Arenas *et al. BMC Bioinformatics* (2015) 16:152
DOI 10.1186/s12859-015-0599-8

BMC Bioinformatics

RESEARCH ARTICLE Open Access

MSCA: a spectral comparison algorithm between time series to identify protein-protein interactions

Ailan F Arenas[1*†], Gladys E Salcedo[2†], Andrey M Montoya[2] and Jorge E Gomez-Marin[1]

Abstract

Background: The interactions between pathogen proteins and their hosts allow pathogens to manipulate host cellular mechanisms to their advantage. The identification of host proteins that are targeted by virulent pathogen proteins is crucial to increase our understanding of infection mechanisms and to propose new therapeutics that target pathogens. Understanding the virulence mechanisms of pathogens requires a detailed molecular description of the proteins involved, but acquiring this knowledge is time consuming and prohibitively expensive. Therefore, we develop a statistical method based on hypothesis testing to compare the time series obtained from conversion of the physicochemical characteristics of the amino acids that form the primary structure of proteins and thus to propose potential functional relation between proteins. We called this algorithm the multiple spectral comparison algorithm (MSCA); the MSCA was inspired by the BLASTP tool and was implemented in R code. The algorithm compares and relates multiple time series according to their spectral similarities, and the biological relation between them could be interpreted as either a similar function or protein-protein interaction (PPI).

Results: A simulation study showed that the MSCA works satisfactorily well when we compare unequal time series generated from ARMA processes because its power was close to 1. The MSCA presented a 70% average accuracy of detecting protein interactions using a threshold of 0.7 for our spectral measure, indicating that this algorithm could predict novel PPIs and pathogen-host interactions (PHIs) with acceptable confidence. The MSCA also was validated by its identification of well-known interactions of the human proteins MAGI1, SCRIB and JAK1, as well as interactions of the virulence proteins ROP16, ROP18, ROP17 and ROP5. We verified the spectral similarities for human intraspecific PPIs and PHIs that were previously demonstrated experimentally by other authors. We suggest that human GBP (GTPase group induced by interferon) and the CREB transcription factor family could be human substrates for the complex of ROP18, ROP17 and ROP5.

Conclusions: Using multiple-hypothesis testing between the spectral densities of a set of unequal time series, we developed an algorithm that is able to identify the similarities or interactions between a set of proteins.

Keywords: Hypothesis testing, Protein-protein interactions, Time series, Toxoplasma

Background

The identification of protein-protein interactions (PPIs) is crucial for elucidating protein function and further understanding various biological processes in cells. Similarly, the identification of interactions between the proteins of infectious pathogens and their hosts (PHIs) may enable researchers to gain crucial insight into infection mechanisms. However, the general methodology for searching for new PPIs and PHIs, such as large-scale yeast two-hybrid approaches or coimmunoprecipitation methods, is time consuming and expensive [1]. Therefore, the design of computational tools, which can provide an efficient method of identifying potential protein interacting partners, is beneficial for minimizing the number of experiments. Current computational methods can be classified into two main approaches. The first approach is based

*Correspondence: aylanfarid@yahoo.com
†Equal contributors
[1]Gepamol, Universidad del Quindío, Carrera 15 Calle 12N, Armenia, Colombia
Full list of author information is available at the end of the article

on the genomic [2] or structural information of proteins [3,4]. However, these methods cannot be implemented when prior information about the proteins is not available. The second approach is based on protein primary sequences [5-7]. The latter type of approach is beneficial because most of the protein information in protein databases is the protein primary structure. The pattern of amino acid positions in protein primary structures give rise to an assumption that the amino acid sequence alone might be sufficient to estimate the propensity for interactions between two proteins for specific biological functions [8]. Accordingly, predicting PPIs and PHIs based only on sequence information is an ideal approach for computational techniques. Most of the methods for PPI prediction based on the primary structure information of proteins have been developed using a learning algorithm-support vector machine (SVM) combined with a kernel function to perform training and extract features from known pairs of PPIs to construct a universal model that separates positive PPIs from false PPIs [6,9-12]. Herein, we developed an algorithm to compare the spectral similarities between proteins using statistical hypothesis testing. In our case, we do not construct a general model for a protein-protein relation; instead, we compare the different spectral functions obtained using different descriptors for a query protein sequence against our own constructed database. This approach has the advantage of using only the protein primary sequence and not requiring either previous information from datasets or training.

The primary structure of proteins is a linear chain of amino acids that are each represented by one of 20 letters of the alphabet; thus, this alphabetic sequence can be translated into a numerical sequence using different physicochemical properties for each amino acid, such as the electron ion interaction potential (EIIP), hydrophobicity, polarity, polarizability, Van der Walls volume, ionization constant, and accessible solvent surface area.

However, because the distance between consecutive CA atoms in a protein sequence is $3.8A^{\circ}$, the points in some corresponding numerical sequences are considered equidistant, and the corresponding numerical sequence can be considered a time series. When two proteins are compared using some bioinformatics techniques, such as pairwise sequence alignments, the similarity between the proteins can be observed simply by looking for the amino acids that are conserved in specific positions of the two proteins. Otherwise, if we transform the same proteins into two time series, the proteins can be compared using mathematical techniques that allow us to see some hidden patterns that cannot be observed through the conventional alignment methods or motif searching patterns. However, when proteins are represented as a time series, we can use methods for extracting information from signals via spectral analysis. For instance, if there are certain periodicities or repetition patterns in two signals, prominent peaks appear in their spectra, and each one of these peaks carries relevant information that can represent either a functional or interaction relation [13-15]. This type of analysis is called an information spectrum method (ISM), and it has been successfully used to characterize protein-protein interactions between the gp120 HIV protein and its CD4, CCR5 and CXCR host cell receptors [16]. To further develop the ISM technique in this work, we propose the multiple spectral comparison algorithm (MSCA) to identify similarities between a query and our own set of proteins. The algorithm was inspired by the BLASTP tool in the sense that for each pair of proteins, we compare the spectral densities of all of their alignments rather than the amino acids themselves.

