Supplementary material to “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[*]

The properties of k-motif networks can reveal important characteristics of the message encrypted in the original data, as the analysis of topological quantities (clustering coefficient, average path length and degree distributions) has helped to understand various linguistic features in networks of words co-occuring in sentences [S1, S2], and also to model how language has evolved in networks of conceptually-related words [S3]. We present here details and further results on the application of the method described in the main article to three different datasets: proteomic sequences, short text messages acquired from *Twitter*, the well-known social network and microblogging platform, and ensembles of sequences derived from dynamical trajectories of the standard map by means of a symbolic dynamics approach.

BIOLOGICAL SEQUENCES

Methods to study over- or under-representation of particular motifs in a complete genome [S4, S5] or in a proteome [S6], have already been proposed, and the results have been used to make functional deductions. Although the information contained in strings deviating from expectancy is useful for the analysis of many biological mechanisms [S7], it 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. The spatial structure of proteins provides an example: when a protein folds, segments distant on the sequence come to be close to each others in the space. This can happen because two (or more) segments need to physically interact in order to perform the biological function the protein is supposed to go through. Such a mechanism translates into a statistical correlation between short motifs of aminoacids, which is well captured by an analysis in terms of k-motif networks.

Human proteome

In our application, we have considered the ensemble of sequences relative to the human proteome [S8]. It consists of 34180 aminoacidic sequences of variable size, with an average length of 481 letters. For this dataset, we have computed the probabilities p^{obs} and p^{exp} for each of the $20^3 = 8000$ possible strings of three aminoacids, and we have selected as 3-motifs the strings satisfying $\frac{p^{obs}}{p^{exp}} > \left\langle \frac{p^{obs}}{p^{exp}} \right\rangle + 2\sigma$, hence creating the dictionary Z_3 [S9]. The entries of the dictionary are the nodes of the 3-motif network. The node X is then linked to Y with a directed arc if the number of times that motif Y follows motif X within the same protein is statistically significant, according to the relation: $\frac{p^{obs}(Y|X)}{p^{exp}(Y|X)} > \left\langle \frac{p^{obs}(Y|X)}{p^{exp}(Y|X)} \right\rangle + 2\sigma$. The statistical significance $\frac{p^{obs}(Y|X)}{p^{exp}(Y|X)}$ is also the weight of the arc. In this way we obtain the 3-motif graph of 199 nodes and 1302 directed links, shown in Fig. 1 of the main article. The graph has 86 isolated nodes (not displayed in Figure), while the remaining 113 nodes are organized into 10 weak components. The largest component of the graph contains 5 clusters, detected by means of the MCl algorithm [S10]. Therefore, 15 different communities are present in the graph. In Table I we report, for each community, the number of nodes and its total internal weight, defined as the sum of the weights of links between nodes of the communities normalized by the sum of the weights of links incident in nodes of the community. By submitting a query to the Prosite database [S11] we have obtained, for each couple of

TABLE I. List of communities in the 3-motif network of the human proteome. Community labels as in Fig. 1 of the main text, number of nodes, total internal weight, associated domain, and the domain specificity are reported.

	# nodes	Internal weight	Domain	Domain recognition
1	6	83,30%	Olfactory receptor	171/175
2	25	74,91%	—	
3	43	94,13%	Zinc Finger	1345/1364
4	6	55,42%	G-protein and CUB-Sushi	9/11
5	3	100%	Cadherin	330/347
6	4	100%	Lipoproteins	16/19
7	2	100%	Homeobox	65/84
8	4	100%	—	
9	4	100%	Collagen	271/482
10	2	100%	Serine protease	22/51
11	2	100%	—	
12	3	60,30%	C-type proteins	3/4
13	5	100%	—	
14	2	100%	—	
15	2	100%	—	

