Detecting rhythms in time series with RAIN Supplementary Online Material

Paul F. Thaben and Pål O. Westermark

Institute for theoretical biology
Charité – Universitätsmedizin Berlin
10115 Berlin, Germany

Algorithm Description

Here, we explain the method used in RAIN, starting with the well-known Mann-Whitney Statistic and extend it through several steps to the special case capable of detecting oscillations. Series of data measurements are expected to be regular, equally spaced, and to come in temporal order. Repeated measurements are grouped together for each time point.

Let $(X_{11}, \ldots, X_{1m_1}), \ldots, (X_{n1}, \ldots, X_{nm_n})$ be a set of n samples of size $m_1, \ldots, m_n$ taken from different populations $F_1(x), \ldots, F_n(x)$ with a total number

$$N = \sum_{c=1}^{n} m_c \,. \tag{1}$$

Empty sets with $m_c = 0$ are also allowed and do not affect the statistic.

There are many methods for testing special relations between these populations against $H_0 : F_i = F_j \, (i \neq j \in 1, ..., n)$. Many of them use a method described by Mann and Whitney, and independently Wilcoxon, to compare two samples of different populations i and j, defining

$$q_{i_k,j_l} = \begin{cases} 1 & \text{if } X_{ik} < X_{jl} \\ 0 & \text{else} \end{cases} , \tag{2}$$

and

$$U_{ij} = \sum_{k=1}^{m_i} \sum_{l=1}^{m_j} q_{i_k,j_l}, \tag{3}$$

which is the Mann-Whitney U-statistic for comparison of two samples. There are alternative definitions of q_{i_k,j_l} where equality of X_{ik} and X_{jl} gives 0.5, but we use the stricter definition, as the following calculations of discrete probability densities are constructed for discrete values.

This Mann-Whitney U-statistic serves as a core element used for the following tests, dealing with complexer problems.

Jonckheere-Terpstra Test

First, the problem of two sets with different populations is extended to a series of populations. The Jonckheere-Terpstra Test (Jonckheere, 1954; Terpstra, 1952) concerns the alternative hypothesis

$$H_1 : F_1(x) < F_2(x) < \ldots < F_n(x), \tag{4}$$

with test statistic

$$s = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} U_{ij}. \tag{5}$$

To gather exact p-values, an efficient algorithm for the calculation of the exact probabilities for all possible scores s under the null hypothesis $F_1(x) = F_2(x) = \ldots = F_n(x)$ for a given set of samples $(m_1, \ldots, m_n)$ must be constructed.

The total number of possible permutations for each score $f(s)$, the frequency distribution, has to be determined. It is limited by the minimal and maximal values of the given statistic $s \in [0, s_{\max}]$. From the $f(s)$ the probability distribution $p(s)$ is given as

$$p(s) = \frac{f(s)}{\sum_{s_i=0}^{s_{\max}} f(s_i)}. \tag{6}$$

The calculation of $f(s)$ for distinct sample set be facilitated by the use of a generating function $G(z)$ which holds the values of $f(s)$ as coefficients:

$$G(z) = \sum_{s=0}^{s_{\max}} z^s \cdot f(s) \tag{7}$$

This generating function allows us to calculate the frequency distribution of combinations of independent tests. If two independent tests with $f_1(s_1)$ and $f_2(s_2)$ and generating functions $G_1(z)$ and $G_2(z)$ are combined by summation of the scores, the combined frequency distribution $f_c(s_1 + s_2)$ can be recovered from the product of the generating functions:

$$G_c(z) = \sum_{s=0}^{s_{\max}} z^s \cdot f_c(s) = G_1(z) \cdot G_2(z) \tag{8}$$

Following Kendall and Stuart (1961), we write the generating function for the Mann-Whitney U-statistic for two samples $m_1$and m_2 as

$$G(z) = \frac{\prod_{u=1}^{m_1+m_2}(1-z^u)}{\prod_{v_1=1}^{m_1}(1-z^{v_1}) \cdot \prod_{v_2=1}^{m_2}(1-z^{v_2})}. \tag{9}$$

