Networks of Motifs from Sequences of Symbols

Roberta Sinatra,[1,2,*] Daniele Condorelli,[2,3] and Vito Latora[1,2]

[1]*Dipartimento di Fisica ed Astronomia, Università di Catania, and INFN, Via S. Sofia 64, 95123 Catania, Italy*
[2]*Laboratorio sui Sistemi Complessi, Scuola Superiore di Catania, Via San Nullo 5/i, 95123 Catania, Italy*
[3]*Dipartimento di Scienze Chimiche, Sezione di Biochimica e Biologia Molecolare, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy*

(Received 4 February 2010; revised manuscript received 16 August 2010; published 19 October 2010)

We introduce a method to convert an ensemble of sequences of symbols into a weighted directed network whose nodes are motifs, while the directed links and their weights are defined from statistically significant co-occurences of two motifs in the same sequence. The analysis of communities of networks of motifs is shown to be able to correlate sequences with functions in the human proteome database, to detect hot topics from online social dialogs, to characterize trajectories of dynamical systems, and it might find other useful applications to process large amounts of data in various fields.

DOI: 10.1103/PhysRevLett.105.178702 PACS numbers: 89.75.Hc, 87.18.Cf, 89.75.Fb

There are many examples in biology, in linguistics, and in the theory of dynamical systems where information resides and has to be extracted from corpora of raw data consisting in sequences of symbols. For instance, a written text in English or in another language is a collection of sentences, each sentence being a sequence of the letters from a given alphabet. Not all sequences of letters are possible, since the sentences are organized on a lexicon of a certain number of words. In addition to this, different words are used together in a structured and conventional way [1,2]. Similarly, in biology, DNA nucleotides or aminoacid sequence data can be seen as corpora of strings [3–6]. For example, it is well known that proteomes are far from being a random assembly of peptides, since clustering of aminoacids [7] and strong correlations among proteomic segments [8] have been clearly demonstrated. These results give meaning to the metaphor of protein sequences regarded as texts written in a still unknown language [3,9]. Sequences of symbols can also be found in time series generated by dynamical systems. In fact, a trajectory in the phase space can be transformed into a sequence of symbols, by the so-called ''symbolic dynamic'' approach [10]. The basic idea is to partition phase space into a finite number of regions, each of which is labeled with a different symbol. In this way, each initial condition gives rise to a sequence of symbols representing the initial cell, the cell occupied at the first iterate, the cell occupied at the second iterate, and so forth.

In all the examples mentioned above, the main challenge is to decipher the message contained in the corpora of data sequences and to infer the underlying rules that govern their production. In order to do this, one needs (i) to detect the fundamental units carrying information, like words do in language, and (ii) to study their combination syntax in the ensemble of sequences. In fact, information in its general meaning is located not only at the level of strings, but also in their correlation patterns [11,12]. In this Letter, we introduce a method to transform a generic corpus of strings, such as written texts, protein sequence data, sheet music, or a collection of dance movement sequences [13], into a network representing the significant and fundamental units of the original message together with their relationships. The method relies on a statistical procedure to detect patterns carrying relevant information, and works as follows. We first construct a dictionary of the recurrent strings of k letters, called k-motifs. Recurrent strings play, in this more general context, the same role as words in written or spoken languages. We then construct a k-motif network, a graph in which each node is one entry of the dictionary, and a directed arc between two nodes is drawn when the ordered co-occurence of the two motifs is statistically significant in the data set analyzed. We will show how the analysis of topological properties of networks of k-motifs, such as the detection of community structures [14,15], allows us to extract important information encoded in the original data. In particular, we will consider the application of the method to data sets in three different domains: namely, biological sequences of proteins, messages from online social networks, and sequences of symbols generated by the trajectories of a dynamical system.

Let us consider an ensemble $\mathcal{S}$ of S sequences of symbols. Each sequence s ($s = 1, 2, \ldots, S$) is a string of letters from an alphabet $\mathcal{A}$ of A letters, $\mathcal{A} \equiv \{\sigma_1, \sigma_2, \ldots, \sigma_A\}$. In general, the strings can have different lengths. We indicate by l_s the length of sequence s, and by $L = \sum_{s=1}^{S} l_s$ the total length of the ensemble. An example is provided by proteomes. A proteome is a collection of $S \approx 10^4$ proteins of a species. Each protein is a sequence of length l_s, ranging from 10^2 to 10^3, made of symbols from an alphabet $\mathcal{A}$ with $A = 20$ letters, $\mathcal{A} \equiv \{\sigma_1, \sigma_2, \ldots, \sigma_{20}\}$, where each σ labels one of the aminoacids a protein can be made of. We define as k-string a segment of k contiguous letters $x_1 x_2 \cdots x_k$, where $x_i \in \mathcal{A}\ \ \forall\, i$. The number of all possible k-strings is A^k, while from the ensemble of sequences $\mathcal{S}$ we

