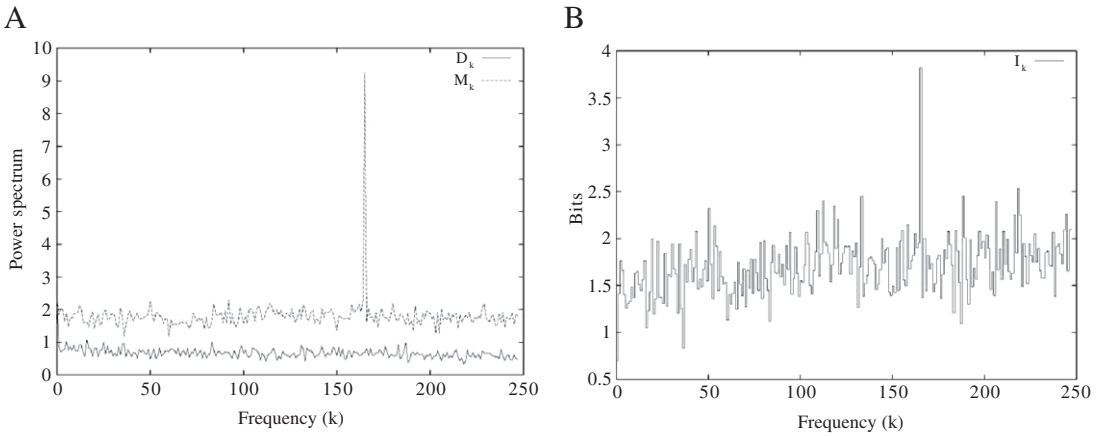

Figure 3. A. Average power spectra of the mean signal ($\hat{M}_k$) and the difference signal ($\hat{D}_k$) for a test set. **B.** Useful information contents I_k per frequency k.

By adding the information contents (Equation 16) of all coefficients, we obtain an estimate of 1.70 bits of useful information for homology matching. This number can be compared to the upper limit of 1.99 bits/base for the information contents of DNA, derived from Equation 3 under the assumption that each base was chosen independently, according to its observed frequency.

CONCLUSIONS

We have described a method for estimating the amount of information contained in a DNA sequence that can be used to identify homologous blocks, in spite of mutations and acquisition errors. This parameter (the mutual information content) allows us to estimate the probability of false positives - strings that are not homologous to the given sequence, but are just as similar to it as the homologous ones.

ACKNOWLEDGMENTS

Research supported partially by the Brazilian agencies CNPq and CAPES. We thank Gabriel Landini, pioneer in Fourier analysis of DNA, for motivating us to pursue this line of research.

REFERENCES

Anasstassiou, D. (2002). *Digital Signal Processing of Biomolecular Sequences*. Technical Report, CU/EE/TR2000-20-042. Department of Electrical Engineering, Columbia University, NY, USA.

Cheever, E.A., Karunaratne, W., Searls, D.B. and **Overton, G.C.** (1989). Using signal processing techniques for DNA sequence comparison. *Proceedings of the Northeast Bioengineering Conference*, Boston, MA, USA, pp. 173-174.

Lathi, B.P. (1968). *Communication Systems.* John Wiley & Sons, New York, USA.

Leitão, H.C.G. and **Stolfi, J.** (2002). A Multi-Scale Method for the Re-Assembly of Two-Dimensional Fragmented Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*TPAMI*) *9*: 1239-1251.

Meidanis, J. and **Setubal, J.C.** (1997). *Introduction to Computational Molecular Biology.* PWS Publishing Company, Pacific Grove, CA, USA.