Correlation Finder

General

CORRELATION FINDER is a free software which allows to exhaustively seek correlations between nucleotides in genomic sequences. It permits to analyze generic DNA sequences and genic sequences where the codon phase needs to be taken into account. Its graphic interface allows to easily set the parameters that characterize the motifs being sought.

Background

A large number of genomic sequences have recently become available. According to information theory, if a sequence has no function it is not written observing a language that follows syntactical rules [Shannon 1948]. So it is expected that succession of nucleotides of the sequence is erratic; that is, all nucleotides and all words appear with about the same frequency. Moreover correlations among the nucleotides or among the words are not expected. Instead unexpected patterns could lie in non-genic sequences and this can be demonstrated by the existence of correlations between nucleotides at various distances [Peng 1995]. Methods like fast Fourier transform (FFT) or detrended fluctuation analysis (DFA) reveal the presence of nucleotidic relationships but do not show the structures of the motifs responsible for the regularities.
Genic sequences also exhibit correlations [Luo 1991, Peng 1995] due to the degeneration of the genetic code that does not specify all genic bases and so synonymous codons can be used with different frequencies in the coding sequences. This phenomenon is called polarization of the genetic code or 'codon bias' and seems to be species-specific [Sharp et al. 1988]. Moreover there are correlations between codons and single near nucleotides that give rise to regularities known as codon bias and context-dependent codon bias [Fedorov et al. 2002]. These characteristics are the redundancy of a language and could be of use to increase the robustness or to specify further languages coexisting with the aminoacidic one. Degenerate and non-degenerate nucleotides seem to make up a context specifying the information for the splicing process [Pagani 2003]. Sequence regularities may be involved in other functions like chromatin organization, cell differentiation, regulation of mRNA lifetime, transport, folding, and translation velocity. Correlation Finder was developed to reveal correlations between nucleotides.

Systems

Correlation Finder is written in Borland Delphi v.6 and runs on ix86 compatible processors under Microsoft Windows as well as on Apple Macintosh, Linux and Unix-based platforms using Windows emulator software with one of the required Microsoft Windows versions.

Installation

CORRELATION FINDER is compound from the following files:
Correlation_finder.exe : Executable file
Correlation_finder.chm, Correlation_finder.hhp : Help files
fasta_input.fas : Example of an input fasta format file
plain_text_input.txt : Example of an input plain text format file
output.txt : Example of an output text format file

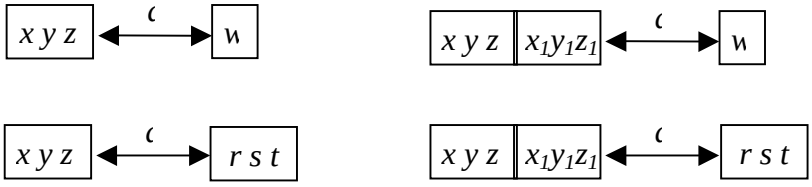
On the web there are zip archive and separete files. To install the software you should download and uncompress the files into a directory. Alternatively you can download files separately. At least you should run the executable file.

Characteristics

We made the software so it would also optimize the processing time.
Primarly we have implemented a RAM buffer to minimize hard disk access, so when the processing starts, a large amount of data are read from the file and stored in RAM memory, rather than to read one sequence at a time.
Secondarily we implemented indexing procedures to quick search substrings into the sequences.
This tool can handle sequences up to about one thousand million of nucleotides and on the whole datasets of several thousand millions of nucleotides. The software can also process a dataset of a single sequence. The possibility of working with very long sequences gives the opportunity to analyze also a long and single chromosomal stretch. The possibility of working with a wide dataset gives the opportunity to analyze many and relatively long sequences like human introns.

What it does

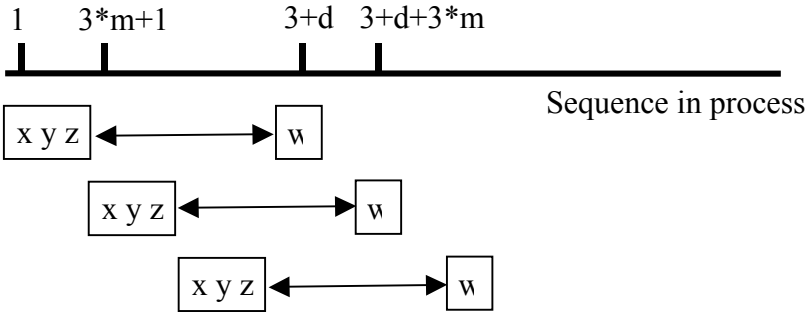
Correlation Finder was developed to reveal correlations between nucleotides. It reads both fasta and plain text format files containing one or more sequences to be analyzed and it is not case sensitive. The sequences can have different length. It is possible to set the kind of correlation to search: between triplets; triplet and nucleotide; two consecutive triplets and a nucleotide; two consecutive triplets and another triplet.



