ACTA FACULTATIS
MEDICAE NAISSENSIS

UDC: UDC: 577.2:004.9
DOI: 10.5937/afmnai38-28322

Review article

Protein–Protein Interaction Networks and Protein-Ligand Docking : Contemporary Insights and Future Perspectives

Aleksandar Velesinović, Goran Nikolić

University of Niš, Faculty of Medicine, Deaprtment of Chemistry, Niš, Serbia

SUMMARY

Traditional research means, such as *in vitro* and *in vivo* models, have consistently been used by scientists to test hypotheses in biochemistry. Computational (*in silico*) methods have been increasingly devised and applied to testing and hypothesis development in biochemistry over the last decade. The aim of *in silico* methods is to analyze the quantitative aspects of scientific (big) data, whether these are stored in databases for large data or generated with the use of sophisticated modeling and simulation tools; to gain a fundamental understanding of numerous biochemical processes related, in particular, to large biological macromolecules by applying computational means to big biological data sets, and by computing biological system behavior. Computational methods used in biochemistry studies include proteomics-based bioinformatics, genome-wide mapping of protein-DNA interaction, as well as high-throughput mapping of the protein-protein interaction networks. Some of the vastly used molecular modeling and simulation techniques are Monte Carlo and Langevin (stochastic, Brownian) dynamics, statistical thermodynamics, molecular dynamics, continuum electrostatics, protein-ligand docking, protein-ligand affinity calculations, protein modeling techniques, and the protein folding process and enzyme action computer simulation. This paper presents a short review of two important methods used in the studies of biochemistry – protein-ligand docking and the prediction of protein-protein interaction networks.

Key words: in silico, **protein–protein interaction, protein-ligand docking, molecular modeling**

Corresponding author:
Aleksandar Veselinović
e-mail: aveselinovic@medfak.ni.ac.rs

ALGORITHMS AND OTHER RESOURCES INVOLVED IN THE COMPUTATIONAL PREDICTION OF THE NETWORKS FOR PROTEIN–PROTEIN INTERACTION

Protein-protein interactions (PPIs) regulate almost all the processes occurring within the cell, such as DNA transcription and replication, diverse signaling cascades, or metabolic cycles, among others. It is crucial to understand the specificity of these PPIs as they perform other types of cellular functions with other proteins. There has been an exponential elevation of the genomic sequence information amount, particularly over the recent years. However, protein sequence annotation tends to lag in terms of its quality and quantity. The gap between the appropriate biochemical and medical information and raw sequence information is overcome by utilizing high-throughput functional genomic approaches and multi-pronged approaches. Employing computational methods becomes essential in the event that high throughput methods are unable to yield rele-

Table 1. *Computational methods for the prediction of protein-protein interactions (general overview)*

Method		Features
Methods based on genomic context and structure information	Gene fusion	• Mostly employed for small scale proteomes. • In general, does not apply to all genes. • Fusion events are not plentiful, particularly when it comes to prokaryotes. • Highly reliable.
	Gene neighboring	• Mostly utilized for small scale proteomes. • Relatively straightforward. • Prone to producing false negatives. • Results dependent on used genome number and distribution.
	Phylogenetic similarity	• Requires a complete genome. • Results dependent on used genome number and distribution. • Does not apply to essential proteins.
Methods based on machine learning algorithms with the utilization of multiple genomic/proteomic features	Sequence and primary structure	• Relatively straightforward. • May be used for large scale proteomes. • Needs the interpretation of importance features.
	Structure based	• Tends to be rather limited when it comes to scale. • Enables a detailed PPI analysis.
	Decision tree and random forest	• Confronts well with high-dimensional data. • Confronts well with missing values. • The data pattern can be easily explained.
	KNN	• Straightforward. • Does not require any kind of training. • The memory requirement and cost of computation quickly increase with the increasing vector dimension feature.
	MLR	• Solid possibilities for generalization. • Resembles a black box.
	Naïve Bays	• Founded on the independence assumption for the explored features. • Straightforward and not difficult for interpretation. • Easily tackles the missing values.
Other methods	Text mining methods	• Results may be affected by network completeness and false positives. • Use of network topology for predicting the protein-protein interaction. • Results may not be as reliable as in the case of manually curated data. Neverthless, the rapid development of published biomedical literature can lead to these methods being better grounded.