Results and discussion

Results

First, we determine whether MSCA can identify some interactions of the human proteins MAGI1 and SCRIB using eight different physical/chemical amino acid descriptors and four different sets of human proteins with MAGI1 and SCRIB as the query proteins. The algorithm identified the following protein relations: MAGI1-NET1 [17-19], MAGI1-FZD4 [19,20], MAGI1-ESAM [19,21], MAGI1-ABC1, MAGI1-CYSLTR2, MAGI1-ARHGAP6, MAGI1-TMEM215 and MAGI1-MARCH3 [19] with a similarity measure greater than 0.7. The MAGI1-NET1 and MAGI1-CYSLTR2 interactions were found using three different descriptors (Additional file 1: Supplementary material S1). MSCA also detected the protein-protein interactions SCRIB-βPIX [22], SCRIB-GLUT7, SCRIB-TANC1, SCRIB-MARCH3, SCRIB-ABC1, SCRIB-ARHGAP6, SCRIB-TAX, SCRIB-CYSLTR2 and SCRIB-TMEM215 [19]. The interactions SCRIB-MARCH3, SCRIB-TMEM215 and SCRIB-βPIX were also detected using three descriptors (Additional file 1: Supplementary material S2). The position and frequency of the interaction partners of MAGI1 and SCRIB did not change dramatically when we used the 4 different datasets. For this analysis, the interactions MAGI1-NET1, MAGI1-ARHGAP6, SCRIB-ARHGAP6, SCRIB-TMEM215 and SCRIB-βPIX appear to be the most consistent findings; these interactions exhibit the same frequency in the 4 datasets (Additional file 1: Supplementary material S1 and S2). We also evaluated JAK1 interaction partners using JAK1 as the query protein. In this case, the majority of proteins related to JAK1 were kinases, including TYK2 and SYK, which have been demonstrated to interact with JAK1 [23,24]. The most common transcription factor family was the STAT family, which are well-proven substrates for JAK1; STAT1 and STAT5b were the most frequent JAK1 partners in the 4 analyzed datasets [25] (Additional file 1: Supplementary

material S3). The MSCA also found the previously known interaction JAK1-TRAF6 [26]. All of the candidates had a similarity measure greater than 0.7 for all descriptors.

Next, we assessed the pathogen-host interactions (PHIs) for some well-studied ROPKs and host proteins. The interactions between ROP16 and STATs were used to validate the MSCA, and similar to JAK1, most of the proteins related to ROP16 are kinases. ROP16 is a kinase protein that has all of the key amino acids for the phosphotransferase function [27] (Additional file 1: Supplementary material S4). Similarly, the third most frequent group that was identified to be related to ROP16 was the STAT transcription factor family; STAT5A and STAT5B were the most frequent partners, with each occurring five times (Additional file 1: Supplementary material S4). The MSCA also detected the experimentally demonstrated interactions ROP16-STAT3, ROP16-STAT6 and ROP16-STAT1 (Additional file 1: Supplementary material S4). The ROP18 have also been described that phosphorilates a member of the mouse GTPase family IRGa6 [28-30], and we also aimed to evaluate this interaction using the MSCA. The kinases were largely represented, but members of the immunity-related GTPase family (IRGs) were found 21 times (Additional file 1: Supplementary material S5). The protein ROP5 was also shown to act as a cofactor of ROP18. A recent work concluded that ROP17, ROP18 and ROP5 function as a complex and that the host substrates for ROP17 and ROP18 are members of the mouse immunity-related GTPase family [28-31]. The MSCA results for this complex showed that aside from proteins with kinase activity, the second most frequent family of proteins is the mouse IRG family, which was significantly related to ROP17 and ROP5, with 20 and 17 instances, respectively (Additional file 1: Supplementary materials S5, S6 and S7).

Finally, we obtained an average accuracy of 70% for detecting protein interactions using a threshold of 0.7 for our p-value measure. However, the specificity of the test was improved when we increased the threshold to 0.8 (Additional file 1: Supplementary material S8). In general, we considered 70% accuracy acceptable for finding novel PPI or PHIs. Furthermore, we assumed that if the functional protein families found and described below from the sequence query with a probability higher than 0.7, the families would have some relationship with the query sequence; these proteins share spectral similarity derived from physicochemical features, and thus, this information could be interpreted as either a common function or PPI.

