Information-Based Similarity Index

Albert C.-C. Yang, Ary L Goldberger, C.-K. Peng

This document can also be read on-line at http://physionet.org/physiotools/ibs/doc/. The most recent version of the software described within is freely available from http://physionet.org/physiotools/ibs/src/.

Background

Physiologic systems generate complex fluctuations in their output signals that reflect the underlying dynamics. Therefore, finding and analyzing hidden dynamical structures of these signals are of both basic and clinical interest. One approach to discovering such hidden information is to analyze the repetitive appearance of certain basic patterns that are embedded in the complete signals. We developed a novel information-based similarity index to detect and quantify such temporal structures in the human heart rate time series using tools from statistical linguistics [1].

Methods

Human cardiac dynamics are driven by the complex nonlinear interactions of two competing forces: sympathetic stimulation increases and parasympathetic stimulation decreases heart rate. For this type of intrinsically noisy system, it may be useful to simplify the dynamics via mapping the output to binary sequences, where the increase and decrease of the inter-beat intervals are denoted by 1 and 0, respectively. The resulting binary sequence retains important features of the dynamics generated by the underlying control system, but is tractable enough to be analyzed as a symbolic sequence.

Consider an inter-beat interval time series, $\{x_1, x_2, \cdots, x_N\}$, where x_i is the i-th inter-beat interval. We can classify each pair of successive inter-beat intervals into one of the two states that represents a decrease in x, or

an increase in x. These two states are mapped to the symbols 0 and 1, respectively

$$I_n = \begin{cases} 0, & \text{if } x_n \leq x_{n-1} \\ 1, & \text{if } x_n > x_{n-1}. \end{cases} \quad (1)$$

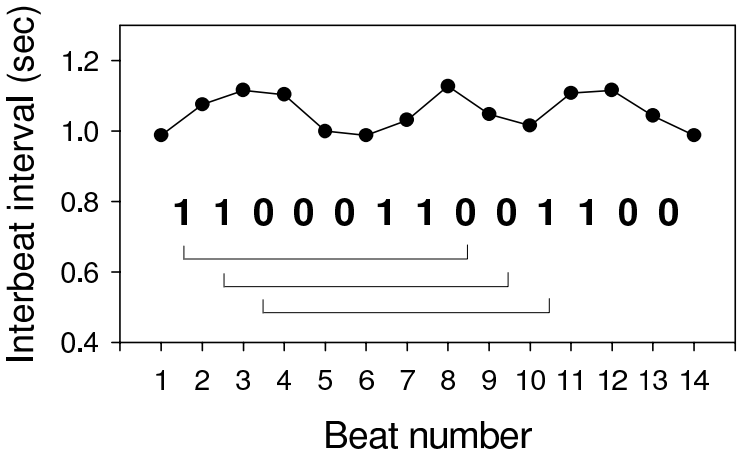

Figure 1: Schematic illustration of the mapping procedure for 8-bit words from a heartbeat time series.

We map $m + 1$ successive intervals to a binary sequence of length m, called an m-bit "word." Each m-bit word, w_k, therefore, represents a unique pattern of fluctuations in a given time series. By shifting one data point at a time, the algorithm produces a collection of m-bit words over the whole time series. Therefore, it is plausible that the occurrence of these m-bit words reflects the underlying dynamics of the original time series. Different types of dynamics thus produce different distributions of these m-bit words.

In studies of natural languages, it has been observed that authors have characteristic preferences for the words they use with higher frequency. To apply this concept to symbolic sequences mapped from the inter-beat interval time series, we count the occurrences of different words, and then sort them in descending order by frequency of occurrence.

The resulting rank-frequency distribution, therefore, represents the statistical hierarchy of symbolic words of the original time series. For example, the first rank word corresponds to one type of fluctuation which is the most frequent pattern in the time series. In contrast, the last rank word defines the most unlikely pattern in the time series.