- KNN – K-nearest neighbors algorithm
- MLR – Multiple linear regression

vant information with regard to the interactions studied. Over the past several decades, a number of computational approaches have been devised for the interaction discovery of protein-protein. Such methods differ in the type of information utilized for PPI prediction (1 - 4). Several databases have been developed for the purpose of holding and retaining large amounts of data concerning the PPIs of numerous organisms, most of which are available publicly (5). The mentioned databases have been grounded in the novel and rapid high-throughput technological advances. These may represent a major source of data for the evaluation of prediction methods. At present, over 100 repositories with regard to PPIs have been uploaded to the web and are readily available for reference purposes online. The most significant of these include BioGRID (the General Repository for Interaction Database, representing one of the most comprehensive databases for protein-protein interactions established through experimentation), DIPTM (which stands for the Database of Interacting Proteins, devised at the University of California, Los Angeles, containing a multitude of varying source data, aimed at forming a unique and consistent PPI), MINT (the Molecular Interaction Database), BIND (the Biomolecular Interaction Network Database), HPRD (the Human Protein Reference Database), as well as IntAct.

Overall, the computational methods available for the prediction of protein-protein interactions may be classified into four major categories: methods utilizing network topology for predicting protein-protein interaction, structural information and genomic context methods, methods involving literature and text mining (or database searching facilities) for the detection of protein-protein interaction, as well as the methods which employ the use of machine learning algorithms working with heterogeneous genomic/proteomic characteristics. Table 1 presents an overview of the methods in question.

Genomic context-based methods and structure information headings

Gene neighboring. Considering the genomic context and the notion that the genome contains related genes in proximity to one another, one of the first and simplest methods, the co-localization of genes or gene neighboring for PPI prediction was developed. The elevation in the numbers of genomes has made this method more reliable, similar to other genome context approaches. Simplicity is the most important feature of this method. However, some false negative results could be obtained with the application of this method, since it lacks recognition of the manner in which distantly located genes interact with each other. A second flaw of the gene neighboring method is that its performance is influenced by the reference genome choice (6 - 9).

Phylogenetic relationship. The main principle on which this method detects PPIs is the "phylogenic profile" similarity. The binary vector reflecting the existence or lack of a studied protein within a set of organisms represents the phylogenetic profile of a certain protein. According to the mentioned approach, the Phylogenetic Relationship (PR) method can be considered as a sort of gene neighboring method which is more flexible and more superior since it is able to uncover the kind of interaction which cannot be discovered with the gene neighboring method. The chief notion on which the PR method develops conclusions is that the genes which are functionally related can be found together across a multitude of distant species, and that they play the same biological process role. Unfortunately, there are three crucial flaws to the PR method: 1) the used genome distribution and numbers have a dramatic influence on the obtained results; 2) the PR method does not apply to essential proteins, present in nearly every organism; 3) the PR method can be applied only to complete genomes (10).

Gene fusion. The main principle used in the gene fusion (GF) method is the application of comparative genomics and evolutionary information. For this reason, the GF method can be regarded as supplemental to the methods which involve the phylogenetic profile and gene neighboring (11, 12). Since the GF method is based on the occurrences of fusion in the existing genes, the obtained results provide relevant information of reliability and a functional relationship. The GF method's major drawback is the lack of fusion events, particularly in prokaryotes (13).

3D structure-based

This method is based on the information about the 3D structure of the studied proteins used for the purpose of predicting their interactions. Accurate 3D structures of studied proteins are necessary to obtain the valid results. In recent years, a genome-wide scale method was introduced. Its basis was the

structure information for predicting PPI with the use of homology models. When compared with other types of methods, this method yields results with a greater number of details, such as defining the biophysical traits of the mentioned interaction, as well as the interacting residue (14, 15). In addition, the obtained information about PPIs can be used to predict the interaction between the new proteins equiavalent to the interacting proteins previously predicted.

Topology network-based methods

PPIs in various organisms have common topological features like many real-world networks, rendering them dissimilar to random networks. Such topological features are made use of to detect true positives and false positives between PPIs. If one is to understand the network dynamics and underlying evolutionary mechanisms shaping the network in a better way, it is essential to consider the topological perspective during PPI network analysis. To determine how significant topological properties are within a specific PPI network, a comparison needs to be made between the topological features and random network features. Afterwards, the PPIs are to be assigned confidence scores (16 - 18). At the end of the process, certain interactions may be eliminated on the basis of the obtained scores, while there is a possibility of adding others to the network (19).