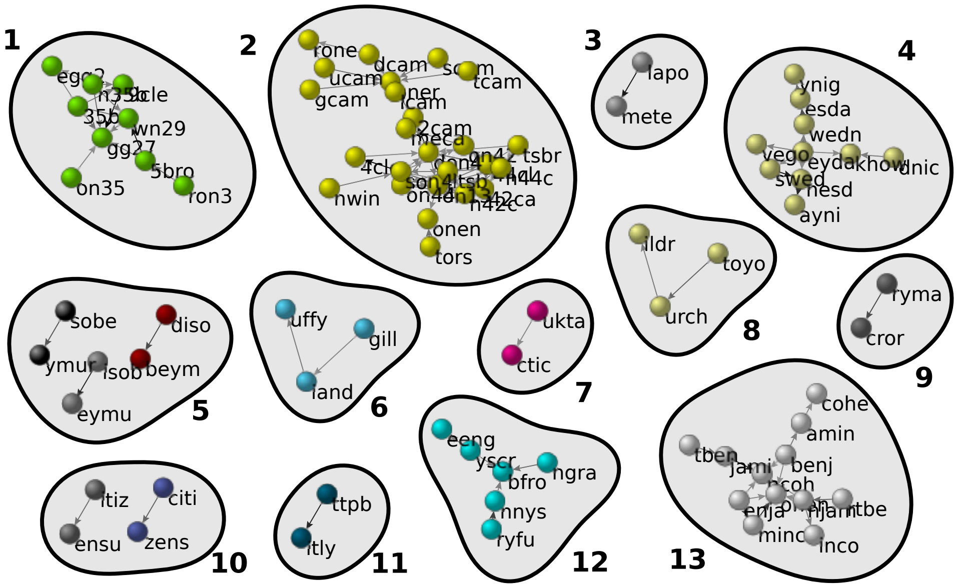

FIG. 1. Components of the 4-motifs network of the twitter dataset. Each component and its associated topic are described in table II.

connected motifs belonging to the same community, the list of all proteins, classified by domain, where the two motifs co-occur. The results show that linked couples of motifs belonging to the same community, all co-occur in the same kind of domains. In addition to this, one can associate 9 of these 15 communities just to one protein domain, since the majority of co-occurrences emerge in proteins matching a well-defined function. In Table I we report, when possible, the association to a single protein domain, together with the ratio between the number of times the couple of motifs with the highest weight occurred in that specific domain, and the total number of co–occurrences in the database.

Analogous results were also found for the 4-motif graph [S12], while it is not possible to derive the same kind of information by using lower order Markov models to construct dictionaries. For example, the 3-motif network constructed with a dictionary based on a lower order approximation rather than on a 2-bodies Markov chain, exhibits a community structure with just four communities, none of which could be identified with a functional protein domain.

SOCIAL NETWORKS AND MICROBLOGGING

By means of k-motif networks, information can also be retrieved from datasets of social dialogs and microblogging websites. Although in these cases, in principle, a dictionary is a-priori known, not all terms used in the Internet language are always listed in a dictionary: abbreviations, puns, leet language words [S13], names of websites or names of public figures, are just some examples. Moreover, some expressions or combinations of terms appear more frequently in some periods or contexts due to the interest to some hot topics. In addition to this, the method of k-motif networks turns to be very useful in all those contexts where it is necessary to process and compact information from large amount of symbolic data. This is the case of Internet, where the amount of text data provided by blogs, dialogs in social networks, forums, etc. is growing and growing.

In the following, we provide details on how network of motifs are able to deduce information about hot topics and cascades [S15, S16] in a dataset extracted from Twitter, a well-know platform for social networking and microblogging.

Twitter

Twitter [S14] is a social networking and microblogging service which allows users to send short messages known as *tweets*. Tweets are composed only of text, with a strict limit of 140 characters: they are displayed on the author's profile page and delivered to the authors subscribers, who are also known as "followers". The dataset we have analyzed is a collection of 28143 tweets, crawled on two days, from the 23rd to 24th April 2010, and selected through the Twitter Streaming API [S17] if they contained the string *#leadersdebate*. The choice of such a keyword, called in Twitter also *hashtag*, was aimed to select all those tweets concerning electoral campaign in UK, where general election to elect the members of the House of Commons would have taken place two weeks later. We have analyzed the dataset removing all blank spaces between words and all symbols that where not numbers or letters (punctuation, symbols like $, @, *, etc.) and not distinguishing between lower- and upper-case letters. From these sequences, dictionaries of motifs $\mathcal{Z}_3$ and $\mathcal{Z}_4$ have been extracted, selecting respectively the 10% and 1% of most significant strings of 3 and 4 letters. As described in the main text, we have constructed networks whose nodes represent the entries of a dictionary, and an arc is drawn from the node representing string X to the node standing for string Y, if $p^{obs}(Y|X)/p^{exp}(Y|X)$ is greater than a certain threshold. In Fig. 1, we show the 4-motifs network when the threshold is set equal to 400 (isolated nodes not reported). Such a high threshold is chosen to have a small network that can be easily visualized and studied. More information can be obtained by setting the threshold to lower values or analyzing networks made up of motifs of different length k. Searching in the original dataset the connected motifs, it is possible to associate each component to a particular tweet which generated a cascade or with a specific expression, related to a specific hot topic discussed by users of the microblogging platform. For all components of Fig. 1, we report in Table II the tweet or expression associated and its meaning. For example, component 1 and 4 can be associated to two exit polls disclosed on those days by two different journals, or component 6 to the name “Gillian Duffy”, a 65-years old pensioner involved in a political scandal with British PM Gordon Brown during the election tour (Brown’s remarks of her as a “bigoted woman” were accidentally recorded and broadcast).