The same authors state the formula for the Jonckheere-Terpstra test:

$$G(z) = \frac{\prod_{u=1}^{N}(1-z^u)}{\prod_{d=1}^{n}\prod_{v=1}^{m_d}(1-z^v)}. \tag{10}$$

$G(z)$ in this form (Equation (9) or (10)) has to be expanded into the polynomial form of Equation (7), where the coefficients of $G(z)$ represent $f(s)$. An efficient way to to do this was presented by Harding (1984). $G(z)$ in the form of Equation (9) or (10) is a product of a finite number of factors and only two different kinds of factors occur, $(1-z^u)$ and $(1-z^u)^{-1}$. The expansion is carried out by a successive multiplication of these factors. After each multiplication a new intermediate polynomial $G_r(z)$ is obtained. To simplify the calculation of each step, the polynomial is represented by a series of integers. Each number represents a coefficient of the polynomial.

$$\text{Example: } 1\,2\,3\,2 \rightarrow G_r(z) = 1z^0 + 2z^1 + 3z^2 + 2z^3$$

It is possible to simplify the calculation by only operating on these coefficients. The successive multiplications with the factors can be conducted by simple shifting and adding operations on this series of integers:

multiplication by $(1-z^u)$**:** Subtract from every element the element which is u positions to its left.

multiplication by $(1-z^u)^{-1}$**:** For this operation add to every element the sum off every u^{th} element to the left. This operation follows from the Taylor series $(1-x)^{-1} = \sum_{n=0}^{\infty} x^n$ with $z^u = x$.

The series of integers, although in general infinite, could be limited to $\lceil s_{\max}/2 \rceil$ for this algorithm, as no other elements are necessary for the calculation.

As an example, we calculate $f(s)$ for a Mann Whitney statistic $m_1 = m_2 = 2$ which gives $s_{\max} = 4$. Using Equation (9), we derive the product form of $G(z)$:

$$\frac{\prod_{u=1}^{4}(1-z^u)}{\prod_{v=1}^{2}(1-z^v) \cdot \prod_{v=1}^{2}(1-z^v)} = \frac{\prod_{u=3}^{4}(1-z^u)}{\prod_{v=1}^{2}(1-z^v)} = z^0 \cdot (1-z^3) \cdot (1-z^4) \cdot (1-z^1)^{-1} \cdot (1-z^2)^{-1} \tag{11}$$

Thereby, the factor $z^0 = 1$ at the beginning of the product acts as a first iteration to initiate the series:

$$G_o(z) = z^0 = 1z^0 + 0z^1 + 0z^2 + \ldots \rightarrow 1\,0\,0\ldots \tag{12}$$

Now we calculate the next iterations of the series stepwise for each factor, ending up with the representation of $G(z)$:

1	0	0	0	0	$z^0 = G_0(z)$
1	0	0	-1	0	$(1 - z^3) \cdot G_0(z) = G_1(z)$
1	0	0	-1	-1	$(1 - z^4) \cdot G_1(z) = G_2(z)$
1	1	1	0	-1	$(1 - z^1)^{-1} \cdot G_2(z) = G_3(z)$
1	1	2	1	1	$(1 - z^2)^{-1} \cdot G_3(z) = G(z)$

Thus, the generating function for the example is $G(z) = 1 \cdot z^0 + 1 \cdot z^1 + 2 \cdot z^2 + 1 \cdot z^3 + 1 \cdot z^4$, and in turn the probability distribution function is $p(0) = 1/6$, $p(1) = 1/6$, $p(2) = 1/3$, $p(3) = 1/6$, $p(4) = 1/6$.

