JDREAM. Journal of interDisciplinary REsearch Applied to Medicine
JDREAM (2019), v. 2 i. 2, 27-33
ISSN 2532-7518
DOI 10.1285/i25327518v2i2p27
http://siba-ese.unisalento.it, © 2018 Università del Salento

Can graph theory discriminate between aptamer-protein configurations?

Emanuela Cianci[1], *Rosella Cataldo*[1], *Eleonora Alfinito*[2]

[1] Department of Mathematics and Physics "Ennio de Giorgi", University of Salento, Via Monteroni, Lecce, Italy
[2] Department of Innovation Engineering, University of Salento, Via Monteroni, Lecce, Italy

Corresponding author: Rosella Cataldo
rosella.cataldo@unisalento.it

Abstract

Graph theory has been extensively applied in the study of proteins, although few applications still exist on biomolecules like aptamer-protein complexes, whose structure is in silico obtained. Aptamers represent a challenging field of research, especially for their involvement in therapy, diagnosis and early detection of illness; furthermore, the in vivo procedures to synthetize them are quite expensive and more time-consuming than those in silico. This paper is focused on the question if and how general network parameters are able to anticipate some features of those biomolecules. The rationale resides in the fact that, studying a large set of aptamer-angiopoietin complexes, two different types of conformers are manifest. Both types could be present in a real sample with their relative amount reflecting, in a typical population shift scenario, the affinity of the whole sample.

Keywords: graph theory, in silico structure of biomolecules, aptamer-angiopoietin complex.

Introduction

In recent years, several techniques have been developed for the early diagnosis and the ongoing follow up of several fatal diseases. The main prerequisites are: to be minimally invasive, to make use of biodevices based on nanosized materials, to show high biocompatibility and long-term stability. Aptamers are small fragments of ssDNA or RNA, entitled to play a primary role in this challenge. They are artificially assembled for achieving high binding affinity to their targets (from proteins to simple ions). Furthermore, low dimension, short half-time, low immunogenicity and production costs make them formidable competitors of antibodies in both diagnosis and therapy.

Nevertheless, it has to be pointed out that acceptance and use of aptamers in industry and by (bio)pharmaceutical companies is still rare (Famulok and Mayer 2014). The reason is primarily linked to the lack of information concerning the binding mechanism with target, intrinsically interesting, since a single referencing model does not exist and different possible scenarios have been drawn to explain the formation of aptamer-receptor complex, including conformational shift, induced fit, lock and key (Koshland 1995; Kinghorn et al. 2017). Furthermore, the in vivo exploration (Tuerk and Gold 1990) is quite difficult and can be extremely time-consuming. Public data, such as the European Bioinformatics Institute (EMBL-EBI https: //www.ebi.ac.uk), provide free and open access to a range of bioinformatics applications for sequence analysis, but, mainly for aptamers, crystallographic characterization is still in its infancy.

In parallel with the in vitro techniques, many computational procedures have been developed for predicting 3D structure from sequence information. In general, aptamer-protein complexes are obtained passing through two main steps: first, aptamers have to be folded, thus the folded (3D) sequences have to be docked with the protein. Docking tools provide a ranking for the obtained configurations, taking into account many parameters, such as the interaction energies between the two molecules, the desolvation and solvation energies associated with the interacting molecules and the entropic factors that occur upon binding (Trott and Olson 2010).

However, the ranking assignment continues to be an open question (Kaufmann et al. 2017; Cataldo et al. 2018). In a recent investigation

(Cataldo et al. 2019), a new estimate was proposed to complement the computational ranking outcomes, based on maximum likelihood criteria of the topological and electrical properties of aptamer-protein complexes. It was applied on a set of anti-angiopoietin(Ang2) aptamers, whose performances are known from the experiments (Hu et al. 2015). From the analysis, two principal types of conformers were identified and a deep discussion of the possible scenarios linked to those results was performed. The present paper aims to formulate a novel and simplified procedure to analyse all those different conformers, looking at topological characteristics of the networks representing the biomolecules. We show that important features can be extracted, another type of arrangement aptamer-Ang2 is observed, and finally, we suggest on which flaws and imperfection research has to focus its attention.

Background

Angiopoietins are part of a family that have a prominent role in vascular disease (Fagiani and Christofori 2013) and vasculature upregulation of many types of tumours (White et al. 2003). The most extensively studied angiopoietins are Ang1 and Ang2, especially for their involvement in cancer therapy. Ang2 is a compact protein with three domains, named A, B, and P. The P domain is the most divergent (Figure 1), both in sequence and in structure, among the fibrinogen homologs, and it is the site of ligand binding for most fibrinogen domain-containing proteins (Barton et al. 2005).