In the scheme above, *xyz*, $x_1y_1z_1$, *rst*, are generic codons and *w* a generic nucleotide.

The user can set the minimum and maximum distance *d* between the words taking part of the correlation being sought. The first sequence can also be sought by analyzing the in frame sequences of the input file, which is especially useful when investigating coding sequences. This is done setting the flag 'Implier in frame'. In this case it considers that input sequences start from phase 0.
The program can remove the bias due to the different nucleotide frequency *w* as a function of the phase. By setting the flag 'Reject sequences with STOP codons in frame', input sequences having nonsense codons in frame can be rejected; this is helpful when input datasets involuntarily contain coding sequences that are not in frame. Maximum distance d is 60 nucleotides. An output text file is produced showing for each motif: R, C, F(*xyz*,*d*,*w*), F_d(*xyz*) and F_d(*w*) values. Windows on the left side of the panel show the progress of the computing and the work that remains to be done.

Let's define the motif *(xyz,d,w)* as a triplet *xyz* and a nucleotide *w* at a distance d downstream *xyz*. For example: (cag,3,a) corresponds to the motif 'cag..a'.
The following scheme shows an example of sliding of the windows used to compute the correlation of a generic motif *(xyz,3,w)*. The sliding is of three nucleotides starting from the beginning so the position of *xyz* corresponds to the codon. The windows run until the end of each row is reach.



When seeking correlations between a triplet and a nucleotide not considering the phase, the software computes these values:

the frequency of the triplet *xyz*

$$F_d(xyz) = \frac{\sum_{col=1}^{\max\_col-d-2;} \sum_{row=1}^{\max\_row;} O(xyz_{col})}{\max\_row * (\max\_col - d - 2)}$$

the frequency of the nucleotide *w*

$$F_d(w) = \frac{\sum_{col=3+d}^{\max\_col;} \sum_{row=1}^{\max\_row;} O(w)}{\max\_row * (\max\_col - d - 2)}$$

the frequency of the motif (*xyz*,*d*,*w*)

$$F(xyz,d,w)=\frac{\sum_{col=1}^{\max\_col-d-2;\ \max\_row;}\ \sum_{row=1} O(xyz_{col},d,w)}{\max\_row*(\max\_col-d-2)}$$

the relative abundance R of the motif (*xyz,d,w*) as

$$R(xyz,d,w)=\frac{F(xyz,d,w)}{F_d(xyz)*F_d(w)}$$

the conditional probability C as

$$C(xyz,d,w)=\frac{F(xyz,d,w)}{F_d(xyz)}$$

where

Max_col: maximum length of each sequence of the input file

Max_row: number of sequences of the input file

O(*xyz*): number of ocurrences of the *xyz* codon in a sequence

O(xyz_{col}): number of ocurrences of the *xyz* codon in the position *col* of a sequence. This can be 0 or 1.

O(*w*): number of ocurrences of the nucleotide *w* in a sequence

O(*xyz,d,w*): number of ocurrences of the motiv *(xyz,d,w)* in a sequence

xyz : aaa, aac, aag, ... ttt

w: a,c,g,t.

d: it ranges from the minimum to the maximum distance setted in the program

When seeking of correlations between a triplet and a nucleotide considering the phase, the software computes these values:

the frequency of the triplet *xyz*

$$F_d(xyz)=\frac{\sum_{col=1\ldots1+3*m}^{DIV[(\max\_col-d),3]*3-2;\ \max\_row;}\ \sum_{row=1} O(xyz_{col})}{\max\_row*DIV[(\max\_col-d),3]}$$

the frequency of the nucleotide *w*

$$F_d(w)=\frac{\sum_{col=3+d\ldots3+d+3*m}^{DIV[(\max\_col-d),3]*3+1;\ \max\_row;}\ \sum_{row=1} O(w)}{\max\_row*DIV[(\max\_col-d),3]}$$

the frequency of the motif (*xyz*,*d*,*w*)