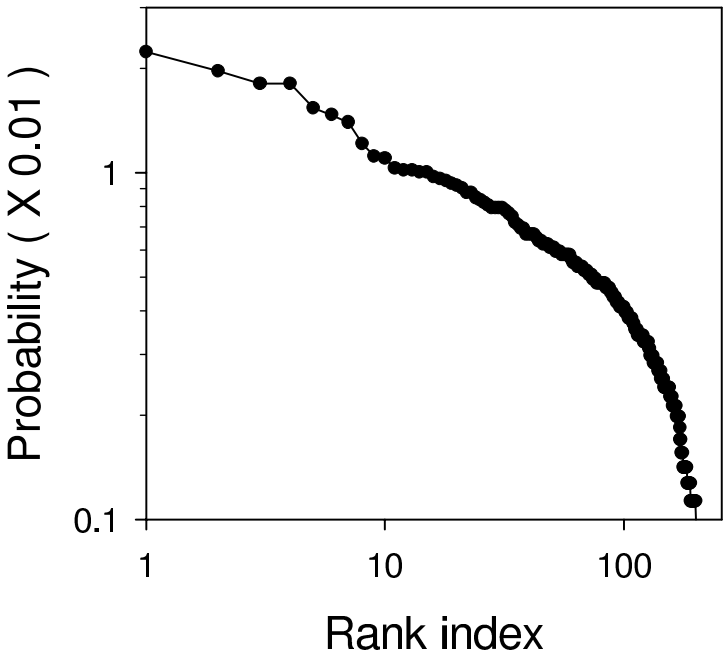

Figure 2: The linear regime (for rank ≤ 50).

To define a measurement of similarity between two signals, we plot the rank number of each m-bit word in the first time series against that of the second time series.

If two time series are similar in their rank order of the words, the scattered points will be located near the diagonal line. Therefore, the average deviation of these scattered points away from the diagonal line is a measure of the "distance" between these two time series. Greater distance indicates less similarity and vice versa. In addition, we incorporate the likelihood of each word in the following definition of a weighted distance, D_m, between two symbolic sequences, S_1 and S_2.

$$D_m(S_1, S_2) = \frac{1}{2^m - 1} \sum_{k=1}^{2^m} |R_1(w_k) - R_2(w_k)| \, F(w_k) \tag{2}$$

where

$$F(w_k) = \frac{1}{Z}[-p_1(w_k) \log p_1(w_k) - p_2(w_k) \log p_2(w_k)]. \tag{3}$$

Here $p_1(w_k)$ and $R_1(w_k)$ represent probability and rank of a specific word,

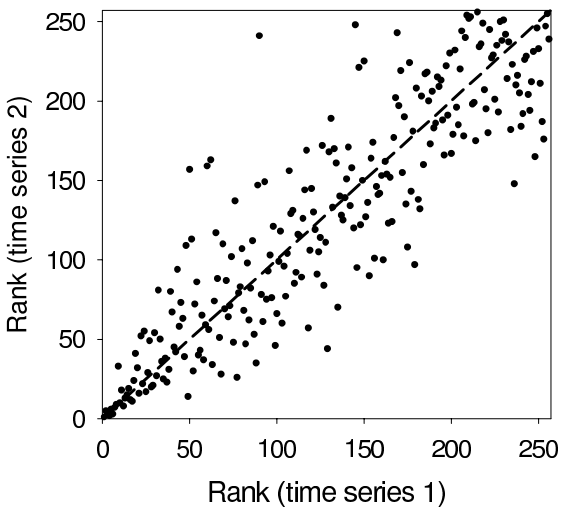

Figure 3: Rank order comparison of two cardiac inter-beat interval time series from same subject. For each word, its rank in the first time series is plotted against its rank in the second time series. The dashed diagonal line indicates the case where the rank-order of words for both time series is identical.

w_k, in time series S_1. Similarly, $p_2(w_k)$ and $R_2(w_k)$ stand for probability and rank of the same m-bit word in time series S_2. The absolute difference of ranks is multiplied by the normalized probabilities as a weighted sum by using Shannon entropy as the weighting factor. Finally, the sum is divided by the value $2^m - 1$ to keep the value in the same range of [0, 1]. The normalization factor Z in Eq. 3 is given by $Z = \sum_k [-p_1(w_k) \log p_1(w_k) - p_2(w_k) \log p_2(w_k)]$.

Example

Here we provide a concise example to demonstrate how to calculate an information-based similarity index between two time series. The following figure illustrates a sample heartbeat interval time series from a healthy subject (left panel) showing complex variability. In contrast, a time series from a CHF subject (right panel) shows less variability. Both sample time series contain 1000 inter-beat intervals. (See RR Intervals, Heart Rate, and HRV Howto for information on how to obtain additional inter-beat interval time series in this format.)