can select only $L-(k-1)S$ overlapping k-strings, so that some of the possible k-strings do not occur, some of them occur once, others more than once, either in the same or in different sequences of symbols. We define as

$$p^{\mathrm{obs}}(x_1x_2\cdots x_k)=\frac{c(x_1x_2\cdots x_k)}{\sum_{(x_1x_2\cdots x_k)\in\mathcal{A}^k}c(x_1x_2\cdots x_k)} \quad (1)$$

the *observed probability* of a string $x_1x_2\cdots x_k$. This probability is obtained by counting the total number of times, $c(x_1x_2\cdots x_k)$, the string actually occurs in the sequences of the ensemble. To assess for the statistical significance of the string, the probability in Eq. (1) has to be compared with the *expected probability* $p^{\mathrm{exp}}(x_1x_2\cdots x_k)$ of the string occurrence. The latter can be evaluated under different assumptions. In fact, the joint probability $p(x_1x_2\cdots x_k)$ can be written as

$$p(x_1x_2\cdots x_k)=p(x_1x_2\cdots x_{k-1})p(x_k|x_1x_2\cdots x_{k-1}),$$

and different approximations for the conditional probabilities $p(x_k|x_1x_2\cdots x_{k-1})$ lead to different values of the expected probability $p^{\mathrm{exp}}(x_1x_2\cdots x_k)$. Namely, if we assume that the occurrence of a letter does not depend on any of the previous letters, i.e., $p(x_k|x_1x_2\cdots x_{k-1})=p(x_k)$, the expected probability is simply given by the product of the relative frequencies of the string's component letters: $p^{\mathrm{exp}}(x_1x_2\cdots x_k)=p^{\mathrm{obs}}(x_1)\cdots p^{\mathrm{obs}}(x_k)$ [16,17]. By using instead a first order Markov approximation, i.e., $p(x_k|x_1x_2\cdots x_{k-1})=p(x_k|x_{k-1})$, the expected probability can be expressed in the form $p^{\mathrm{exp}}(x_1x_2\cdots x_k)=p^{\mathrm{obs}}(x_1)p^{\mathrm{obs}}(x_2|x_1)\cdots p^{\mathrm{obs}}(x_k|x_{k-1})$, where $p^{\mathrm{obs}}(x_j|x_i)$ is extracted from the countings as $p^{\mathrm{obs}}(x_j|x_i)=c(x_ix_j)/\sum_{x_j}c(x_ix_j)=p^{\mathrm{obs}}(x_ix_j)/p^{\mathrm{obs}}(x_i)$. This latter assumption is based on the fact that there is a minimal amount of memory in the sequence: a symbol of the sequence is correlated to the previous one only. Here, we go beyond the approximation of Markov chains of order 1, by retaining as much memory as possible [4]. We assume

$$p^{\mathrm{exp}}(x_1x_2\cdots x_k)=p^{\mathrm{obs}}(x_1x_2\cdots x_{k-1})p^{\mathrm{obs}}(x_k|x_2\cdots x_{k-1}), \quad (2)$$

where the conditional probabilities can be evaluated from countings as

$$p^{\mathrm{obs}}(x_k|x_2\cdots x_{k-1})=\frac{c(x_2x_3\cdots x_k)}{\sum_{x_k}c(x_2x_3\cdots x_k)}, \quad (3)$$

or can be expressed in terms of the observed probability for shorter sequences as

$$p^{\mathrm{obs}}(x_k|x_2\cdots x_{k-1})=\frac{p^{\mathrm{obs}}(x_2\cdots x_k)}{p^{\mathrm{obs}}(x_2\cdots x_{k-1})}. \quad (4)$$

By using the latter expression, we can finally write the expected probabilities in a more compact form:

$$\begin{aligned}p^{\mathrm{exp}}(x_1)&=p^{\mathrm{obs}}(x_1)\\ p^{\mathrm{exp}}(x_1x_2)&=p^{\mathrm{obs}}(x_1x_2)\\ p^{\mathrm{exp}}(x_1x_2x_3)&=p^{\mathrm{obs}}(x_1x_2)\frac{p^{\mathrm{obs}}(x_2x_3)}{p^{\mathrm{obs}}(x_2)}\\ &\cdots=\cdots\\ p^{\mathrm{exp}}(x_1x_2\cdots x_k)&=p^{\mathrm{obs}}(x_1\cdots x_{k-1})\frac{p^{\mathrm{obs}}(x_2\cdots x_k)}{p^{\mathrm{obs}}(x_2\cdots x_{k-1})}.\end{aligned} \quad (5)$$