Discussion

Discovering protein interaction partners is a difficult task because it is time consuming and experimentally expensive; thus, it is necessary to generate algorithms to develop computational tools that help researchers who are deciding how to better understand the pathogen-human interaction system and decrease the number of experiments that must be performed. Previous experimental information, curated databases or 3D structural information is necessary to find potential interactions between proteins. Most of the bioinformatic programs that predict PPIs and PHIs require already characterized or experimentally proven PPIs and PHIs to transfer this information to new sequences. Therefore, our motivation was to develop a program that we called the MSCA, which will identify PPI or PHI relations between proteins from the primary sequence information itself. Each spectrum contains the information for each particular physicochemical descriptor for all of the proteins in this study. The MSCA relies on a spectral comparison of the protein sequences, but the comparison was formally designed through hypothesis testing. The MSCA confirmed all known PPIs using a similarity threshold of greater than 0.7. If a query protein has significant spectral similarities with another protein (using several descriptors) but the proteins are functionally different (in our case, we compared toxoplasma ROP kinases vs. transcription factors), this would mean that some spectral information is shared and would suggest an interaction between the proteins. When comparing series, commonality of some frequency and amplitude peaks along the spectra suggests a relation between the series. In the case of biological sequences, commonality of particular frequency peaks that arise from the periodical interaction interfaces of the proteins would suggest an interaction relationship. However, the MSCA can also identify the functional similarity (as shown in all tables). Indeed, many human kinase proteins appeared close to the ROP queries because they are also kinases. MSCA detected the human PPI between MAGI1 and SCRIB. Accordingly, in our analysis, the domains related to the G protein Rho are the third most abundant for MAGI1 and SCRIB and appeared 11 and 13 times, respectively (Additional file 1: Supplementary materials S1 and S2). The second most abundant group is the proteins-related to cell-cell adhesion and integral membrane proteins. Additional experimental studies suggested that the MAGI1 and SCRIB proteins are closely associated with cell-cell adhesion and that these proteins act as scaffolds that assemble proteins close to membranes to regulate G protein Rho signaling [19,20,32,33]. Similarly, the ribosomal protein S6 kinase RPS6K and MAPK3 can interact with MAGI1 and SCRIB, respectively [34,35]. For the JAK1 validation, the transcription factor family STAT is the third most frequent family, and 10 experimental JAK-STAT interactions that had already been experimentally proven were found (Additional file 1: Supplementary

material S3). Moreover, toxoplasma ROP16, ROP17 and ROP18 are grouped as active kinases, and these proteins are not highly divergent from one another [27]; however, other protein groups are also related to each ROP. Our ROP16 analysis showed that in addition to kinases, the STAT transcription factor family was represented frequently and appeared a total of 18 times (Additional file 1: Supplementary material S4). Although the experimentally proven interactions are not the most frequent, the MSCA found the ROP16-STAT3 and ROP16-STAT1 interactions one time each and found the ROP16-STAT6 twice. Following this concept, when we analyzed ROP18 and ROP17, the group with the second most frequent occurrences was the immunity-related GTPases (IRGs) (Additional file 1: Supplementary materials S5 and S6). Furthermore, the CREB human transcription factor family was identified frequently during ROP18 and ROP5 queries with 15 and 16 occurrences, respectively; CREB1 was the most frequently found member of the CREB family (Additional file 1: Supplementary materials S5 and S7). Experimental evidence also demonstrated that ROP18 interacts with the ATF6β factor, which belongs to the CREB family [36]. Another human protein group related to the ROP18, ROP5, and ROP17 complex is the SMAD family, which is a group of signal transducers and transcriptional modulators that belong to the (TGF-β) pathway and mediate cell differentiation [37]. Finally, an interesting group that is also related to the complex ROP18/ROP17/ROP5 is the human GBP GTPase family, which consists of guanylate-binding proteins induced by interferon. These types of proteins promote inflammasome responses to pathogenic bacteria [38,39]. We consider GBPs to be possible human substrates for the ROP complex because the ROP complex is able to interact with mouse IRGs, which are also GTPases that are induced by interferon. Furthermore, human GBPs are highly induced after microbial infection and are associated with T. gondii [40]. Although the MSCA relies on spectral information methods, it compares the complete spectrum rather than only the frequency peaks that are shared among the proteins. Moreover, formal statistical testing was used. In summary, the MSCA results included highly well-known and experimentally identified PPIs as well as some new candidates that have a sound theoretical basis for an interaction. These candidates merit further experimental validation. In agreement with the accumulating evidence, the MSCA identifies some direct candidates for PPIs, PHIs and protein-function relations. At minimum, the MSCA can reduce the number of sequences in a large database that should be further studied, and the sequences that remain should be true candidates for relationships with the query protein. The MSCA provides the advantage of analyzing a large number of sequences, and the method can be generalized for any type of protein from any organism.

Conclusions

Using multiple-hypothesis testing between the spectral densities of several time series, we developed an algorithm that can identify similarities or interactions between a set of proteins. A simulation study that compared different series generated from autoregressive moving average processes showed that the approach works satisfactorily. We also could identify some well-known interactions between proteins from a toxoplasma-host infection model. Considering the obtained accuracy, we choose a threshold of 0.70 that guarantees an interaction with the query protein.

Methods

Time series analysis

A time series is a set of observations $\{x_t\}$, where each x_t is recorded at a different time. If the observations are recorded at discrete points, we have a discrete time series; this type of time series is used most frequently in practice. A more formal definition of a time series can be obtained using the theory of stochastic processes. In this context, the time series $\{x_t,\ t = 1, 2, \ldots, T\}$ represents a realization of a stochastic process $\{X_t,\ t \in \tau\}$. Stationarity is an interesting property of stochastic processes and can be either strong or weak. Processes are strongly stationaries if their finite-dimensional distributions are time invariants, and processes are second-order stationaries when the unconditional expectations and variances are time invariants and if the correlation structures between observations x_t and x_s depend solely on the delay $k = |s - t|$.