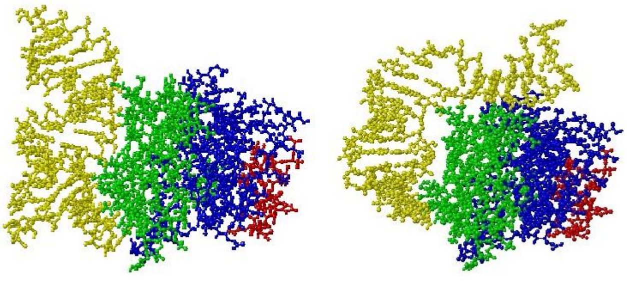

Figure 1. Hair (left) and Belt (right) aptamer-Ang2 conformers: in the Hair, the binding site of the aptamer (yellow) is primarily the P domain (green), while in the Belt, the binding site of the aptamer (yellow) is also in the B domain (blue).

Hu et al. (2015) selected and studied the affinity properties of a set of five sequences: two of them were natural anti-Ang2 and anti-Ang1 aptamers, while the other three were mutated sequences; the affinity for the target was evaluated by means of measurements with a surface plasmon resonance (SPR) biosensor. Conflicting results concerning the performances of the sequences were observed; in particular, the sequence with the best computational rank among the five considered (named Seq2_12_35) was one of the worst performing in experiments.

In (Cataldo et al. 2018) a procedure to cope with the problem of docking of ligands characterized by a large number of flexible freedom degrees was presented, by using as case-study the five aptamers proposed in (Hu et al. 2015). The 3D structure of the aptamers was obtained by using the free SimRNA software (Boniecki et al. 2015), docking was performed by means of rigid or flexible roto-translations in the free AutoDock Vina tool (Trott and Olson 2010). For scoring the complexes, a single and energetic quantity called "effective affinity" (EA) was proposed, putting together the docking energy provided by AutoDock Vina and the SimRNA energies.

In (Cataldo et al. 2019) a novel kind of scoring was addressed, starting from the same aptamers in (Hu et al. 2015). The study revealed a different type and abundance of conformers, which would best represent each aptamer-protein complex. The one, named Hair, mimics the binding with its natural protein target (Tie2), the other, named Belt, hangs the protein also far from the original binding domain (Figure 1). The results highlighted that: a. the effective affinity (Cataldo et al. 2018) seems not to be able to discriminate between these two types of conformers, because their values are quite similar; b. the Ang2-specific aptamer prefers Hair conformers; c. the Ang1-specific aptamer prefers Belt conformers; d. mutated sequences do not show a distinct preference. In other terms, it seems that the Hair conformer characterizes the high affinity complexes, while the Belt conformer characterizes the low affinity complexes.

Graph analysis

Starting from some seminal papers (Watts and Strogatz 1999; Albert and Barabási 2000), graph analysis has received increasing attention with applications to several fields of research, from physics to psychology. In fact, graph analysis is able to capture part of the complexity, which is inherent in most of natural phenomena. The inner idea is to explain multiscale-multiphase-many body phenomena by using interactions, thus in terms of networks.

Network has a topological structure given by the associated graph *G(N,L),* with *N* the node set and *L* the link set (Van Mieghem et al 2014). In proteins, the C_α atom of each amino acid is considered as a node, and two amino acids are considered nearest neighbors if the distance between their C_α atoms is less than an assigned threshold value. This distance or cut-off radius R_C is a free parameter; whose tuning produces a graph more or less connected (Alfinito et al. 2017). Once all the Euclidean distances between the couples of C_α atoms are calculated, a distance matrix is obtained; from the distance matrix, the graph description of the network is represented through its adjacency matrix *A* of size *N x N*, with element:

$a_{i,j}$ = 1 (there is a link), if the distance between node *i* and *j* is less than the assigned R_C;

$a_{i,j}$ = 0 (there is no link), if the distance between node *i* and *j* is greater than the assigned $R_{C.}$

We assume that no self-loops (hence $a_{i,i} = 0$) and no overlapping links exist, i.e. there cannot be more than one link between $a_{i,j}$, therefore, we deal with a *simple graph* (Albert and Barabási 2000).

Among the various parameters that measure the graph characteristics, we point out on the following, giving a short definition.