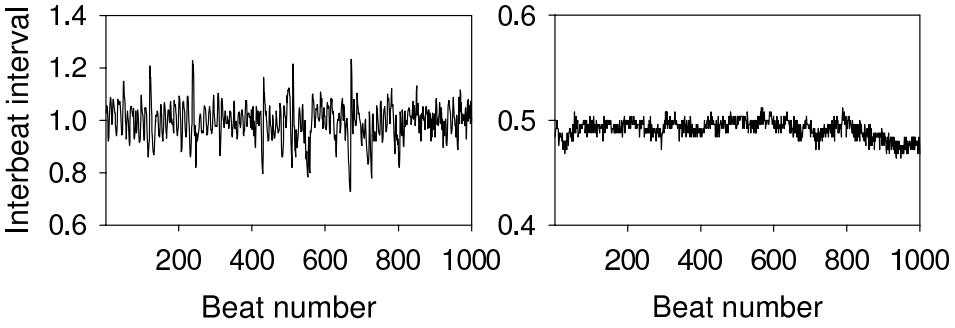

Figure 4: Representative inter-beat interval time series for (a) a healthy subject, and (b) a subject with congestive heart failure (CHF).

We first map each signal to a binary sequence according to the increment of consecutive inter-beat intervals. Suppose we set the m equal to 8, then there will be $2^8 = 256$ different 8-bit words. We count the occurrences of each 8-bit word, and then sort them by descending frequency. The resulting rank-frequency distribution represents the statistical hierarchy of repetitive patterns of a given time series. For example, the top-ranked 8-bit words correspond to the most frequently occurring patterns in a given heartbeat time series. In contrast, the last ranked word defines the rarest patterns.

8-bit words w_k	$R_1(w_k)$	$R_2(w_k)$	$p_1(w_k)$	$p_2(w_k)$	$H_1(w_k)$	$H_2(w_k)$
00000000	25	117	0.014423	0.001603	0.061138	0.010314
00000001	8	132	0.002404	0.004006	0.014497	0.022115
00000010	93	80	0.023237	0	0.087418	0
00000011	3	179	0.004808	0.007212	0.025661	0.035568
00000100	44	35	0.002404	0.003205	0.014497	0.018407
00000101	91	92	0.004808	0.000801	0.025661	0.005713
00000110	47	140	0.029647	0	0.104311	0
00000111	1	176	0.003205	0.009615	0.018407	0.044658
00001000	66	23	0.003205	0.004006	0.018407	0.022115
00001001	64	81	0	0.00641	0	0.032371
...	...	...	...	...	...	...

Table 1: $H(w_k) = -p(w_k) \log p(w_k)$ is the Shannon entropy.

The rank order difference between two time series can be visualized by

plotting the rank number of each 8-bit word in the first time series against that of the second time series. The dashed diagonal line indicates the case where the rank order of words for both time series is identical.

As demonstrated by the above rank order comparison map, the "distance" (or dissimilarity) between any two time series can be quantified by measuring the scatter of these points from the diagonal line in the rank order comparison plot. By applying Eq.1 to the rank-order frequency list obtained from the sample time series, we obtained an information-based similarity index equal to 0.412725. Using the example data files provided with the `ibs` software, this result may be reproduced by running the command

```
ibs 8 healthy.txt chf.txt
```

Applications

Phylogenetic Tree of Human Heart Beats

We applied this distance measurement to RR interval time series, each at least 2 hours in length, from 40 ostensibly healthy subjects with subgroups of young (10 females and 10 males, average 25.9 years) and elderly (10 females and 10 males, average 74.5 years), a group of subjects (n = 43) with severe congestive heart failure (CHF) (15 females and 28 males, average 55.5 years) and a group of 9 subjects with atrial fibrillation (AF). We measured the average distance between subjects across different groups. We defined the inter-group distance of groups A and B as the average distance between all pairs of subjects where one subject is from group A and the other subject is from group B. We calculated the inter-group distances among all groups of our time series as well as a group of 100 artificial time series of uncorrelated noise (white noise group).

The method for constructing phylogenetic trees is a useful tool to present our results since the algorithm arranges different groups on a branching tree to best fit the pairwise distance measurements. Here we show the result of a rooted tree for the case of $m = 8$.

We note that the structure of the tree is consistent with the underlying physiology: the further down the branch the more complex the dynamics are. The groups are arranged in the following order (from bottom to top as shown in the above figure): 1) Time series from the healthy young group represent dynamical fluctuations of a highly complex integrative control system. 2) The