Methods involving both literature and text mining

The significance of biomedical literature mining approaches is related to the fact that the PubMed database is being augmented at an incredible rate, with two papers published every minute. For PPI prediction, some computational methods employ algorithms for literature and text mining to obtain data related to the co-occurrence of the proteins cited in PubMed abstracts. The literature mining approach is threefold: 1) Named Entity Recognition (NER) is a recognition step in which the name of the studied protein is defined, and this step is of extreme importance for conducting further analyses; 2) Zoning, a step involving splitting the text into basic constituents and extracting sentences from the text; 3) PPI extraction, a step that requires the use of different algorithms for the determination of protein-protein interaction. There are three categories of mining approaches employed in biomedical literature required to detect protein-protein interactions in current practice: 1) Computational natural language processing (NLP) and methods grounded in linguistics, by which protein-protein interaction is detected through the use of parsers and grammar definitions; 2) Methods which are rule-based, in which the deductions of protein-protein interaction are made by applying patterns or a set of context-specific rules; 3) Machine learning approaches involve the use of classifiers to learn the pattern that enables the identification of PPI from the training set (19 - 22). Although the results which automated data mining yields may be less reliable than manually curated data, these methods can be made more reliable by the rapid development of the published biomedical literature (23).

Machine learning algorithm-based methods with the utilization of heterogeneous genomic/proteomic features

Heterogeneous biological data, like gene expression, k-let count (length k subsequences) codon usage, and amino acids' physicochemical properties are utilized in a number of methods for developing a PPI prediction model. For the purpose of learning about and predicting PPIs, the mentioned methods involve the integration of biological data sources provided by high-throughput technologies to feature vectors and apply machine learning techniques. For the most part, a classifier, i.e. machine learning algorithm used for the prediction of protein-protein interaction includes numerous descriptors or features of the proteins or their pairs with known interactions and non-interactions, in the form of a learning set for establishing which proteins do or do not interact. After the model is established, new protein pairs can be classified by this algorithm to interacting classes or non-interacting classes (24 - 26). Computational biology and bioinformatics rely on the wide use of support vector machines (SVM) or kernel machines for the classification of protein-protein interaction, along with the classification of other sorts of biological data. The idea of margin maximization is the key concept for the development of the SVM classifier. The certainty of an object's classification is in direct relation to the object margin. For this reason, objects which have a correctly assigned label and are highly certain will have large margins. Conversely, small margins are found in

objects with uncertain classification. The SVM method establishes a training model with the application of a dataset for labeled training. This entire set is designated so as to belong to one of the two classes, so the developed model is able to make predictions regarding class labels for new cases. The positive feature of SVM is its extremely powerful classification algorithm that can be used with arbitrary complexity. Unfortunately, the SMV algorithm is quite complex and demands large computational memory, which leads to rather slow training and evaluation. What is more, the initial parameters can have a big impact on the results obtained with this classifier (27, 28). Artificial neural networks (ANNs or NNs) have been envisaged to mathematically model the intellectual abilities of humans by employing designs which are plausible in biological terms. The multilayer perception (MLP) is among the most popular NN models. It represents a tool for PPI modeling which has demonstrated admirable performance levels. Nevertheless, MLP has received a lot of criticism and has been considered a black-box classifier owing to the difficulty in establishing the actual model parameter meaning (29, 30). As a feed-forward artificial neural network, MLR has multiple layers, each of which is fully linked to the following layer with the use of weighted edges. In general, three layers comprise the MLR method: the input layer, the hidden (intermediate) layer and the output layer. These layers' main topology can be summarized since every node at the output layer, which is hidden, is a neuron that has an activation function, and the processing units of an MLP are contained within the mentioned node. A probabilistic classifier based on Bayes' theorem, Naïve Bayes is quite simple, computationally efficient, and easy to interpret, making it very popular. The main source of Naïve Bayes' simplicity lies in the assumption that independent variables are statistically independent. Naïve Bayes can be used quite adequately when tackling problems concerning normal distributions, which tend to be increasingly common in real-world problems. For a small training dataset, Naive Bayes classifiers can be efficiently trained for achieving a learning approach under supervision. On the other hand, when it comes to more complex problems of classification, the above mentioned may prove to be inadequate. Despite the mentioned fact, the method has seen extensive use in the prediction problem of PPIs (31). One of the most straightforward machine learning classifiers is K-Nearest Neighbors (K-NN). In this method, the approach used for classifying objects is based on label assignation to each separate object on the basis of the K closest objects (the user is the one who sets the K parameter). No explicit training is required to use K-NN (K optimization can be considered as a type of learning owing to the relevance of the choice of K in this method). The memory requirements and computational costs increase quickly in the event that numerous features or large data sets are present. For this reason, the method has not been used extensively in PPI prediction (32, 33). The random forest (RF) algorithm represents a method of classification involving various decision trees. Each of these is built on random feature vectors which have been independently sampled from a single data set during the training phase. Then, a small fraction of the variables is selected randomly for every node in a tree, after which each classification tree is fully developed. Following this, the input vector is placed down in every tree of the forest for the purpose of classifying a new object. Finally, one class is assigned to the object according to majority voting. Once a considerable number of features or a big dataset are used, leaving no need for feature deletion or selection, RF becomes a practical classifier. Moreover, RF is able to rank features on the basis of classification relevance. RF can also be of use in the missing data recovery. Owing to the unique features of the random forest and decision trees, they are frequently utilized in computational biology and bioinformatics for the classification of biological data, especially those required for the prediction of PPIs (34, 35).