Degree: the degree *k* is defined as the total number of network connections. The average of k_i over all *i* nodes is called the average degree *[k]* of the network. The spread in node degree is characterized by a distribution function *P(k),* which gives the probability that a randomly selected node has exactly *k* links (Wang 2002).

Assortative mixing and coefficient of assortativity (*r*): a network is said to show 'assortative mixing', if the high-degree nodes tend to be connected with other high-degree nodes, and 'disassortative' when the high-degree nodes tend to connect with low-degree nodes. It is the Pearson correlation coefficient of the degrees at either ends of a link and lies in the range $-1 \leq r \leq 1$; $r = 1$ means perfect assortativity, $r = -1$ means perfect disassortativity, $r = 0$ means no assortativity (random linking). If a network has perfect assortativity ($r = 1$), then all nodes connect only with nodes with the same degree.

Diameter and Shortest path length and average path length: the diameter *D* of a network is the maximal distance between any pair of its nodes. The distance L_{ij} between two nodes *i* and *j* is defined as the number of links along the shortest path connecting them. Then, the average path length L of the network is defined as the distance between two nodes, averaged over all pairs of nodes. L determines the effective size of a network, i.e. the typical separation of pairs of nodes (Wang 2002).

Clustering coefficient: clustering coefficient *C* can be defined as the average fraction of pairs of neighbors of a node, which are also neighbors of each other. Suppose that a node *i* in the network has k_i links, which connect it to k_i other nodes. These nodes are the neighbors of node *i.* Clearly, at most $k_i(k_i - 1)/2$ links can exist between them, and this occurs when every neighbour of node *i* is connected to every other neighbour of node *i.* The clustering coefficient *Ci* of node *i* is defined as the ratio between the number *Ei* of links that actually exist between these *ki* nodes and the total number $k_i(k_i - 1)/2$, namely

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

The clustering coefficient *C* of the whole network is the average of *Ci* over all *i* (Wang 2002).

Largest eigenvalue (*lev*) depends on the highest degree in the graph. For any *k* regular graph *G* (a graph with *k* degree on all the vertices), the eigenvalue with the largest absolute value is *k*. Generally, *lev* increases if the graph contains vertices of high degree, and decreases gradually from the graph with highest degree 6 to the one with highest degree 2 (Vishveshwara 2002).

Materials and Methods

In Table 1 are listed the five different RNA-aptamers employed for the research, they are the same sequences described in (Hu et al. 2015).

Name	Sequence
Seq1	AAAAAACUAGCCUCAUCAGCUCAUGUGCCCCUCCGCCUGGAUCAC
Seq16	AAAAAACUCGAACAUUUCCACUAACCAACCAUACUAAAGCACCGC
Seq2_12_35	AAAAAUUAACCAUCAGAUCAUGGCCCCUGCCCUCUCAAGCACCAC
Seq15_12_35	AAAAAGAGGACGAUGCCGACUAGCCUCAUCAGCUCAUGUCCCCCUC
Seq15_15_38	AAAAAGAGGACGAUGCGGAUUAGCCUCAUCAGCUCAUGUGCCGCUC

Table 1. The five studied sequences, from (Hu et al., 2015). Seq1 is an Ang2-specific aptamer, Seq16 is an Ang1-specific aptamer, the other three are Ang2-mutated sequences.

In Table 2 are listed the percentage/number of Hair and Belt configurations took into account in this paper.

Name	Hair		Belt		
	%	n	%	n	Total n
Seq1	24	43	76	134	177
Seq16	44	103	56	133	236
Seq2_12_35	32	102	68	219	321
Seq15_12_35	35	55	65	102	157
Seq15_15_38	29	26	71	63	89

Table 2. The percentage/number (n) of the studied Hair and Belt configurations.

In the next, the results will be presented on violin plots; outliers, i.e. values more than 1.5 times the interquartile range, or approximately 3 standard deviations in a Gaussian distribution, have been dropped out. In general, we pay careful attention to the outliers, because they can signal skewed distributions rather than a statistical error in a unimodal and symmetric distribution. The RC value best representing our networks was 11.3 Å; this value was sufficiently large to have a connected network, but not so large to hide the node peculiarities (Alfinito et al. 2017).

Hereafter, the terms "network" and "graph" have to be considered synonyms.