This way, the expected probability of a given k-string is evaluated based on observations for strings of up to $(k-1)$ symbols. Therefore, by predicting the probability of appearance with a high order Markov model, our method allows us to highlight the true k-body correlations subtracting from them the effects due to $(k-1)$ and lower order correlations. Based on observed and expected probabilities, a test of statistical significance, for instance a Z-score, is then performed for each k-string. We define k-motifs or recurrent k-strings, the statistically relevant strings whose observed and expected number of occurrences are such as to validate the statistical test adopted, and we indicate as $\mathcal{Z}_k$ the dictionary composed by all the selected k-motifs [18].

Once we have constructed a lexicon of fundamental units, the next goal is to represent in a graph the way they are combined together. Recurrent k-strings can be distributed differently along the sequences: they can appear in single sequence or in more than one sequence, alone or in clusters. To extract the nontrivial patterns of correlated appearance of k-motifs, we need to evaluate the probability for the random co-occurrence of two motifs, when these are uncorrelated. We estimate first the expected probability that motif X is followed by motif Y within a generic sequence of the ensemble $\mathcal{S}$, then we sum over all the sequences of $\mathcal{S}$. We denote as $p(X)$ and $p(Y)$ the probabilities of finding the two motifs in $\mathcal{S}$. In sequence s, motif X can occupy positions ranging from the first to the (l_s-2k)th site, where l_s is the length of s and k is the length of the motif. We have assumed that the two motifs cannot overlap. For each fixed position i of X on s, with $i=1,\ldots,(l_s-2k)$, there are $(l_s-2k+1-i)$ possibilities for Y to appear in the sequence. Hence, the number of expected co-occurences of X and Y within s is given by: $\sum_{i=1}^{l_s-2k}(l_s-2k+1-i)p(X)p(Y)$. In order to obtain the expected number of co-occurrences, we have to sum over all the sequence in the ensemble $\mathcal{S}$. We finally get

$$\begin{aligned}N^{\mathrm{exp}}(Y|X)&=p(X)p(Y)\sum_{s=1}^{S}\sum_{i=1}^{l_s-2k}(l_s-2k+1-i)\\ &=\frac{1}{2}p(X)p(Y)\sum_{s=1}^{S}(l_s-2k+1)(l_s-2k+2).\end{aligned} \quad (6)$$

For each value of k, we are now able to construct the k-motif network of the ensemble $\mathcal{S}$, i.e., a directed network whose nodes are motifs in the dictionary $\mathcal{Z}_k$, and an arc points from node X to node Y if the number of times Y

follows X in the ensemble of sequences is statistically significant. Furthermore, a weight can be associated to the arc from X to Y, based on the extent to which the co-occurrence of the two motifs deviates from expectation.

This approach is able to represent the correlation patterns encrypted in the ensemble of sequences into a single object, the k-motif network. Then, graph theory allows us to extract information from the structural properties of the network and to retrieve the main message encoded in the original sequences. In particular, it is interesting to study the components of the k-motif network or, if the graph is connected, its community structures, i.e., those groups of nodes tightly connected among themselves and weakly linked to the rest of the graph [15].

In the following, we will consider the application of the method to three different data sets, belonging to three contexts as diverse as biology, social dialogs, and dynamical systems. We will show how the community analysis of the related k-motif networks enables one to extract functional domains in proteomes, social cascades and hot topics in Twitter, and the increase of chaoticity in deterministic maps.

In the biological context, many methods based on strings deviating from expectancy in genome [4,19] or in a proteome [20] have already been used to make functional deductions. Although they provide insight into many biological mechanisms [17], this approach turns out to be not sufficient for a complete and exhaustive interpretation of the genomic and proteomic message. A fundamental key to its comprehension is in fact hidden in the correlations among recurrent patterns of strings, which are perfectly represented at a global scale in terms of k-motif networks. Various features of these correlations translate into structural properties of k-motif networks. In Fig. 1 we illustrate, as an example, the 3-motif graph derived from the ensemble of human proteins (see [21] for details about the data set). We have detected 15 different communities in the graph, labeled in the figure with different colors and numbers. By means of a research in biological databases, we can show that linked couples of motifs belonging to the same community all co-occur in the same kind of protein domains and that one can associate 9 of these 15 communities just to one domain (see Table I in [21]). These results are outstanding compared to the current methods to extract functional protein domains, all based on multialignment of sequences, and cannot be obtained if one uses a lower order Markov model, meaning that it is fundamental to take into account both short- and long-range correlations (for more details on the k-motif networks in proteomes, see [21]).