With this algorithm for the calculation of the null distribution, tests against the alternative hypothesis of strictly rising or falling trends in a time series can be formulated, but moreover with reordering of the evaluated data, also other relationships could be evaluated as done by JTK_CYCLE for sinusoidals (Hughes et al., 2010). There is, however, a remaining problem with this approach: Every time point is tested against all other time points, so in the oscillating case, the rising and falling part of the oscillation are tested against each other according to a specific pre-determined wave form (a sinusoid in the default implementation). This limitation is alleviated by the umbrella approach using a combination of two almost independent Jonckheere-Terpstra tests.

Umbrella Alternatives

General Umbrella

Mack and Wolfe (1981) constructed the test for umbrella alternatives as an extension of the Jonckheere-Terpstra test with the alternative hypothesis that an index $1 < e < n$ exists such that

$$H_1 : F_1(x) < F_2(x) < \ldots < F_e(x) > \ldots > F_n(x) \tag{13}$$

For this case, the statistic is calculated as the sum of two independent Jonckheere-Terpstra statistics

$$s = \sum_{i=1}^{e-1} \sum_{j=i+1}^{e} U_{ij} + \sum_{i=e}^{n-1} \sum_{j=i+1}^{n} U_{ji}. \tag{14}$$

With Equation (8) and $N_{ij} = \sum_{k=i}^{j} m_k$, the generating function is the product of the generating functions of two independent Jonckheere-Terpstra tests:

$$G(z) = \frac{\prod_{u_1=1}^{N_{1e}} (1 - z^{u_1})}{\prod_{d=1}^{e} \prod_{v=1}^{m_d} (1 - z^v)} \cdot \frac{\prod_{u_2=1}^{N_{en}} (1 - z^{u_2})}{\prod_{d=e}^{n} \prod_{v=1}^{m_d} (1 - z^v)} \tag{15}$$

The calculation of $p(s)$ follows the Harding algorithm, as the factors are the same as for the Jonckheere-Terpstra test.

This is the general principle used in RAIN: An oscillation is modeled as a time series consisting of a rising part followed by a falling part. Some more refinements of the method are necessary to correctly handle the trough of the oscillation and to fit different phases.

Ring Shaped Alternatives

The general umbrella can be generalized to variable "partial order alternatives", as described by Steitberg and Röhmel (Streitberg and Röhmel, 1988). In particular, a ring-shaped extension of the umbrella alternatives is described, where there is no only a largest $F_e(x)$ but also a smallest population $F_1(x)$, which extends the alternate hypothesis to

$$H_1 : F_1(x) < F_2(x) < \ldots < F_e(x) > \ldots > F_n(x) > F_1(x). \tag{16}$$

The statistic is extended by a Mann-Whitney test between all groups of the falling part except $F_e(x)$ and the smallest element $F_1(x)$.

$$s = \sum_{i=1}^{e-1} \sum_{j=i+1}^{e} U_{ij} + \sum_{i=e}^{n-1} \sum_{j=i+1}^{n} U_{ji} + \sum_{j=e+1}^{n} U_{j1}. \tag{17}$$

This additional Mann-Whitney test reflects as an additional factor in the generating function:

$$G(z) = \frac{\prod_{u_1=1}^{N_{1e}}(1-z^{u_1})}{\prod_{d=1}^{e}\prod_{v=1}^{m_d}(1-z^v)} \cdot \frac{\prod_{u_2=1}^{N_{en}}(1-z^{u_2})}{\prod_{d=e}^{n}\prod_{v=1}^{m_d}(1-z^v)} \cdot \frac{\prod_{u_3=1}^{N_{(e+1)n}+m_1}(1-z^{u_3})}{\prod_{v=1}^{m_1}(1-z^v)\cdot\prod_{v=1}^{N_{(e+1)n}}(1-z^v)} \tag{18}$$

To detect rhythms based on the ring shaped alternatives the original data are regrouped into sets $F_1(x) \dots F_n(x)$. The period is given by the length n of this set, the shape of the peak by e, and different phases by the data put into the group $F_e(x)$. The preceding and following data are put into the preceding and following groups, in a circular manner ($F_{n+i}(x) \to F_i(x)$). Thereby, if data series cover more than one period, the groups contain data from more than one time point.