There are two common approaches for analyzing a stationary time series depending on the domain under consideration. In the *time domain*, the analyses are largely based on the autocorrelation function (acf) given by

$$\rho_x(k) = \frac{\gamma_x(k)}{\sigma^2}, \quad k = 0, \pm 1, \pm 2, \ldots,$$

where $\gamma_x(k) = Cov(x_t, x_{t-k}) = E[(x_t - \mu)(x_{t-k} - \mu)]$ is the autocovariance function, μ is the unconditional expectation and σ^2 is the unconditional variance of the process. The function $\rho_x(k)$ measures the linear dependence between pairs of observations separated by a lag k. If $\rho_x(k) = 0$ for all $k \neq 0$, the process lacks memory. In the *frequency domain*, the correlation structure is represented by the spectral density function defined as

$$f_x(\lambda) = \sum_{k=-\infty}^{\infty} \gamma_x(k) \exp(-i2\pi\lambda k), \quad \lambda \in [-1/2, 1/2],$$

where λ is measured in cycles per unit of time. $f_x(\lambda)$ is the Fourier transform of $\gamma_x(k)$ and describes the properties of the process in terms of periodic components at different frequencies.

For a set of observations $\{x_t,\ t = 1, 2, \ldots, T\}$, the *discrete Fourier transform (DFT)* defined for the discrete Fourier frequencies $\lambda_j = j/T,\ j = 0, 1, 2, \ldots, [T/2]$ is given by

$$d_x(\lambda_j) = \frac{1}{T}\sum_{t=1}^{T} x_t \exp(-i2\pi\lambda_j t).$$

An estimator of the spectral density $f_x(\lambda)$ is the periodogram $I_x(\lambda_j)$, which is defined as the squared modulus of the DFT,

$$I_x(\lambda_j) = |d_x(\lambda_j)|^2.$$

The value of the periodogram at each frequency represents the amount of time series variance related to this frequency or its power.

Time series metrics

The classification and comparison of time series are problems with applications in biology, medicine, seismology, economics, and other fields, and different metrics have been proposed for classifying a set of time series. Through a simulation study, Caido et al. [41] evaluated several metrics to compare two stationary time series; most of these metrics were based on Euclidian distances. The metrics studied in the time domain were the Euclidian distance between the two time series, the two autocorrelation functions, the two partial autocorrelation functions and the Euclidian distance between their respective autoregressive parameters. In the frequency domain, the analyzed metrics were the Euclidean distance between the normalized periodograms and the Kullback-Leibler distance. A simulation study showed that distances based on the autoregressive parameters and normalized periodograms were the best metrics. Maharaj [42] proposed the p-value of the hypothesis testing of the equality of autoregressive parameters as a metric to compare two stationary time series. Accordingly, our algorithm uses the p-value of the hypothesis testing of the equality of spectral densities.

The hypothesis testing

The issue of comparing two or more stationary time series is equivalent to evaluating whether the series were generated by the same stationary process. Stationary time series are similar if they have the same correlation structure. Coates and Diggle [43] provided some conditions for comparing two time series in the frequency domain.

Let $f_x(\lambda)$ and $f_y(\lambda)$ be the spectral density of $\{x_t,\ t = 1,\ldots,T\}$ and $\{y_t,\ t = 1,\ldots,T\}$, respectively, and let $I_x(\lambda_k)$ and $I_y(\lambda_k)$ be their respective periodograms. For $k = 1,\ldots,K,\ k << T, \frac{k}{T} \approx \lambda$ and $\frac{k}{T} \neq 0, \pm\frac{1}{2},\ldots$, when $T \to \infty$, and when the time series are independent,

$$J(\lambda_k) = \frac{I_x(\lambda_k)}{I_y(\lambda_k)} \xrightarrow{d} U(\lambda)F_{2,2},$$

where $\xrightarrow{d}$ represents a distribution convergence, $U(\lambda) = \frac{f_x(\lambda)}{f_y(\lambda)}$ and $F_{2,2}$ is the Fisher distribution with two degrees of freedom in both its numerator and denominator. Furthermore, $z_k = \ln(1+J^{-1}(\lambda_k)) \xrightarrow{d} U(\lambda_k)\exp(1)$. Thus, when the spectral densities are equal, $U(\lambda_k) = 1$, and when they are asymptotically equivalent, z_k is exponentially distributed with a mean of 1. Consequently, the statistics $c_j = \sum_{k=1}^{j} z_k$ describe the points of a Poisson process of mean 1, and

$$\left\langle o_j = \frac{c_j}{c_m} \right\rangle,\ j = 1,\ldots,m = [T/2],$$

is a vector of the order statistics from a uniform distribution over $(0,1)$. Then, we can test the hypothesis H_0 : $f_x(\lambda) = f_y(\lambda),\ \forall\lambda \in (-1/2, 1/2)$ if the statistics o_j's follow a uniform distribution over $(0,1)$; for instance, we can use a Kolmogorov-Smirnov test.

The multiple spectral comparison algorithm (MSCA)

Our approach for comparing a query and a set of proteins with unequal sizes and for sorting the proteins according to their similarities follows the steps outlined below:

- Step 1: The set of proteins is transformed into a set of time series according to some amino acid properties (see properties in the Additional file 1: Supplementary material).
- Step 2: Prior to each alignment between the query and a protein, hypothesis testing for equality of their spectral densities is performed and provides a p-value of the testing. Each alignment is understood as a match between the query and each protein using translations of order 1. The spectral similarity is represented by the mean of these p-values.
- Step 3: Because the p-value from an equality testing of two time series represents a similarity measure between the two time series and satisfies the properties of a semi-metric [42], the set of proteins is sorted according to the p-values obtained in Step 2. The similarity between a protein and itself results in a p-value of 1, and this protein has the highest score. Similarities close to 1 indicate that two proteins are strongly related.