TABLE II. We report the number of nodes, the number of links, the tweet or the expression containing the motifs and the related topic for each of the13 communities represented in Fig. 1.

Comm.	Nodes	Links	Expression or Tweet	Topic
1	9	13	*GUARDIAN ICM POLL Cameron 35% Brown 29% Clegg 27%*	poll results from various websites, journals, tv channels, etc
2	25	33	*Brown wins on 44%, Clegg is second on 42%, Cameron 13% None of them 1%*	poll results from various websites, journals, tv channels, etc
3	2	1	www.slapometer.com	A funny website on the election
4	10	11	*hey Dave, Gordon and Nick : how about a 4th debate on Channel 4 this wednesday night without the rules?!*	Proposal for a 4th debate among leaders, made by a journalist on his Twitter page
5	6	3	#disobeymurdoch	Twitter hashtag
6	3	2	Gillian Duffy	Woman branded a ’bigot’ by Gordon Brown in general election campaign [S20]
7	2	1	Tactical voting	*Strategy that when a voter misrepresents his or her sincere preferences in order to gain a more favorable outcome [S21]*
8	6	5	*Cameron: I believe that if you’ve inherited hard all your life you should pass it on to your children*	Electoral campaign from David Cameron
9	2	1	Henry Macrory	Head of press for the Conservatives, owner of a twitter account
10	4	2	#citizensuk	Twitter hashtag
11	2	1	http:// ... .ly	Format of shortened weblinks in twitter
12	6	5	*Very funny screengrab from the LeadersDebate*	About a funny picture of the leaders debate on BBC [S19]
13	12	14	Benjamin Cohen	Journalist of Channel 4 News [S18]

SYMBOLIC DYNAMICS

Symbolic dynamics is a general method to transform trajectories of dynamical systems into sequences of symbols. The distinct feature in symbolic dynamics is that time is measured in discrete intervals. So at each time interval the system is in a particular state. Each state is associated with a symbol and the evolution of the system is then described by a sequence of symbols. The method turns to be very useful in all those cases where system states and time are inherently discrete. In case the time scale of the system or its states are not discrete, one has to set a coarse-grained description of the system. Different initial conditions usually generate different trajectories in the phase space, which map onto different sequences of symbols. A large number of initial conditions produces an ensemble of sequences whose analysis can be addressed with the method based on networks of motifs, as described in the main article.

In the following, we will describe the application of the method to the standard map, and we will show how the related networks of motifs shape according to its chaotic behavior.

Standard Map

The standard map, also known as Chirikov map, is a bidimensional area-preserving chaotic map. It maps a square with side 2π onto itself [S22]. It is described by the equations:

$$\begin{cases} x_{t+1} = p_t + a \sin x_t \mod 2\pi \\ p_{t+1} = p_t + x_{t+1} \mod 2\pi \end{cases} \qquad (1)$$

where t represents time iteration and a is a parameter assuming real values. The map is increasingly chaotic as a increases (see inset of Fig. 2 in the main article to see a plot of the Lyapunov exponent as a function of the parameter a). For $a = 0$, the map is linear and only periodic and quasiperiodic orbits are allowed. When evolution of trajectories are plotted in the phase space (the *xp* plane), periodic orbits appear as closed curves, and quasiperiodic orbits as necklaces of closed curves whose centers lie in another larger closed curve. Which type of orbit is observed depends on the map’s initial conditions. When the nonlinearity of the map increases, for appropriate initial conditions it is possible to observe chaotic dynamics.