This approach is used in RAIN as the preset 'individual', as it provides a quite good power for different curve shapes in simulations. It has proved to be the most stable and strongest approach for individual measured data with the benefit that the rising and the falling part are treated independently from each other, so that no particular symmetry is assumed. For longitudinal data with strong additional time dependent effects, e.g. damping or underlying trends, a different approach could be used, as outlined in the next section.

Series of Umbrellas for Longitudinal Time Series

It is possible to define a series of inflections $1 = e_0 \leq e_1 \dots < e_g = n$ defining alternate rising and falling slopes, effectively representing a series of umbrellas:

$$H_1 : F_1(x) < \dots < F_{e_1}(x) > \dots > F_{e_2}(x) < \dots \tag{19}$$

In the case of an initial falling series $e_0 = e_1 = 1$, the first rising part is skipped and the following formulas are still valid. The calculation of the statistic is the summation of multiple Jonckheere-Terpstra statistics:

$$s = \sum_{\alpha=1}^{g} \sum_{a=e_{\alpha-1}}^{e_\alpha - 1} \sum_{b=a+1}^{e_\alpha} \begin{cases} U_{ab} & \text{if } \alpha \text{ odd} \\ U_{ab} & \text{if } \alpha \text{ even} \end{cases}. \tag{20}$$

Analogously, the calculation of the generating function extends to a product of the generating functions for the individual slopes

$$G(z) = \prod_{\alpha=1}^{g} \frac{\prod_{u_2=1}^{N_{e_{\alpha-1}e_\alpha}}(1-z^{u_2})}{\prod_{d=e_{\alpha-1}}^{e_\alpha}\prod_{v_d=1}^{m_n}(1-z^{v_d})}. \tag{21}$$

To test for an oscillation, the measurement groups are the groups in their original order. The oscillation itself is defined by the series of inflections. Thereby, all odd inflections (e_1, e_3, ...) represent peaks and all even inflections (e_2, e_4, ...) troughs. The oscillatory behavior is tested by a strict periodic setup of these inflections. The phase is represented by the first peak position, the period length by the space between peaks and the peak shape is the time from peak to trough.

Although this 'longitudinal' approach is weaker as the 'individual' approach, it was implemented in the R package, because disturbances by damping, underlying trends and the like, are better dealt with, and in some data sets we found some data missed by the previous algorithm. On the other hand, the differentiation between strong asymmetric behavior and pure trends is weak when using the 'longitudinal' mode.

Implementations for Oscillation Detection

When testing a time series for oscillatory behavior, a number of different parameters (phases, amplitudes, and peak shapes) are possible and have to been considered. All the tests for different parameter settings are done independently from each other. For each of these tests, a probability distribution function $p(s)$ is calculated in order to obtain the exact p-value. Thus, we can choose the best p-value and thereby the best matching parameter set. For the returned p-value, we have to take into account that we do multiple testing on the data, but that the different tests are not independent from each other. To correct for multiple testing, RAIN uses the adaptive BH algorithm presented by Benjamini and Hochberg (2000) for partially dependent multiple testing. The corrected p-value together with the phase, period and peak shape are returned by RAIN.

References

Benjamini Y, Hochberg Y (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J Educ Behav Stat*, 25:60–83.

Harding EF (1984) An efficient, minimal-storage procedure for calculating the Mann-Whitney U, generalized U and similar distributions. *J R Stat Soc, Ser C, Appl Stat*, 33:1–6.

Hughes ME, Hogenesch JB, Kornacker K (2010) JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms*, 25:372–380.

Hughes ME, Hong HK, Chong JL, Indacochea AA, Lee SS, Han M, Takahashi JS, Hogenesch JB (2012) Brain-specific rescue of clock reveals system-driven transcriptional rhythms in peripheral tissue. *PLoS Genet*, 8:e1002835.