Simulation study

Our algorithm is based on a multiple-hypothesis testing of signals of different lengths; thus, to assess its power for finite samples, we performed some simulations where we compared series generated from autoregressive moving average (ARMA) processes. $\{X_t, t \in \tau\}$ is a stationary ARMA(p,q) process of zero mean if $X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \theta_1 a_{t-1} + \ldots + \theta_q a_{t-q} + a_t$, where the roots of the characteristic polynomials $1 - \phi_1 B - \ldots - \phi_p B^p$ and $1 - \theta_1 B - \ldots - \theta_q B^q$ are outside the unit circle and $\{a_t\}$ is a white noise process.

Then, in the first case, we only compared time series AR(1) or MA(1) and generated series of length $T = 1000$ from AR(1) processes with ϕ_1 varying from $\{0.2, 0.3, \ldots, 0.9\}$. These series were compared with series AR(1), where $\phi_1 = 0.2$ remains fixed and $T = 800$. Analogously, we generated series from processes MA(1) with θ_1 varying from $\{0.2, 0.3, \ldots, 0.9\}$ and length $T = 1000$, and we compared these series with series MA(1), where $\theta_1 = 0.2$ remains fixed and $T = 800$. In both cases, we simulated 2000 replications and considered a nominal size of 5%. We calculate this size when both signals are generated from the same process where the parameter is 0.2; in the other cases, we calculate the power. Figure 1(a) and 1(b) show the estimated power function when we compare the signals from the AR and MA processes, respectively. In general, the test performance is similar when we compare either the AR or MA time series, and the test is reasonably good because the estimated power function increases rapidly to 1 when the processes are different. In both cases, the estimated size is 0.045.

In the second case, the algorithm was used for 2000 replications of the following group of stationary time series: two time series from AR(1) process with $\phi_1 = 0.8$ and $T = 500$, two time series from MA(1) process with $\theta_1 = -0.6$ and $T = 400$, and two time series from ARMA(1,1) process with the parameters $\phi_1 = 0.8$, $\theta_1 = -0.6$ and $T = 300$. In this case, we compare ARMA(1,1) with the other series. However, due to the similarity between the parameters, the series from the AR(1) or MA(1) processes could eventually be classified as similar to series from the ARMA(1,1) process, and we obtained less power due to the misclassification. The estimated size was 0.054, and the power was 0.9995 because only 1 series was misclassified.

Finally, we augmented the previous group with two time series from the AR(2) process with the parameters $\phi_1 = 0.8$, $\phi_2 = -0.3$ and $T = 200$ and with two time series from the MA(2) process with the parameters $\theta_1 = -0.6$, $\theta_2 = 0.3$ and $T = 100$. In this case, the estimated size and power were 0.048 and 0.883, respectively. The loss of power was due to 1 or 2 misclassifications with a frequency of 0.112 or 0.005, respectively, in the 2000 replications. However, for a nominal size of 10%, the estimators were 0.106 and 0.95 for size and power, respectively, with only 1 misclassified series with a frequency of 5%.

In general, the MSCA performs reasonably well because in all cases, the estimated size was close to the nominal value, and for different series, the power rapidly increased to 1, as can be expected.

Validation study

For intra-specific (PPI) validation, we compared the different spectra for three well-studied human proteins, MAGI1, SCRIB and JAK1, against our own spectral

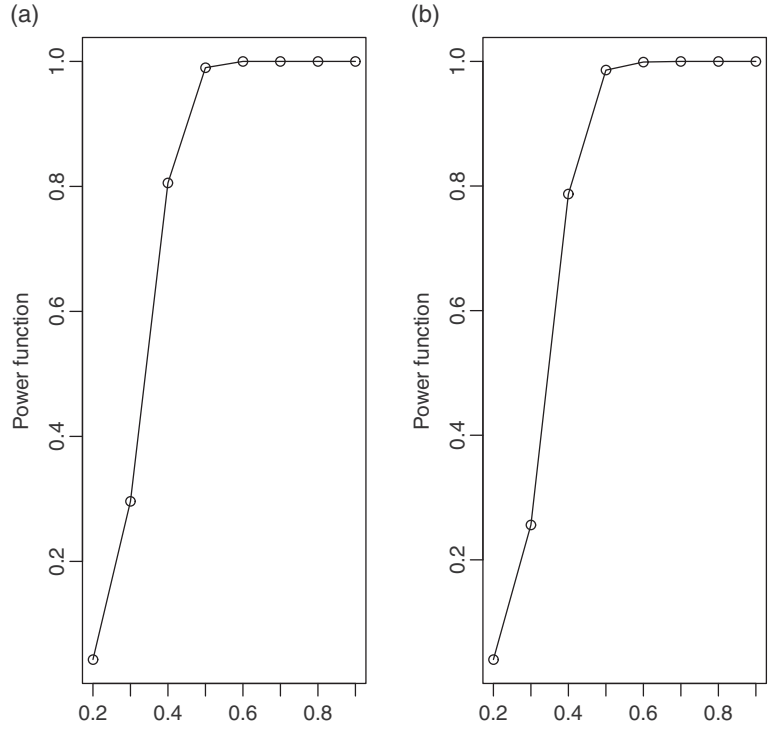

Figure 1 Power function for comparison of AR(1) processes **(a)** and MA(1) processes **(b)**. This figure shows the estimated power function from the hypothesis testing of Step 1.