$$F(xyz,d,w)=\frac{\sum_{col=1\ldots 1+3*m}^{DIV[(\max\_col-d),3]*3-2;}\ \sum_{row=1}^{\max\_row;} O(xyz_{col},d,w)}{\max\_row*DIV[(\max\_col-d),3]}$$

the relative abundance R of the motif (*xyz*,*d*,*w*) as

$$R(xyz,d,w)=\frac{F(xyz,d,w)}{F_d(xyz)*F_d(w)}$$

the conditional probability C as

$$C(xyz,d,w)=F(w\,|\,(xyz,d))=\frac{F(xyz,d,w)}{F_d(xyz)}$$

DIV(a,b) is the integer division without remainder

Input

Example of input text format file:

```
gtaagagggcgcccaccacgtggccagggcgggacaccgaggcactgacgcctccctgcccccag
gtaaagtcagcttactaactttggggaaaggaaattaagtcatcataccaaaagttttttttattatag
gtaaggggaatgggtgtcctacagaggggttgcagcggggatgggtgctcagtggtcttctcccgatcag
gtaagatgtggaggggaagagggtgggaggaggagtggccggtgctgaccacccccctactgggcccccccag
gtaatttaattctgttctctttatttttgttcaatataagggcttgcttctaactggggcatttattgtag
gtaagtgggccctgggagtgggtggggtggggtgggtccaggcctcttgctgaaggctagacttccacgcag
gtaagtttaaaataacccaaattgctccttggatttccttcagtttattaaactctgttgcttccctttcag
gtaagaaggcagacacttggcattttggttctaattttggtagcctttcttaatcattgcatttatattttag
gtaagaggaggctgagggtcaagcagggcatccaagggggccagcctgacttccttccctccatgtcccacag
```

Example of input Fasta format file:

```
>intron_1
gtaagagggcgcccaccacgtggccagggcgggacaccgaggcactgacgcctccctgcccccag
>intron_2
gtaaagtcagcttactaactttggggaaaggaaattaagtcatcataccaaaagttttttttattatag
>intron_3
gtaaggggaatgggtgtcctacagaggggttgcagcggggatgggtgctcagtggtcttctcccgatcag
>intron_4
gtaagatgtggaggggaagagggtgggaggaggagtggccggtgctgaccacccccctactgggcccccccag
>intron_5
```

```
gtaatttaattctgttctctttatttttgttcaatataagggcttgcttctaactggggcatttattgtag
>intron_6
gtaagtgggccctgggagtgggtggggtggggtgggtccaggcctcttgctgaaggctagacttccacgcag
>intron_7
gtaagtttaaaataacccaaattgctccttggattttccttcagtttattaaactctgttgcttcctttcag
>intron_8
gtaagaaggcagacacttggcattttggttctaattttggtagcctttcttaatcattgcatttatattttag
>intron_9
gtaagaggaggctgagggtcaagcagggcatccaagggggccagcctgacttccttccctccatgtcccacag
```

Output

Example of output file:

xyz	d	w	R	C	F(xyz,d,w)	F(xyz)	F(w)
aaa	1	a	4,7925	0,3818	0,0084	0,0221	0,2373
aaa	1	c	3,3280	0,1501	0,0033	0,0221	0,2443
aaa	1	g	0,4486	0,2247	0,0050	0,0221	0,2561
aaa	1	t	0,6996	0,2434	0,0054	0,0221	0,2623
aac	1	a	1,4852	0,3524	0,0036	0,0101	0,2373
aac	1	c	1,1300	0,2760	0,0028	0,0101	0,2443
aac	1	g	0,2165	0,0554	0,0006	0,0101	0,2561
aac	1	t	1,2049	0,3161	0,0032	0,0101	0,2623
aag	1	a	0,0136	0,2769	0,0045	0,0161	0,2373

Interpreting the results

C*(xyz,d,w)* is the probability that there is the nucleotide *w* at distance *d* from *xyz*, given that there is *xyz*. It can range from 0 to 1 if the secondary event respectively never or always occurs. So we could have that *aag* codon has at distance 3 the nucleotides *a*, *c*, *g*, *t* respectively with frequencies of 35%, 27%, 20%, 18%. It means that the codon *aag* prefers the nucleotide *a* at distance 3 rather than *t*.

R*(xyz,d,w)* represents the ratio between observed and expected motif frequency, under the hypothesis of indipendence of the observed positions. It can range from 0 to a higher than 1 value. It is 0 if there is not the motiv *(xyz,d,w)* in the dataset. It is 1 or higher if the observed frequency is respectively equal or higher to the expected. Values far from 1 point out the presence of correlations.