DOCKING AND SCORING – METHODS AND APPLICATIONS

In situations when the chemical compositions of both the target enzyme (receptor) and the studied small molecule (ligand) are known as molecular docking (MD), the frequnetly utilized *in silico* method may be of use when defining the most likely geometry of the ligand within the active site of the receptor. Furthermore, MD studies can be used to calculate binding energies between amino acids and ligands, and from the active site of the receptor. Thus, these values can be correlated to "scoring functions". Since the studied compounds' inhibitory effect may be correlated to such interactions, these methods have high applicability in biochemistry-related research, especially in pharmacology and

drug design (36-38). In brief, in the hit identification and lead optimization of drug candidates, small molecules are 'docked' into macromolecular target structures, and they are potential complementarily to the active sites 'scored' with the application of "scoring functions" which are based on energy calculations. Most MD studies are aimed at performing two tasks: precise structural modeling and accurate activity prediction. Identifying the molecular features which account for certain biological activities or the future modification predictions regarding the studied ligands, which will enhance their inhibitory potential, is often a highly complex issue and, therefore, challenging to simulate with the use of a computer. To overcome this problem, MD is derived from several multistep processes, where every single step presents one or more additional degrees of complexity (39, 40).

The first step in the MD studies is applying docking algorithms to "pose" as small molecules in the active site of the receptor. Since a significant number of conformational degrees of freedom may be contained even in relatively simple organic molecules, this step is challenging in itself. "Scoring" functions have been developed for the purpose of predicting the biological activity by evaluating the interactions between amino acids and ligands inside a receptor's active site. During the early stages of MD studies, relatively simple scoring functions are applied, and after the most preferable conformers are determined, more complex scoring schemes are used for further evaluation. These "scoring" functions can include electrostatic interactions, van der Waals interactions, as well as the inclusion of at least a certain degree of salvation or entropic effects, since a mixture of enthalpic and entropic effects drives ligand-binding events, and either enthalpy or entropy can be dominant in certain interactions. The representation of both receptors and ligands must be considered for the evaluation of various docking methods. In the current practice, the basic representations of the receptor are threefold: grid, surface, and atomic representations. Only during final ranking procedures is atomic representation used in the combination involving a potential energy function. In protein-protein docking studies, surface-based docking programs are the ones predominantly employed in the *in silico* approach. The main attempt of this method is the alignment of points on surfaces through the minimization of the angle between opposing molecule surfaces, and for this reason, the standard for a great number of techniques for protein-protein docking is still a rigid body approximation. The fundamental idea of grid representation is the storage of information regarding the energetic contributions of the receptor on the grid points, due to which it is only to be read during ligand scoring, where electrostatic and van der Waals are the two main energetic potentials stored by grid points.

The next step in MD studies is related to search methods and molecular flexibility. Ligand flexibility is treated within one of the three fundamental categories: random methods (genetic algorithms, Monte Carlo); simulation methods (energy minimization, molecular dynamics), as well as systematic methods (databases, conformational search, incremental construction). Protein flexibility treatment is considerably less sophisticated in comparison to that of ligand flexibility, though a number of approaches have been flexibly applied to at least one part of the target model, including Monte Carlo calculations, protein ensemble grids, molecular dynamics and rotamer libraries. The final step in MD studies is evaluating and ranking the predicted ligand conformations. This step is essential in the virtual screening based on structure. For this reason, the design of reliable schemes and scoring functions is of utmost importance. Free-energy simulation techniques have been developed as some of the most sophisticated methods for the quantitative modeling of protein-ligand interactions, as well as the binding affinity prediction. Nonetheless, these expensive calculations are sometimes inaccurate and impractical for evaluating greater numbers of protein-ligand complexes. Three kinds or scoring function classes are presently utilized to overcome this issue: knowledge-based, empirical and force-field-based scoring functions. These "scoring" functions take into account numerous simplifications and assumptions during the assessment of modeled complexes, and do not present a full account of many physical phenomena dictating molecular recognition.