dataset. MAGI1 (membrane-associated guanylate kinase, contains a PDZ domain), which has six PDZ domains, was found to be located to adherens and tight junctions in epithelial and endothelial cells [21,44], where MAGI1 appears to be involved in the maintenance of the junctions and in cell signal propagation. SCRIB (scribbled planar cell polarity protein), which has four PDZ domains, is known to be involved in the establishment of adherens and tight junctions as well as in the regulation of cell polarity and cell migration [45-47]. JAK1 (Janus kinase 1) is involved in the interferonα/β and interferonγ signal transduction pathways. Furthermore, for (PHIs), we used the pathogen-host infection model (Toxoplasma gondii-Host). This pathogen is an obligate intracellular parasite that is able to infect any mammalian cell [48]. T. gondii is a highly successful parasite that can manipulate and control a variety of host processes due to secreted factors that interact with the host cell proteins [49-51]. Consequently, rhoptry proteins are vital for the Toxoplasma infection process and for its survival. There are a few well-documented host target proteins for toxoplasma rhoptry kinases (ROPKs) that are involved in host cell modulation. A proteomic study of rhoptry contents led to the identification of 38 rhoptry proteins [52], after a screening of a database (ToxoDB) for ROPKs revealed 44 ROPKs in the T. gondii genome using hidden Markov models (HMMs) and a phylogenomic approach [51]. ROP16 activates the STAT family transcription factors STAT1, STAT3 and STAT6 that influence the JAK/STAT pathway [53-55]. In a recent study, the authors found that ROP18 forms a complex with ROP5 and ROP17, which phosphorylate mouse immunity-related GTPase family members (IRGs) [28].

Datasets

The MSCA was validated by searching for protein interaction partners that had been experimentally proven for MAGI1, SCRIB and JAK1 (PPIs) and for the aforementioned interactions between ROP16, ROP18, ROP17, ROP5 and host STAT and IRG proteins (PHIs). We downloaded 930 proteins from the UniProt database (http://www.uniprot.org) and 250 kinases belonging to seven host signal transduction pathways from the KEGG database (www.genome.jp/kegg/pathway), the MAPK, JAK-STAT, NF-, TNF, HIF-1, PI3K-Akt and mTOR pathways, as well 450 transcription factors, 100 membrane proteins related to cell-cell adhesion and 130 proteins related to different activities. We mixed all of these proteins into four different sets and tested some query proteins against each group. Group 1 had 279 sequences, and the other three groups had 218 different sequences each; the queries were MAGI1 and SCRIB. For the JAK1 query, group 1 had 262 sequences, and the other three groups had 212 sequences; for ROP16, ROP18, ROP17 and ROP5, only one group of 332 proteins was considered for each validation (see Additional file 1: Supplementary material). MSCA sorted the proteins of each group according to the spectral similarity measure (the global p-value) of each protein with each of the following query proteins: MAGI1, SCRIB, JAK1, ROP16, ROP18, ROP17 and ROP5. (All the sequences used in the validation study were uploaded in the Additional file 2).

Feature conversions

PPIs can be categorized into four interaction modes: electrostatic interactions, hydrophobic interactions, steric interactions and hydrogen bonds. Here, 6 physicochemical features of amino acids were selected to transform the alphabetic sequences into a numerical series to reflect these interaction modes. These features were hydrophobicity (hydro), volume of side chains (VSC), polarity (P1), polarizability (P2), solvent accessible surface area (SASA) and the net charge index of side chains (NCISC) [56]. Furthermore, we considered 5 other physicomathematical characteristics for each amino acid, and these characteristics were successful used in the ISM technique (to look for interaction partners). These characteristics were the electron ion interaction potential (EIIP) and ionization constant (IC), which were used in [57,58] and [59], respectively. The characteristics P001, H085 and H371 were also previously proposed [60].

Measuring the accuracy

We calculated the accuracy (ACC) and F1 scores to assess the accuracy of the MSCA. We downloaded the interaction partners for JAK1, MAGI1 and STAT3 from STRING 9.1 (string-db.org) [61]. Each protein was analyzed separately, and we designed 3 sets of negative interactions for each analysis.

Additional files

Additional file 1: Supplementary material. The supplementary material contains all tables of the validation study.

Additional file 2: Supplementary material. The supplementary material contains all the sequences used in the validation study.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AFA and GES designed the approach and implemented and validated the algorithm; AMM helped with the simulation study; JGM contributed to the discussion and revision of the manuscript; and all authors read and approved of the final manuscript.

Acknowledgements

This work was supported by COLCIENCIAS through grant 111356933319 and by the University of Quindio. The authors thank Luis Hernando Hurtado for his advice on this work.

Author details
[1]Gepamol, Universidad del Quindío, Carrera 15 Calle 12N, Armenia, Colombia. [2]Grupo de Investigación y Asesoría en Estadística, Carrera 15 Calle 12N, 460 Armenia, Colombia.