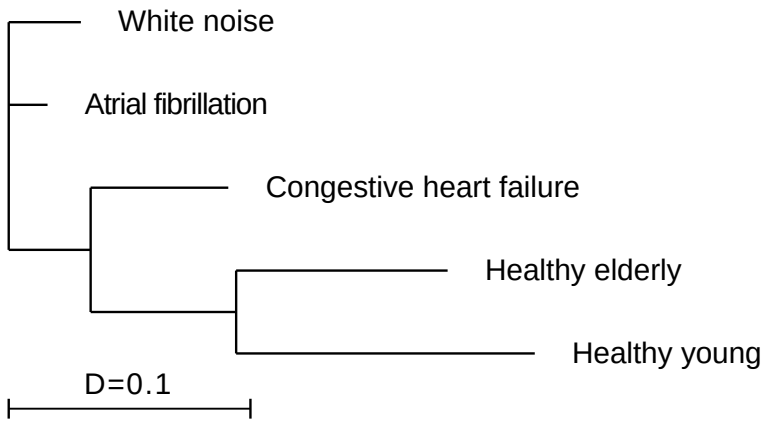

Figure 5: A rooted phylogenetic tree generated according to the distances between different groups. White noise indicates simulated uncorrelated random time series.

healthy elderly group represents deviation from the "optimal" youthful state, possibly due to decoupling (or drop-out) of components in the integrative control system. 3) Severe damage to the control system is represented by the CHF group. These individuals have profound abnormalities in cardiac function associated with pathologic alterations in both the sympathetic and parasympathetic control mechanisms that regulate beat-to-beat variability. 4) The AF group is an example of a pathologic state in which there appears to be very limited external input on the heartbeat control system. 5) The artificial white noise group represents the extreme case in that only noise and no signal is present. This example demonstrates that the physiologic complexity of human heart beat dynamics can be robustly described by our information categorization method.

Written texts and genetic sequences

The generic concept underlying the information categorization method makes it applicable to a wide range of problems. Recently, as a further proof of principle application, we applied this approach to address a long-standing authorship debate related to Shakespeares plays [2]. This work was featured in the Boston Globe (Aug. 5, 2003) and was the basis for the award-winning entry in the Calvin & Rose G. Hoffman Marlowe Memorial Trust 2003 Prize.

In addition to being a new approach to forensic text analysis, this method has potential applications in genetic sequence analysis [3].

Information-Based Similarity Software

The software may be obtained from `http://www.physionet.org/physiotools/ibs/`, where you will find `ibs.c` (the source for the software), a `Makefile`, two data files (`healthy.txt` and `chf.txt`) and a file named `ibs.expected`. Download all of these files.

If you have a `make` utility, you can use it to compile and test the software, simply by typing "`make check`" (look in `Makefile` to see what this command does). Otherwise, compile `ibs.c` and link it with the C standard math library (needed for the `abs` and `log` functions only). For example, if you use the GNU C compiler (recommended), you can do this by:

```
gcc -o ibs -O ibs.c -lm
```

Test the program by running the command:

```
ibs 8 healthy.txt chf.txt
```

If the current directory is not in your PATH, you may need to type the location of `ibs`, as in

```
./ibs 8 healthy.txt chf.txt
```

The output should match the contents of `ibs.expected`. For brief instructions about how to run the program, type its name at a command prompt:

```
ibs
```

which should produce a message similar to:

```
usage: ibs M SERIES1 SERIES2
  where M is the word length (an integer greater than 1), and
  SERIES1 and SERIES2 are one-column text files containing the
  data of the two series that are to be compared.  The output
  is the information-based similarity index of the input series
  evaluated for M-tuples (words of length M).

  For additional information, see
         http://physionet.org/physiotools/ibs/.
```

This program reads two text files of numbers, which are interpreted as values of two time series. Within each series, pairs of consecutive values are compared to derive a binary series, which has values that are either 1 (if the second value of the pair was greater than the first) or 0 (otherwise). A user-specified parameter, m, determines the length of "words" (m-tuples) to be analyzed by this progam.

Within each binary series, all m-tuples of consecutive values are treated as "words"; the function counts the occurrences of each of the 2^m possible "words" and then derives the word rank order frequency (WROF) list for the series. Finally, it calculates the information-based similarity between the two WROF lists, and outputs this number. Depending on the input series and on the choice of m, the value of the index can vary between 0 (completely dissimilar) and 1 (identical).

References

1. Yang AC, Hseu SS, Yien HW, Goldberger AL, Peng CK: Linguistic analysis of the human heartbeat using frequency and rank order statistics. Phys Rev Lett 2003, 90: 108103.

2. Yang AC, Peng C-K, Yien H-W, Goldberger AL: Information categorization approach to literary authorship disputes. Physica A 2003, 329: 473–483.

3. Yang AC, Goldberger AL, Peng CK: Genomic Classification Using a New Information-Based Similarity Index: Application to the SARS Coronavirus. Phys Rev E (submitted).