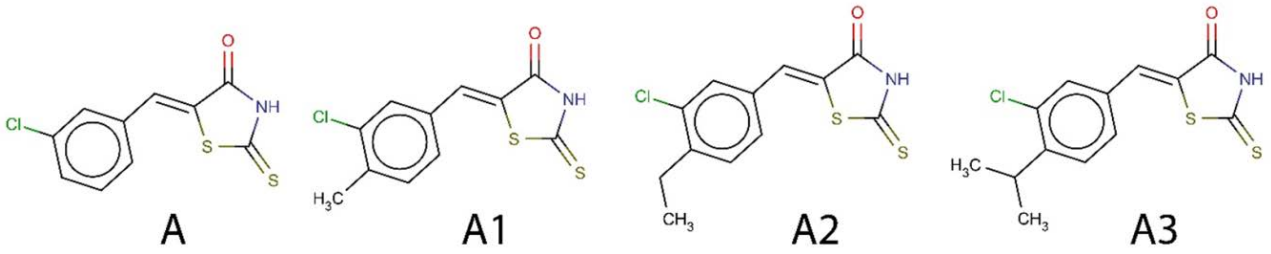

Figure 1. *The chemical structures of the examined compounds*

Table 2. *Score values (kcal/mol) of all compounds designed with the aid of a computer*

Molecule	MolDock Score	Rerank Score	Steric	VdW	HBond	NoHBond	Energy
A	-106.561	-88.0738	-108.013	-28.8384	-3.48477	-3.82655	-108.344
A1	-111.365	-93.2288	-114.508	-31.6348	-2.5	-2.5	-113.699
A2	-121.833	-97.8433	-120.089	-35.2613	-0.27274	-1.79198	-119.881
A3	-128.117	-101.857	-127.337	-32.6603	-0.54586	-1.43879	-128.642

At present, many types of software can be used successfully for MD studies, both free and license based. Molegro Virtual Docker software (hereinafter: MVD) is one of them, and it has unique features in comparison to others. As all MD softwares, MVD can be used for attaining relevant geometrical orientation of the ligand inside the active site of the enzyme being studied. What is more, MVD can be used for the determination of hydrogen bonds, in addition to the hydrophobic interactions existing between the rigid amino acids from the active site of the enzyme and flexible ligands. Finally, MVD can be used to calculate the so-called "scoring" functions, i.e. adequate binding energies (41). The most important "scoring" functions that can be calculated with MVD include Rerank Score, VdW, Hbond, NoHbond, Pose energy, MolDock and Steric. NoHbond and HBond are used for the calculation of no hydrogen and hydrogen bond interactions, respectively; energies from Steric and Van der Walls interactions are, in turn, calculated with the use of Steric and VdW "scoring" functions. The total energy of the best-calculated pose is determined with Pose energy. Finally, the Rerank Score and MolDock Score are final estimators of amino acids and ligands from the interaction energy of the enzyme's active site. Once calculated, the "scoring" functions may be employed to assess the inhibitory effect of the examined compounds by making a comparison of their interaction energies (42 - 46).

In the presented paper, as an example of the MD method, MVD was applied to a small set of molecules acting as peroxisome proliferator-activated receptor (PPARγ) agonists. The Marvin sketch (Marvin 6.1.0, 2013, ChemAxon) was made use of for drawing the examined molecules, while the MMFF94 force field was the tool required for securing the optimal 3D geometry of the mentioned. The protein data base (PDB: 5TWO) was the source for the Peroxisome proliferator-activated receptor (PPARγ). In addition, the following "scoring" functions were calculated and taken into account while assessing the inhibitory activity of the molecules designed with the aid of a computer, since different interactions between amino acids from the active site and ligands relate to different scoring functions. Still, all these need to be considered: Rerank Score, MolDock Score, NoHBond Score, Hbond Score, Energy, VdW and Steric. Structures of the studied compounds are shown in Figure 1. The calculated values for the selected "scoring" functions are cited in Table 2.