In order to obtain sequences from the standard map (1) by means of the symbolic dynamic approach [S23], one needs to make a coarse graining of the phase space, defining a discrete and finite number of possible states the trajectory can occupy. This way it is possible to associate a symbol to each of the possible states and derive a sequence according to the trajectory originating from an initial condition. We have coarse-grained the phase space into 25 (5×5) squares of equal size and we have derived for different values of the parameter a, 10^4 sequences of 10^3 symbols. In other words, this means to follow for 10^3 time steps the trajectories originating from 10^4 different initial conditions.

The idea is that closed orbits or quasi periodic-ones correspond to correlations between motifs and therefore in links of the graph of motifs. When the map becomes more and more chaotic, closed orbits disappear and, correspondingly, the networks break in many components. In the extreme limit of map highly chaotic ($a > 3$), the network of motifs are completly disconnected, with all nodes isolated. Nevertheless, this scenario is different from the one generated by stochastic sequences, since in this case motifs would not be detected, while this still happens in the chaotic map, although only for small values of k. This result is well depicted in Fig. 3 of the main article, where the number of components of the 3-motif graphs is plotted as a function of the value a of the map generating the ensemble. This curve is shown to have the same behavior of the Lyapunov exponent, as reported in the inset of the same figure.

[*] Corresponding author: roberta.sinatra@ct.infn.it

[S1] R. Ferrer i Cancho and R.V. Solé, *Proc. R. Soc. Lond. B* **268**, 2261 (2001).

[S2] S.M.G. Caldeira, T.C. Petit Lobão, R.F.S. Andrade, A. Neme, and J.G.V. Miranda, *Eur. Phys. J. B* **49**, 523 (2006).

[S3] A. Motter, A.P.S. De Moura, Y.C. Lai, and P. Dasgupta, *Phys. Rev. E*, **65**, 065102 (2002).

[S4] V. Brendel, J.S. Beckmann, and E.N. Trifonov, *Journal of Biomolecular Structure & Dynamics* **4**, 011 (1986).

[S5] M. Caselle, F. Di Cunto, and P. Provero, *BMC Bioinformatics* **3**, 7 (2002); D. Corà, F. Di Cunto, P. Provero, L. Silengo, and M. Caselle, *BMC Bioinformatics* **5**, 57 (2004).

[S6] P. Nicodème, T. Doerks, and M. Vingron, *Bioinformatics* **18**: S161, Suppl.2 (2002).

[S7] A. Giansanti, M. Bocchieri, V. Rosato, and S. Musumeci, *Parasitol. Res.* **101**, 639 (2007).

[S8] Data downloaded from the *Consensus Coding Sequence database* (CCDS), http://www.ncbi.nlm.nih.gov/CCDS/, version Hs35.1.

[S9] With the notation $\langle p\left(x\right)\rangle$, we denote the average of $p(x)$ over all the possible configurations of x and with σ the standard deviation of the distribution.

[S10] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, *Nucleic Acids Research* **30**:1575 (2002).

[S11] http://www.expasy.ch/prosite

[S12] R. Sinatra, D. Condorelli, A. Giansanti, V. Latora, V. Rosato, *Analysis of proteomes by means of k-motif networks*, in preparation.

[S13] http://en.wikipedia.org/wiki/Leet

[S14] www.twitter.com

[S15] K. Lerman and R. Ghosh, in Proc. of ICWSM (2010).

[S16] M. Cha, A. Mislove, B. Adams, K. P. Gummadi, in Proc. of WOSN 08 - USA (2008).

[S17] http://apiwiki.twitter.com/Streaming-API-Documentation

[S18] http://en.wikipedia.org/wiki/Benjamin_Cohen_%28journalist%29

[S19] http://twitpic.com/1jge7b

[S20] http://www.guardian.co.uk/politics/2010/jul/26/gillian-duffy-backs-david-miliband

[S21] http://wiki.electorama.com/wiki/Tactical_voting

[S22] B.V. Chirikov, Phys. Rep. **52**:263 (1979).

[S23] V.M. Aleksev and M.V. Yakobson, *Phys. Rep.* **75**:287 (1981).