Received: 5 November 2014 Accepted: 13 April 2015
Published online: 13 May 2015

References

1. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. 2001;409:533–8.
2. Pazos JD, Valencia F. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci USA. 2008;105:934–9.
3. Singhal M, Resat H. A domain-based approach to predict protein-protein interactions. BMC Bioinformatics. 2007;8:199.
4. Burger L, Van NE. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol. 2008;4:165.
5. Chou KC, Cai YD. Predicting protein-protein interactions from sequences in a hybridization space. J Proteome Res. 2006;5:316–22.
6. Shen JW, Zhang J, Luo XM, Zhu WL, Yu KQ, Chen K, et al. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci USA. 2007;104:4337–41.
7. Guo YZ, Yu LZ, Wen ZN, Li ML. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. 2008;3:3025–30.
8. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181:223–30.
9. Ofran Y, Rost B. Predicted protein-protein interaction sites from local sequence information. FEBS Lett. 2003;544:236–39.
10. Res I, Mihalek I, Lichtarge O. An evolution based classifier for prediction of protein interfaces without using protein structures. Bioinformatics. 2005;21:2496–501.
11. Betel D, Breitkreuz KE, Isserlin R, Dewar DD, Tyers M, Hogue CW. Structure-templated predictions of novel protein interactions from sequence information. PLoS Comput Biol. 2007;3:1783–89.
12. Yu CY, Chou LC, Chang DT. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. BMC Bioinformatics. 2010;11:167.
13. Veljkovic V, Niman HL, Glisic S, Veljkovic N, Perovic V, Muller CP. Identification of hemagglutinin structural domain and polymorphisms which may modulate swine H1N1 interactions with human receptor. BMC Struct Biol. 2009;9:62.
14. Veljkovic V, Veljkovic N, Muller CP, Muller S, Glisic S, Perovic V, et al. Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control. BMC Struct Biol. 2009;9:21.
15. Hu W. Highly conserved domains in hemagglutinin of influenza viruses characterizing dual receptor binding. Nat Sci. 2010;2:1005–14.
16. Krsmanovic V, Biquard JM, Sikorska WM, Cosic I, Desgranges C, Trabaud MA, et al. Investigation into the cross-reactivity of rabbit antibodies raised against nonhomologous pairs of synthetic peptides derived from HIV-1 gp120 proteins. J Peptide Res. 1998;52:410–20.
17. Dobrosotskaya IY. Identification of mNET1 as a candidate ligand for the first PDZ domain of MAGI-1. Biochem Biophys Res Commun. 2001;283:969–75.
18. Fournane S, Charbonnier S, Chapelle A, Kieffer B, Orfanoudakis G, Trave G, et al. Surface plasmon resonance analysis of the binding of high-risk mucosal HPV E6 oncoproteins to the PDZ1 domain of the tight junction protein MAGI-1. J Mol Recognit. 2010;24:511–23.
19. Luck K, Fournane S, Kieffer B, Masson M, Nomine Y, Trave G. Putting into Practice Domain-Linear Motif Interaction Predictions for Exploration of Protein Networks. PLoS ONE. 2011;e25376:6.
20. Yao R, Natsume Y, Noda T. MAGI-3 is involved in the regulation of the JNK signaling pathway as a scaffold protein for frizzled and Ltap. Oncogene. 2004;23:6023–30.
21. Wegmann F, Ebnet K, Pasquier LD, Vestweber D, Butz S. Endothelial adhesion molecule ESAM binds directly to the multidomain adaptor MAGI-1 and recruits it to cell contacts. Exp Cell Res. 2004;300:121–33.
22. Audebert S, Navarro C, Nourry C, Chasserot GS, Lécine P, Bellaiche Y, et al. Mammalian Scribble forms a tight complex with the βPIX exchange factor. Curr Biol. 2004;14:987–95.
23. Domanski P, Yan H, Witte MM, Krolewski J, Colamonici OR. Homodimerization and intermolecular tyrosine phosphorylation of the Tyk-2 tyrosine kinase. FEBS Lett. 1995;3:317–22.
24. Zhou YJ, Magnuson KS, Cheng TP, Gadina M, Frucht DM, Galon J, et al. Hierarchy of protein tyrosine kinases in interleukin-2 (IL-2) signaling: activation of syk depends on Jak3; however, neither Syk nor Lck is required for IL-2-mediated STAT activation. Mol Cell Biol. 2000;12:4371–780.
25. Fujitani Y, Hibi M, Fukada T, Takahashi TM, Yoshida H, Yamaguchi T, et al. An alternative pathway for STAT activation that is mediated by the direct interaction between JAK and STAT. Oncogene. 1997;7:751–61.
26. Motegi H, Shimo Y, Akiyama T, Inoue J. TRAF6 negatively regulates the Jak1-Erk pathway in interleukin-2 signaling. Genes Cells. 2011;2:179–89.
27. Talevich E, Kannan N. Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors. BMC Evol Biol. 2013;13:117.
28. Etheridge RD, Alaganan A, Tang K, Lou HJ, Turk BE, Sibley LD. The Toxoplasma pseudokinase ROP5 forms complexes with ROP18 and ROP17 kinases that synergize to control acute virulence in mice. Cell Host Microbe. 2014;15:537–50.
29. Fleckenstein MC, Reese ML, Könen WS, Boothroyd JC, Howard JC, Steinfeldt T. A Toxoplasma gondii Pseudokinase Inhibits Host IRG Resistance Proteins. PLoS Biol. 2012;10:e1001358.
30. Niedelman W, Gold DA, Rosowski EE, Sprokholt JK, Lim D, Farid Arenas A, et al. The rhoptry proteins ROP18 and mediate Toxoplasma gondii evasion of the murine, but not the human, interferonγ response. PLoS Pathog. 2012;8:e1002784.
31. Reese ML, Shah N, Boothroyd JC. The Toxoplasma Pseudokinase ROP5 Is an Allosteric Inhibitor of the Immunity-related GTPases. J Biol Chem. 2014;40:27849–58.
32. Mino A, Ohtsuka T, Inoue E, Takai Y. Membrane-associated guanylate kinase with inverted orientation (MAGI)-1/brain angiogenesis inhibitor 1-associated protein (BAP1) as a scaffolding molecule for Rap small G protein GDP/GTP exchange protein at tight junctions. Genes Cells. 2000;5:1009–16.
33. Zhang H, Wang D, Sun H, Hall RA, Yun CC. MAGI-3 regulates LPA induced activation of Erk and RhoA. Cell Signal. 2007;19:261–68.
34. Thomas GM, Rumbaugh GR, Harrar DB, Huganir RL. Ribosomal S6 kinase 2 interacts with and phosphorylates PDZ domain-containing proteins and regulates AMPA receptor transmission. Proc Natl Acad Sci USA. 2005;42:15006–11.
35. Elsum IA, Martin C, Humbert PO. Scribble regulates an EMT polarity pathway through modulation of MAPK-ERK signaling to mediate junction formation. J Cell Sci. 2013;126:3990–9.
36. Yamamoto M, Ma JS, Mueller C, Kamiyama N, Saiga H, Kubo E, et al. ATF6 is a host cellular target of the Toxoplasma gondii virulence factor ROP18. J Exp Med. 2011;208:1533–46.
37. Ji H, Tang H, Lin H, Mao J, Gao L, Liu J, et al. Rho/Rock cross-talks with transforming growth factor-β/Smad pathway participates in lung fibroblast-myofibroblast differentiation. Biomed Rep. 2014;6:787–92.
38. Shenoy AR, Wellington DA, Kumar P, Kassa H, Booth CJ, Cresswell P, et al. GBP5 promotes NLRP3 inflammasome assembly and immunity in mammals. Science. 2012;336(6080):481–85.
39. Ohshima J, Lee Y, Sasai M, Saitoh T, Su MJ, Kamiyama N, et al. Role of mouse and human autophagy proteins in IFNγ induced cell-autonomous responses against Toxoplasma gondii. J Immunol. 2014;7:3328–35.
40. Degrandi D, Konermann C, Beuter GC, Kresse A, Würthner J, Kurig, S, et al. Extensive characterization of IFNγ induced GTPases mGBP1 to mGBP10 involved in host defense. J Immunol. 2007;11:7729–40.
41. Caiado J, Crato N, Peña D. A periodogram-based metric for time seires classification. Comput Stat Data Anal. 2006;50:2668–84.
42. Maharaj EA. Comparison and classification of stationary multivariate time series. Pattern Recognition. 1999;32:1129–38.
43. Coates DS, Diggle PJ. Test for comparing two estimated spectral densities. J Time Ser Anal. 1986;7:7–20.
44. Ide N, Hata Y, Nishioka H, Hirao K, Yao I, Deguchi M, et al. Localization of membrane-associated guanylate kinase (MAGI)-1/BAI-associated protein (BAP) 1 at tight junctions of epithelial cells. Oncogene. 1999;18:7810–15.