It does not mean that a motif with a high R value is highly abundant in the dataset, but only that it is more frequent than expected.

Significance analysis

According to previous works [Karlin et al. 1994; Karlin and Ladunga 1994; Karlin and Cardon 1994; Karlin and Burge 1995], $R \leq 0.78$ and $R \geq 1.23$ are considered respectively the extreme under-representation and over-representation values, with $P \leq 0.001$ and with dataset of length $\geq$ 20Kb. Values from 0.79 to 0.82 are considered marginally low and values from 1.20 to 1.23 are considered marginally high.

To test the significance of datasets of length<20Kb the user should perform a Monte Carlo simulation, that is specific for each dataset composition [Fedorov et al. 2002].

Glossary

Correlation: Link between nucleotides. There is a correlation between nucleotides when the presence of a particular base (implier) implies or promotes the presence of another particular (implied) base in a different position. If the distance between correlated nucleotides is relatively short (about less than 10), they can have a functional role, for example be involved in a binding site. Long distance correlations in transcription sequences can be involved in mRNA secondary structure.

Codon bias: Preferential codon usage within synonymous groups. Coding sequences are biased in their usage of synonymous codons. It is hypothesised to be a consequence of either mutational pressure or translational selection. Highly expressed genes have a very strong bias, in which only very few codons are systematically used. High bias can confer high translation efficiency to sequences.

Conditional probability (C): probability that the second event B happens given that the first event A has happened or probability of B given A.
$C = P(B|A) = P(A \text{ and } B) / P(A)$

Information Theory: Efficiency in a language is the ability to transfer or memorize information using the smallest possible number of symbols, whereas redundancy is the loss of efficiency caused by the presence of correlations and different frequencies of symbols or words. According to the information theory, the more chaotic the succession of symbols of a language, the greater its efficiency but the less robust the language in terms of the ability to preserve/transfer information in the presence of noise. Natural languages tend to reach a balance between efficiency and robustness; redundancy is therefore a characteristic of natural languages.

Language: The protein coding regions of genes are highly constrained by the presence of at least two languages, one specifying the amino acid by defining the codon (coding function) and the second regulating the splicing process by defining some codons among their synonyms; this contributes to the formation of enhancer or silencer regulatory elements which allow exons to be recognized as constitutive or alternative. The two languages are able to coexist because the genetic code is degenerate and the splicing language can use bases that are not constrained by the first language. Since not all exonic synonymous mutations affect splicing efficiency, the splicing language obviously does not specify all synonymous codons and can drive splicing by creating alternative signals with equivalent function. Consequently, other constraints or languages can be present in gene coding sequences. In fact, there is evidence that synonymous codon usage is partly

constrained by the isochore composition of the region in which the gene lies. Some codons create motifs to bind proteins involved in transcription, in the export of mRNA towards the cytoplasm, and in translation. In addition, coding sequences seem to be overloaded with functions whereas the opposite is true of introns, as they contain few splicing signals while the rest have a weak regulatory role.

Relative abundance (R): observed / expected motifs ratio. If two events A and B are independent, P(A and B)=P(A)*P(B).
So R = P(A and B) / [P(A)*P(B)] = 1
The more R value is different from 1 the more a correlation between A and B could exist.

REFERENCES

Fedorov A, Saxonov S, Gilbert W. (2002) Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res. **30**:1192-1197.

Karlin S, Cardon LR. (1994) Computational DNA sequence analysis. *Annu Rev Microbiol.* **48**:619-654. Review.

Karlin S, Ladunga I. (1994) Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci U S A.* **91**:12832-12836

Karlin S, Ladunga I, Blaisdell BE. (1994) Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci U S A.* **91**:12837-12841.

Karlin S, Burge C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**:283-290. Review.

Luo LF, Li H. (1991) The statistical correlation of nucleotides in protein-coding DNA sequences. Bull Math Biol.;**53**:345-353

Pagani F, Buratti E, Stuani C, Baralle FE. (2003) Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. J Biol Chem. Jul **18**;278(29):26580-26588.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Simons M, Stanley HE. (1995) Statistical properties of DNA sequences. Physica A.;**221**:180-192

Shannon C. E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal* **27**, 379–423, 623–656

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. (1988) Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Res **16**:8207-8211*

Authors:

Francesco Piva, Giovanni Principato
Polytechnic University of Marche
Institute of Biology and Genetics
ITALY