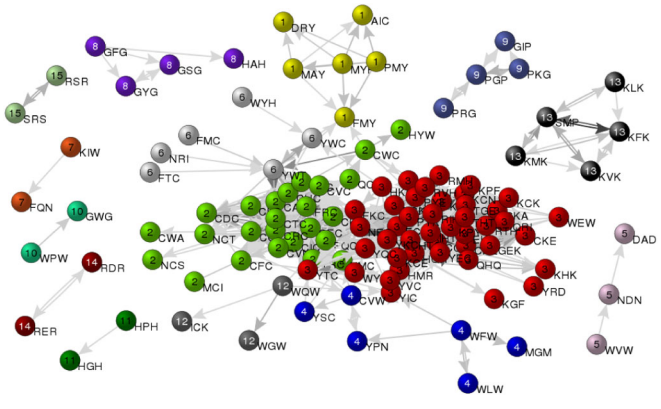

FIG. 1 (color online). The 3-motifs network of the human proteome. Nodes belonging to the same community are labeled by the same number and share the same color. Most of the communities can be associated to a functional domain as described in Table I in [21].

Important information from k-motif networks can also be retrieved from data sets of social dialogs and microblogging web sites, such as Twitter. Although in these cases, in principle, a dictionary is *a priori* known, not all terms used in the Internet language are always listed in the dictionary [22]: abbreviations, "leet language" words, and names of web sites or of public personages are just some examples. Moreover, some expressions or combinations of terms appear more frequently in some periods or contexts due to the interest in some hot topics. We have found that communities of k-motif networks derived from microblogging sequences in Twitter during the United Kingdom election in April 2010 are able to detect exactly those hot topics which generate information cascades [23], as shown in Fig. 1 and Table II of [21]. In Table I we report the links with the highest significance together with the tweet associated to their community. Each tweet was the origin of a cascade and can be associated with a specific topic or event discussed during the election campaign (see [21] for details).

Finally, k-motif networks carry important information on sequences of symbols generated from trajectories of dynamical systems by the so-called "symbolic dynamic"

TABLE I. The eight most significant links found in the Twitter data set [21]. The links belong to five different communities, each corresponding to a specific tweet or expression that generated a topic cascade. The topics are poll results from various journals and TV channels (communities 1 and 2), a satiric web site on the election (community 3), a proposal for a 4th debate among leaders (community 4), and a hashtag (community 5).

Comm.	Motif 1	Motif 2	$\frac{p^{\text{obs}}}{p^{\text{exp}}}$	Expression or tweet
1	9cle 5bro	gg27 wn29	955.3 894.8	GUARDIAN ICM POLL Cameron 35% Brown 29% Clegg 27%
2	son4 don4	4cle 2cam	924.3 881.7	Brown wins on 44%, Clegg is second on 42%, Cameron 13% None of them 1%
3	lapo	mete	892.3	www.slapometer.com
4	swed nesd	nesd ayni	864.7 826.1	hey Dave, Gordon and Nick : how about a 4th debate on Channel 4 this wednesday night without the rules?!
5	isob	eymu	831.4	#disobeymurdoch

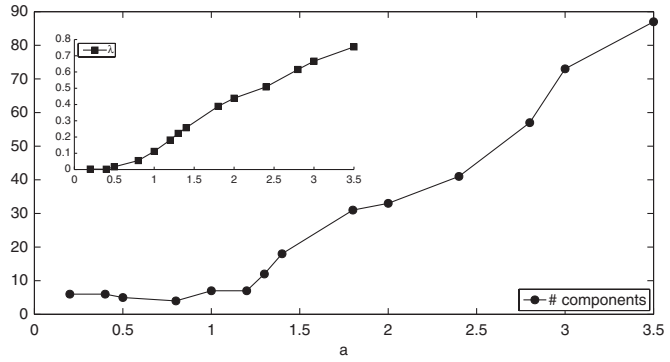

FIG. 2. Standard map. Number of components in the 3-motifs networks (main figure) and the Lyapunov exponent (inset), as a function of the nonlinearity parameter a.

approach [10]. One is able, for instance, to distinguish ensembles of sequences generated by deterministic maps from those generated by stochastic processes, by looking at the number of components and communities in the k-motif network. In fact, the method, when applied to sequences generated by deterministic equations that are increasingly nonlinear, still finds short motifs, while the same does not occur for ensembles of random sequences. Furthermore, we have found that the higher the nonlinearity in a conservative deterministic dynamical system, the more disconnected the corresponding k-motif network. In Fig. 2, we show an example of this behavior for a well-known two-dimensional area-preserving deterministic map, the standard map [24]. Each point in Fig. 2 represents the number of components in the 3-motif network obtained from an ensemble of trajectories produced for a specific value of the nonlinearity parameter a. We observe that the number of components increases with a, and this behavior is similar to that of the positive Lyapunov exponent of the map, shown in the inset (see also [21]).