45. Yoshihara K, Ikenouchi J, Izumi Y, Akashi M, Tsukita S, Furuse M. Phosphorylation state regulates the localization of Scribble at adherens junctions and its association with E-cadherincatenin complexes. Exp Cell Res. 2011;317:413–22.
46. Ivanov AI, Young C, Beste KD, Capaldo CT, Humbert PO, Brennwald P, et al. Tumor suppressor scribble regulates assembly of tight junctions in the intestinal epithelium. Am J Pathol. 2010;176:134–45.
47. Humbert PO, Dow LE, Russell SM. The Scribble and Par complexes in polarity and migration: friends or foes Trends Cell Biol. 2006;16:622–30.
48. Scallan E, Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, et al. Foodborne illness acquired in the United States major pathogens. Emerg Infect Dis. 2011;17:7–15.
49. Plattner F, Soldati FD. Hijacking of host cellular functions by the Apicomplexa. Annu Rev Microbiol. 2008;62:471–87.
50. Blader IJ, Saeij JP. Communication between Toxoplasma gondii and its host: impact on parasite growth, development, immune evasion and virulence. APMIS. 2009;117:458–476.
51. Peixoto L, Chen F, Harb OS, Davis PH, Beiting DP, Brownback CS, et al. Integrative Genomic Approaches Highlight a Family of Parasite-Specific Kinases that Regulate Host Responses. Cell Host Microbe. 2010;8:208–18.
52. Bradley PJ, Ward C, Cheng SJ, Alexander DL, Coller S, Coombs GH, et al. Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in Toxoplasma gondii. J Biol Chem. 2005;280:34245–58.
53. Saeij JP, Boyle JP, Coller S, Taylor S, Sibley LD, Brooke-Powell ET, et al. Polymorphic secreted kinases are key virulence factors in Toxoplasmosis. Science. 2006;314:1780–83.
54. Ong YC, Reese ML, Boothroyd JC. Toxoplasma rhoptry protein 16 (ROP16) subverts host function by direct tyrosine phosphorylation of STAT6. J Biol Chem. 2010;285:28731–40.
55. Rosowski EE, Saeij JP. Toxoplasma gondii Clonal Strains All Inhibit STAT1 Transcriptional Activity but Polymorphic Effectors Differentially Modulate IFNγ Induced Gene Expression and STAT1 Phosphorylation. PLoS ONE. 2012;e51448:7.
56. You ZH, Lei YK, Zhu L, Xia J, Wang B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinformatics. 2013;14:Suppl 8, S10.
57. Cosic I. Macromolecular bioactivity: is it resonant interaction between macromolecules? Theory and applications. IEEE Trans Biomed Eng. 1994;41:1101–14.
58. Cosic I, Pirogova E. Applications of ionization constant of amino, Acids for Protein Signal Analysis within the resonant recognition model. In: Proceedings of 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 1998. p. 1072–5.
59. Nguyen QT, Fablet R, Pastor D. Protein interaction hotspot. Identification using sequence-based frequency-derived features. IEEE Trans Biomed Eng. 2013;11:2993–3002.
60. Pirogova E, Cosic I. Examination of amino acid indexes within the Resonant Recognition Model. Proceeding of the 2nd conference of the Victorian chapter of the IEEE EMBS. Melbourne Australia; 2001. February 2001.
61. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41:808–15.