Graph representation

In Figure 2 two typical networks for Hair and Belt conformers are drawn. Hair conformers seem to describe "golf-club" networks, in which aptamer is positioned above the Ang2 (dark grey zone), while Belt conformers describe a circular structure above the network protein (dark cyan zone). In both conformers, the average number of links is constant in all the sequences, reporting a value of 6150, for R_C = 11.3 Å.

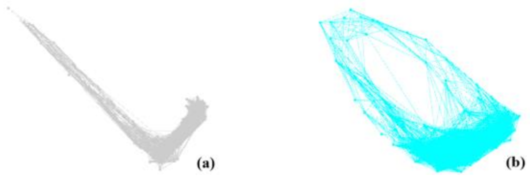

Figure 2. Networks for Hair (a) and Belt (b) conformer. The Hair conformer seems to describe a "golf-club" network, in which aptamer is positioned above the Ang2 (dark grey zone), while the Belt conformer describes a circular structure above the Protein (dark cyan zone).

Assortativity

From Figure 3a, it is evident that the assortativity coefficient assumes a positive value and significantly greater than zero; it varies in the range [0.51- 0.55], showing a very slight increase, as a general trend, in the Hair conformers. Even if seemingly a distinct discrimination does not exist between the two conformations, Figure 3a clearly shows how the shape of the distribution for natural sequences (Seq1 and Seq16) is much more compact (Gaussian) than the mutated ones. These last sequences exhibit a less regular behavior, especially Seq15_12_35 and Seq15_15_38.

Hierarchy

From the findings, it is possible to state that the rank of the network is very high for low degree value; thus, even increasing the rank, the degree decreases exponentially. The average values are almost constant for all the sequences, in both the conformations. Therefore, it seems that it is not possible to discriminate. Figure 3b shows this too weak fluctuation of hierarchy, even if Belt conformations linked to the natural sequences exhibit a more homogeneous shape of the distribution. Under certain aspects, this could be considered as an indicator of the quality of the in silico realization of the aptamer-Ang2 complex.

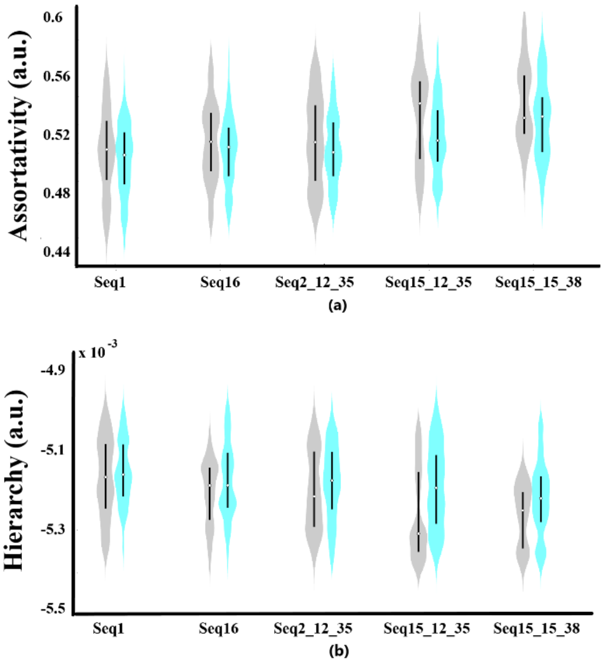

Figure 3. Assortativity (a) and hierarchy (b) coefficient distribution for the five aptamer-Ang2 complexes, gray for Hair and cyan for Belt.

Diameter

The diameter, intended as the maximum distance between two points of the network, could provide important information to discriminate between the two conformers. In particular, it could be hypothesized that the Belt conformers should have the highest values, since the aptamer, positioning itself around the protein, should increase the diameter of the network. However, this difference is not so evident (Figure 4a); the diameter turns out to be comparable for all the sequences, expressing for all the networks a *small-world* behaviour (Watts and Strogatz 1999).

Average length of the path

Figure 4b shows that the average path length in the Hair configurations is greater than the Belt ones, as expected. In fact, when aptamer surrounds protein, it is closer than when it is positioned above, namely in the Hair configuration.