Jonckheere AR (1954) A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41:133–145.

Kendall M, Stuart A (1961) The advanced theory of statistics. vol. 2. Griffin & Co., London.

Mack GA, Wolfe DA (1981) K-sample rank tests for umbrella alternatives. *J Am Statist Assoc*, 76:175–181.

Mauvoisin D, Wang J, Jouffe C, Martin E, Atger F, Waridel P, Quadroni M, Gachon F, Naef F (2014) Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver. *Proc Natl Acad Sci USA*, 111:167–172.

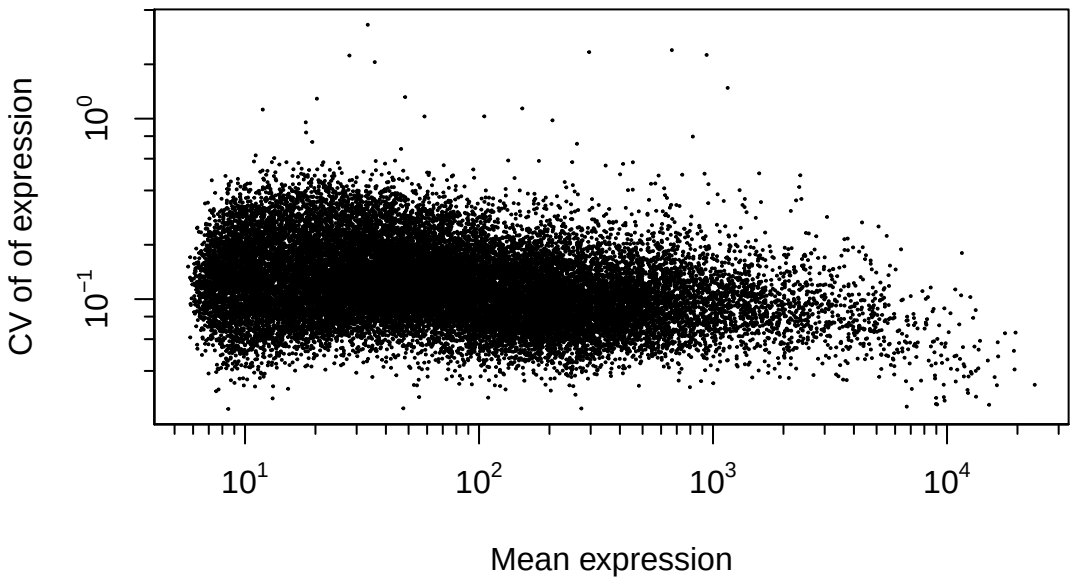
Robles MS, Cox J, Mann M (2014) In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS Genet*, 10:e1004047.

Streitberg B, Röhmel J (1988) Exact nonparametrics for partial order tests. *Comput Stat Q*, 1:23–41.

Terpstra T (1952) The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14:327–333.