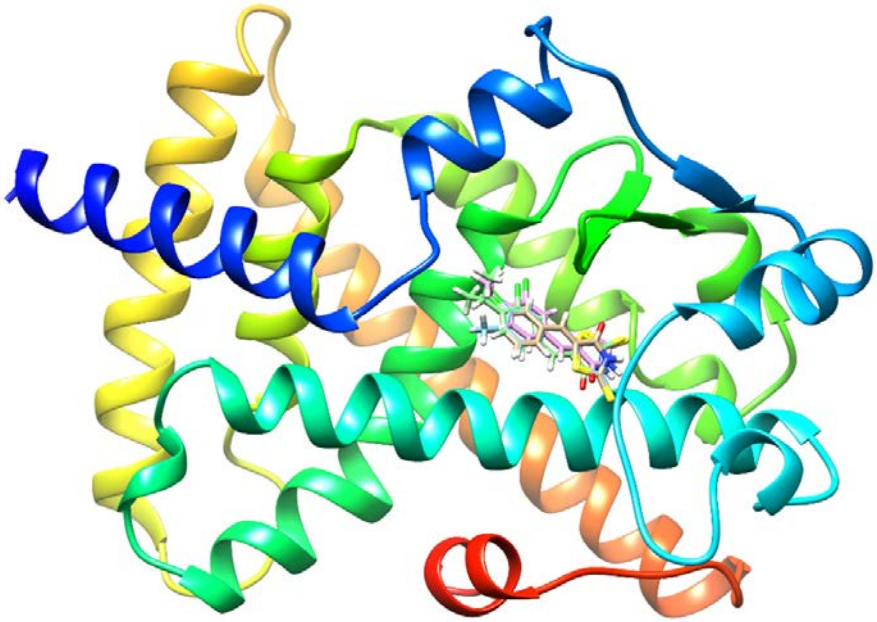

Figure 2. *The best calculated poses for all the compounds studied within the active site of PPARγ*

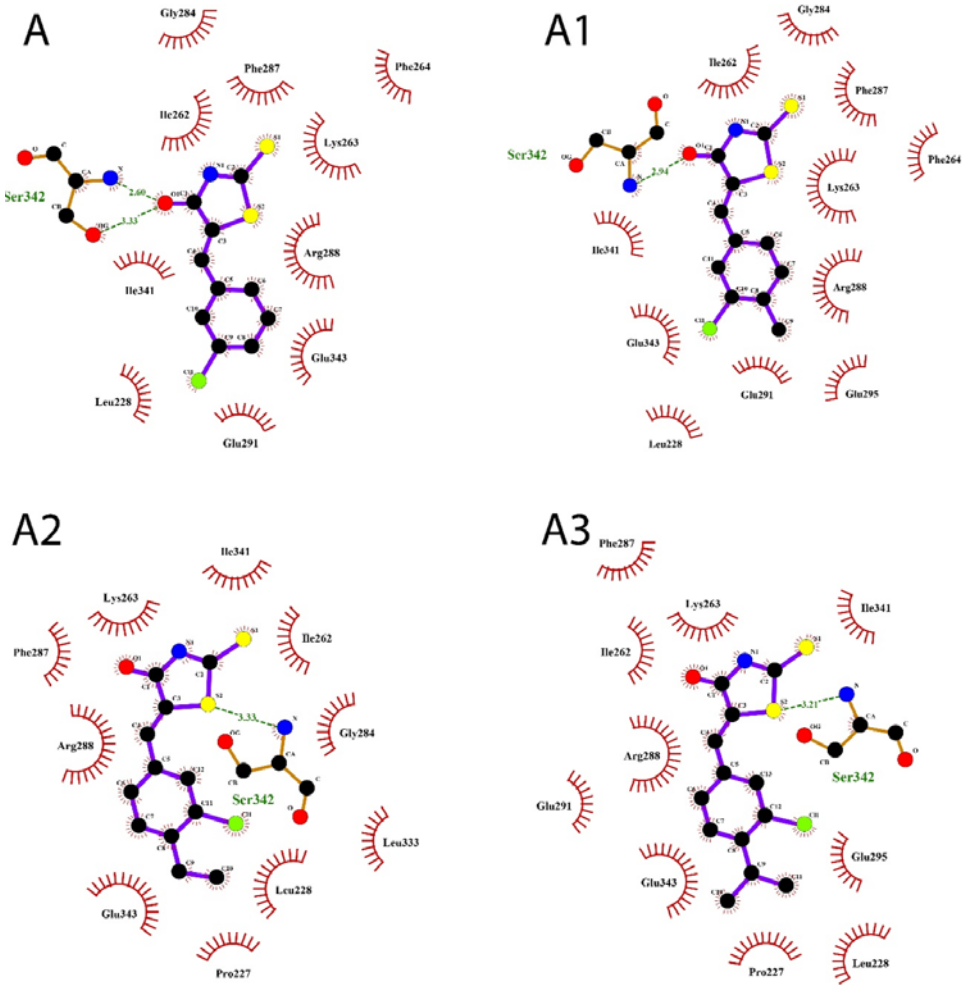

Figure 3. *A two-dimensional representation of the interaction between the evaluated molecules and amino acids within the PPARγ binding pocket*