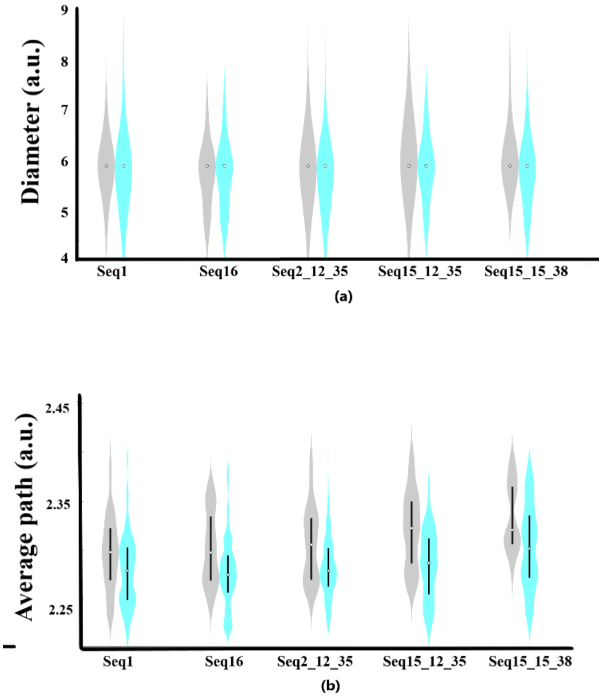

Figure 4. Diameter (a) and average length of the path (b) distribution for the five aptamer-Ang2 complexes, gray for Hair and cyan for Belt.

Average clustering coefficient

The average clustering coefficient is calculated as the average of all the coefficients of local clustering. Figure 5a shows that the Hair configurations values are higher than the Belt ones. In particular, the shape of the distribution points out appreciable variations for the mutant (Seq2_12_35, Seq15_12_35, and Seq15_15_38) compared to the natural (Seq1 and Seq16) sequences.

Maximum eigenvalue

The maximum eigenvalue of the Laplacian matrix was calculated according to the Perron-Fröbenius theorem. This eigenvalue has an associated eigenvector with no negative values for not oriented graphs, as those we examined. This is useful for centrality measures, in a sense that network with high centrality has many nodes with short contacts. This happens, as expected, for Belt conformers, as shown in Figure 5b.

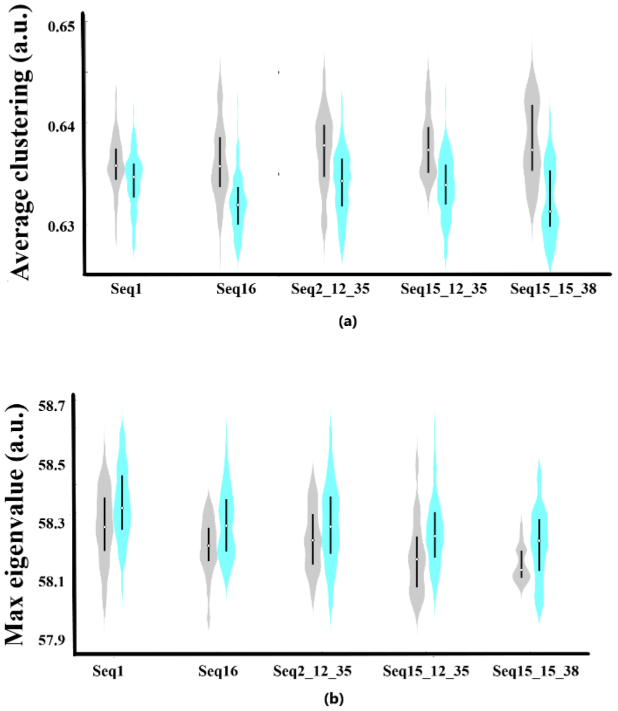

Figure 5. Average clustering coefficient (a) and maximum eigenvalue (b) distribution for the five aptamer-Ang2 complexes, gray for Hair and cyan for Belt.

Conclusions

Graph theory has been extensively applied in several fields of research, but applications to aptamer-protein complexes are still pioneering. The reasons reside in the fact that few structures are experimentally resolved, the results obtained by the application of the theory cannot be compared with a general, well-established model.

In (Cataldo et al. 2019) it has been observed that in silico aptamer-Angiopoietin complexes were arranged in two main conformers: Hair and Belt. Here, we study the same structures by using the concepts of graph theory with the aim to assess if the interpretation of networks could anticipate some characteristics, typical of those configurations.

From the graphs, differences between the two types of conformers are very clearly deduced, starting from the visual representations.

As for other topological parameters, the assortativity and hierarchy coefficients did not highlight a clear difference, but the shape of the distribution clearly reflects two different behaviours, able in appreciably discriminating the conformations. Discrimination becomes more evident in the average length of the path, the clustering coefficient and the maximum value of the eigenvalues of the adjacency matrix. In general, consistent differences are highlighted in the shape of the distribution for all the parameters, this is particularly true in the aptamer-Angiopoietin complexes derived from natural sequences (Seq1 and Seq16), compared to the mutated ones.