Summing up, in this Letter we have introduced a general method to construct networks out of any symbolic sequential data. The method is based on two different steps: first it extracts in a ''natural'' way motifs, i.e., those recurrent short strings which play the same role words do in language, then it represents correlations of motifs within sequences as a network. Important information from the original data are embedded in such a network and can be easily retrieved as shown with different applications (a biological system, a social dialog, and a dynamical system). With respect to previous linguistic methods, our approach does not need the *a priori* knowledge of a given dictionary, and also allows us to compare different ensembles corresponding, for example, to different values of control parameters in dynamical systems. All this makes the method very general and opens up a wide range of applications from the study of written text to the analysis of sheet music or sequences of dance movements. Moreover, the method does not use parameters on the position of motifs in order to correlate them, since co-occurrences are computed within sequences, which represent natural interruptions of a corpora of data (proteins in a proteome, posts in a blog, different initial conditions in a symbolic dynamics, etc.).

We thank A. Giansanti and V. Rosato for stimulating discussions on the biological applications of the method, S. Scellato for providing us with the Twitter data set, and M. De Domenico for his interesting comments on applications to dynamical systems. This work was partially supported by the Italian To61 INFN project.

*Corresponding author.
roberta.sinatra@ct.infn.it

[1] R. Ferrer i Cancho, R. V. Solé, and R. Köhler, Phys. Rev. E **69**, 051915 (2004); R. Ferrer i Cancho and R. V. Solé, Proc. R. Soc. B **268**, 2261 (2001).
[2] A. E. Motter *et al.*, Phys. Rev. E **65**, 065102 (2002); E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, PLoS ONE **4**, e7678 (2009).
[3] D. B. Searls, Nature (London) **420**, 211 (2002).
[4] V. Brendel, J. S. Beckmann, and E. N. Trifonov, J. Biomol. Struct. Dyn. **4**, 011 (1986).
[5] C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[6] N. Scafetta, V. Latora, and P. Grigolini, Phys. Rev. E **66**, 031906 (2002).
[7] V. Rosato, N. Pucello, and G. Giuliano, Trends Genet. **18**, 278 (2002).
[8] H. J. Bussemaker, H. Li, and E. D. Siggia, Proc. Natl. Acad. Sci. U.S.A. **97**, 10 096 (2000).
[9] Z. Solan *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 11 629 (2005).
[10] C. Beck and F. Schlögl, *Thermodynamics of Chaotic Systems* (Cambridge University Press, Cambridge, England, 1993).
[11] L. Lacasa *et al.*, Proc. Natl. Acad. Sci. U.S.A. **105**, 4972 (2008).
[12] J. Zivkovic, M. Mitrovic, and B. Tadic, *Studies in Computational Intelligence*, edited by S. Fortunato *et al.*, Complex Networks Vol. 207 (Springer, New York, 2009) pp. 23–34.
[13] E. Bradley *et al.*, Open Artif. Intell. J. **4**, 1 (2010).
[14] S. Boccaletti *et al.*, Phys. Rep. **424**, 175 (2006).
[15] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[16] L. Ferraro *et al.*, arXiv:q-bio/0410011v2.
[17] A. Giansanti *et al.*, Parasitol. Res. **101**, 639 (2007).
[18] The term motif is chosen in analogy with the concept of network motifs, i.e., recurrent patterns of nodes and links in a graph. U. Alon, Nat. Rev. Genet. **8**, 450 (2007).
[19] M. Caselle, F. Di Cunto, and P. Provero, BMC Bioinf. **3**, 7 (2002); D. Corà *et al.*, BMC Bioinf. **5**, 57 (2004).
[20] P. Nicodème, T. Doerks, and M. Vingron, Bioinformatics, Suppl. 2, **18**, 161 (2002).
[21] See supplementary material at http://link.aps.org/supplemental/10.1103/PhysRevLett.105.178702 for details and supplementary results about the application of the method in the three data sets.
[22] M. Mitrovic and B. Tadic, Eur. Phys. J. B **73**, 293 (2010).
[23] K. Lerman and R. Ghosh, arXiv:1003.2664.
[24] B. V. Chirikov, Phys. Rep. **52**, 263 (1979).