Based on the results yielded by the Rerank and MolDock "scoring" function values, molecule A3 displays the greatest binding potential. According to Steric score function values, the A3 molecule exhibits the highest energy levels from steric interactions. Molecule A2 shows the highest energies from the Van der Walls interactions defined by the appropriate value obtained from VdW. The obtained results indicate that molecule A has the highest interactions concerning hydrogen bonds formed between receptor and ligand. The "scoring" function that summarizes all relevant ligand energies is Energy, and according to the obtained values, molecule A2 has the most preferable interaction in comparison to other studied molecules. Figure 2 presents the foremost calculated poses of all the compounds studied within the active site of PPARγ. A two-dimensional representation of the interactions between the studied compounds and PPARγ are presented in Figure 3.

CONCLUDING REMARKS

Protein-protein interaction network studies can be categorized in two manners: methods based on the mathematical and statistical modeling, as well as the models established in comparative network analysis. Statistical and mathematical modeling applied in protein-protein interaction network studies initially analyzes the network, with the focus on topological features. This step is followed by the production of the statistical models for evolving networks. Finally, the tuning of initial parameters is performed for the purpose of reproducing the properties noted in the experimentally developed networks. An analysis of the species with different complexity levels related to the protein-protein interaction networks is performed in approaches which have been based on the comparative network, followed by the network which is comparative for the future evolution of generated networks. Three main evolutionary events are considered in both approaches. These events are included in the crucial protein-protein interaction network shaping processes. Pharmaceutical research has applied molecular docking calculations over nearly two decades. Virtual screening and molecular docking provide an opportunity for a de novo identification of active compounds, not leaning towards the established leads or hits. There is great diversity among the developed algorithms for scoring methodologies and contemporary posing. The correlation between molecular docking and the obtained "scoring" functions seems to be rather complex, though the produced models with bound ligands have proven to be very reliable. Furthermore, the developments of novel scoring functions will not necessarily lead to the further development of schemes that deal with com-+pound ranking and scoring. The combination of molecular docking and compound filter functions, pharmacophore models, as well as of the two-dimensional similarity-based methods, or three-dimensional ones, has been successfully applied in biochemistry related research. One of the main features of this combination is the reduction in the figure of candidate compounds required for particularly complex calculations for scoring. Even though scoring and docking rely on a lot of approximations, during application, the mentioned techniques improved molecular optimization, frequently combining other computational methods to extend the already applied traditional approaches for a structure-based design.

Acknowledgements

The Ministry of Education and Science of the Republic of Serbia supported this study (Project Number 172044).

References

1. Lambrinidis G, Vallianatou T, Tsantili-Kakoulidou A. *In vitro, in silico* and integrated strategies for the estimation of plasma protein binding. A review. Adv Drug Deliver Rev 2015;86:27-35. https://doi.org/10.1016/j.addr.2015.03.011

2. Cicaloni V, Trezza A, Pettini F, Spiga O. Applications of *in silico* methods for design and development of drugs targeting protein-protein interactions. Curr Top Med Chem 2019;19(7):534-44. https://doi.org/10.2174/1568026619666190304153901

3. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. Plos Comput Biol 2007;3(3):e42. https://doi.org/10.1371/journal.pcbi.0030042

4. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part II. Experimental techniques and databases. Computational methods to predict protein and domain interaction partners. Plos Comput Biol 2007;3(4):e43. https://doi.org/10.1371/journal.pcbi.0030043

5. Orchard S, Kerrien S, Abbani S, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 2012;9(4):345-50. https://doi.org/10.1038/nmeth.1931

6. Zahiri J, Bozorgmehr JH, Masoudi-Nejad A. Computational Prediction of Protein-Protein Interaction Networks: Algorithms and Resources. Curr Genomics 2013;14(6):397-414. https://doi.org/10.2174/1389202911314060004

7. Mercatelli D, Scalambra L, Triboli L, et al. Gene regulatory network inference resources: A practical overview. Biochim Biophys Acta Gene Regul Mech 2020;1863(6):194430. https://doi.org/10.1016/j.bbagrm.2019.194430

8. Li X, Li W, Zeng M, et al. Network-based methods for predicting essential genes or proteins: A survey. Brief Bioinform 2020;21(2):566-83. https://doi.org/10.1093/bib/bbz017