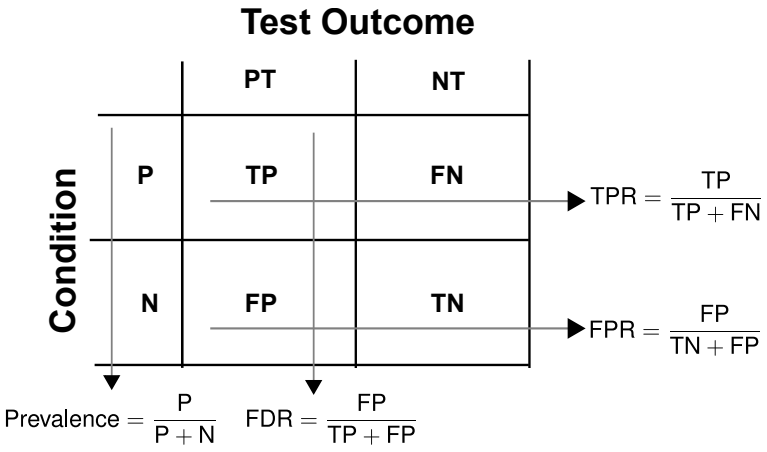
Supplementary Figures

Figure S1



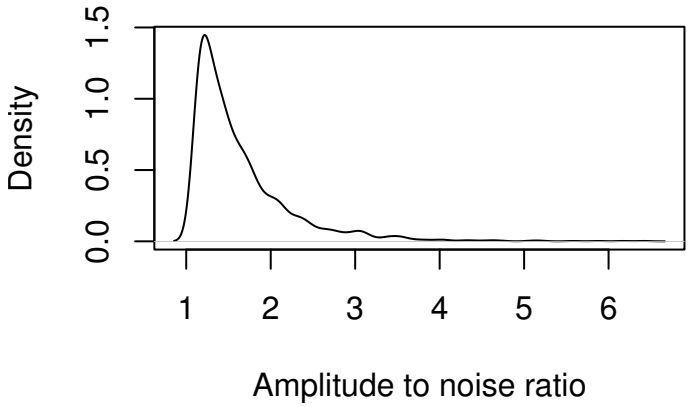
Coefficients of variation (CVs) for mouse liver microarray data. Noise standard deviations were estimated for each of the 25,817 genes in a microarray study encompassing 24 time points (Hughes et al., 2012), as described in the Methods section. The CV (standard deviation divided by mean) lies around 0.1 for most genes, regardless of expression level.

Figure S2



ROC curves and false discovery rates. ROC curves gauge the true positive rate as a function of the false positive rate (horizontal arrows). These pertain to test outcomes only and are independent of the prevalence, i.e. the proportion of true positives in the data. The false discovery rate (FDR, vertical arrow) measures the proportion of false positives in the true test outcomes, and is dependent on the prevalence: A lowered prevalence will increase the FDR for a given TPR, or decrease the TPR if the FDR is held constant.

Figure S3



Amplitude/Noise ratio for microarray data. Circadian relative amplitudes and noise standard deviations were estimated for each of the 25,817 genes in a microarray study encompassing 24 time points (Hughes et al., 2012), as described in the Methods section. Amplitude/noise ratios, i.e. amplitudes divided by the noise standard deviations, were computed for each gene and the resulting density estimate is plotted.

Figure S4 (separate PDF file)

ROC curves for different sampling rates and amplitude/noise ratios. We generated artificial data for noise standard deviation 0.1 and sine curve and sawtooth-shaped curve amplitudes 0 (control) 0.1, 0.2, and 0.3, respectively, as well as with "native" amplitudes sampled from the distribution estimated from microarray data (Methods and Figure S3). In addition, outliers (one random time point in each series is altered to $y(t) + 20$) were present or absent, and different sampling rates were evaluated: 2 hours (24 samples), 3 hours (16 samples), and 4 hours (12 samples), resulting in 24 different diagrams, where the results for RAIN and JTK_CYCLE for sine curves as well as sawtooth-shaped curves are plotted. Each curve is based on 100,000 artificial time series and 100,000 controls (amplitude 0). The light gray straight lines correspond to an FDR of 0.1 and prevalences (proportions of true positives) of 0.05, 0.1, and 0.25, respectively (Equation 2), or alternatively to a prevalence of 0.1 and FDRs of 0.05, 0.1, and 0.25 (by the symmetry of the equation). The obtained true positive rate for a given FDR and prevalence is the intersection point between the straight line and the ROC curve.

Supplementary Tables

Table S1

DAVID output for the 61 proteins detected by RAIN in both proteomics studies (Mauvoisin et al., 2014; Robles et al., 2014). This file can be opened directly in most common spreadsheet programs, or imported to data analysis software directly as plain text. DAVID detects overrepresented clusters of related biological functions, the specific functions are to be found in the second column.