For the sake of completeness, another type of conformer (Figure 6) has been noted during the analysis of the networks. It is different from Hair or Belt and further investigations are needed in this regard.

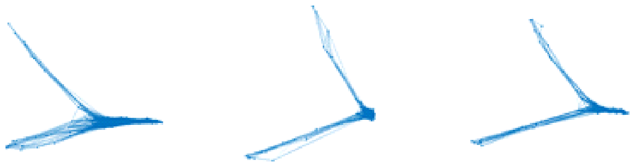

Figure 6. Examples of conformers, different from Hair or Belt configurations.

The achievement of these results can be considered a merit of the application of graph theory and, conversely, a precise indication of the fact that other efforts are required to perfect molecular simulation software, so that a description as close as possible to the real structure of the complexes can be reached.

We consider the proposed procedure effective in the screening process of the outcomes of different computational software, thus, it can be applied as a useful tool for increasing chances of success in designing high-specificity biosensors.

References

- Albert, R., Barabási, A. L. (2002), Statistical mechanics of complex networks, Reviews of modern physics, 74(1), 47.
- Alfinito, E., Reggiani, L., Pousset, J. (2015), Proteotronics: Electronic devices based on proteins, in Sensors (pp. 3-7), Springer, Switzerland.
- Alfinito, E., Reggiani, L., Cataldo, R., et al. (2017), Modeling the microscopic electrical properties of thrombin binding aptamer (TBA) for label-free bio-

sensors, Nanotechnology, 28(6), 065502.

- Barton, W. A., Tzvetkova, D., Nikolov, D. B. (2005), Structure of the angiopoietin-2 receptor binding domain and identification of surfaces involved in Tie2 recognition, Structure, 13(5), 825-832.
- Boniecki, M. J., Lach, G. et al. (2015), SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction, Nucleic acids research, 44(7), e63-e63.
- Trott, O., Olson, A. J. (2010), AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, Journal of computational chemistry, 31(2), 455-461.
- Cataldo, R., Ciriaco, F., Alfinito, E. (2018), A validation strategy for in silico generated aptamers, Computational biology and chemistry, 77: 123-130.
- Cataldo, R., Giotta, L., Guascito, M. R., Alfinito, E. (2019). Assessing the Quality of in Silico Produced Biomolecules: The Discovery of a New Conformer, The Journal of Physical Chemistry B, 123(6), 1265-1273.
- Fagiani E., Christofori G. (2013), Angiopoietins in angiogenesis, Cancer Letter, 328(1), 18-26.
- Famulok, M., Mayer, G. (2014), Aptamers and SELEX in Chemistry Biology, Chemistry Biology, 1054-1058.
- Hu, W. P., Kumar, J. V., Huang, C. J., et al. (2015), Computational selection of RNA aptamer against angiopoietin-2 and experimental evaluation, BioMed research international.
- Kaufmann, A., Butcher, P., Maden, K., et al. (2017), Using in silico fragmentation to improve routine residue screening in complex matrices, Journal of The American Society for Mass Spectrometry, 28(12), 2705-2715.
- Kinghorn, A. B., Fraser, L. A., Lang, S., et al. (2017), Aptamer bioinformatics, International journal of molecular sciences, 18(12), 2516.
- Koshland Jr, D. E. (1995), The key–lock theory and the induced fit theory, Angewandte Chemie International Edition in English, 33(23-24), 2375-2378.
- Tuerk, C., Gold, L. (1990), Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, Science, 249(4968), 505-510.
- Van Mieghem P. (2014), Performance Analysis of Complex Networks and Systems, Cambridge University Press.
- Vishveshwara, S., Brinda, K. V., Kannan, N. (2002), Protein structure: insights from graph theory, Journal of Theoretical and Computational Chemistry, 1(01), 187-211.
- Wang, X. F. (2002), Complex networks: topology, dynamics and synchronization, International journal of bifurcation and chaos, 12.05, 885-916.
- Watts, D.J., Strogatz, S.H. (1999), Collective dynamics of ‘small-world’ networks, Nature, 393, 440–442.
- White, R. R., Shan, S., Rusconi, C. P., Shetty, et al. (2003), Inhibition of rat corneal angiogenesis by a nuclease-resistant RNA aptamer specific for angiopoietin-2, Proceedings of the National Academy of Sciences, 100(9), 5028-5033