9. Koutsoukas A, Simms B, Kirchmair J, et al. From *in silico* target prediction to multi-target drug design: Current databases, methods and applications. J Proteomics 2011;74(12):2554-74. https://doi.org/10.1016/j.jprot.2011.05.011

10. Juan D, Pazos F, Valencia A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci USA 2008;105(3):934-49. https://doi.org/10.1073/pnas.0709671105

11. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struc Biol 2001;12(3):368-73. https://doi.org/10.1016/S0959-440X(02)00333-0

12. Skrabanek L, Saini HK, Bader GD, Enright AJ. Computational prediction of protein-protein interactions. Mol Biotechnol 2008;38(1):1-17. https://doi.org/10.1007/s12033-007-0069-2

13. Enright A, Iliopoulos I, Kyrpides N, Ouzounis C. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402(6757):86-90. https://doi.org/10.1038/47056

14. 14. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 2002;99(9):5896-901. https://doi.org/10.1073/pnas.092147999

15. Hue M, Riffle M, Vert JP, Noble WS. Large-scale prediction of protein-protein interactions from structures. BMC Bioinformatics 2010;11(1):144. https://doi.org/10.1186/1471-2105-11-144

16. Li M, Lu Y, Wang J, et al. A topology potential-based method for identifying essential proteins from PPI networks. IEEE ACM T Comput Biol 2015;12(2):372-83. https://doi.org/10.1109/TCBB.2014.2361350

17. Lei X, Yang X, Fujita H. Random walk based method to identify essential proteins by integrating network topology and biological characteristics. Knowl-Based Syst 2019;167:53-67. https://doi.org/10.1016/j.knosys.2019.01.012

18. Luo J, Kuang L. A new method for predicting essential proteins based on dynamic network topology and complex information. Comput Biol Chem 2104;52:e34-e42. https://doi.org/10.1016/j.compbiolchem.2014.08.022

19. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci USA 2003;100(8):4372-6. https://doi.org/10.1073/pnas.0735871100

20. Tikk D, Thomas P, Palaga P, et al. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. PLoS Comput Biol 2010;6(7):Article number e1000837 https://doi.org/10.1371/journal.pcbi.1000837

21. Bui Q-C, Katrenko S, Sloot PMA. A hybrid approach to extract protein-protein interactions. Bioinformatics 2011;27(2):259-65. https://doi.org/10.1093/bioinformatics/btq620

22. He M, Wang Y, Li W. PPI finder: A mining tool for human protein-protein interactions. PLoS ONE 2009;4(2):Article number e4554 https://doi.org/10.1371/journal.pone.0004554

23. Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D. Integrating protein-protein interactions and text mining for protein function prediction. BMC Bioinformatics 2008;9(Suppl 8):S2. https://doi.org/10.1186/1471-2105-9-S8-S2

24. Shen J, et al. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci USA 2007;104(11):4337-41. https://doi.org/10.1073/pnas.0607879104

25. Ben Hur A, Ong CS, Sonnenburg S, et al. Support Vector Machines and Kernels for Computational Biology. PLoS Comput Biol 2008;4(10):e1000173. https://doi.org/10.1371/journal.pcbi.1000173

26. Rashid M, Ramasamy S, PS Raghava G. A simple approach for predicting protein-protein interactions. Curr Pro Pept Sci 2010;11(7):589-600. https://doi.org/10.2174/138920310794109120

27. Chatterjee P, Basu S, Kundu M, et al. PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. Cell Mol Biol Lett 2011;16(2):264-78. https://doi.org/10.2478/s11658-011-0008-x

28. Zhang M, Su Q, Lu Y, et al. Application of machine learning approaches for protein-protein interactions prediction. Med Chem 2017;13(6):506-14. https://doi.org/10.2174/1573406413666170522150940

29. Zhang L, Yu G, Xia D, Wang J. Protein-protein interactions prediction based on ensemble deep neural networks. Neurocomputing 2019;324:10-9. https://doi.org/10.1016/j.neucom.2018.02.097

30. Li H, Gong X-J, Yu H, Zhou C. Deep neural network based predictions of protein interactions using primary sequences. Molecules 2018;23(8): Article number 1923 https://doi.org/10.3390/molecules23081923

31. Lin X, Chen X-W. Heterogeneous data integration by tree-augmented naïve Bayes for protein-protein interactions prediction. Proteomics 2013;13(2):261-8. https://doi.org/10.1002/pmic.201200326

32. Thuy Phan TT, Ohkawa T. Protein-protein interaction extraction with feature selection by evaluating contribution levels of groups consisting of related features. BMC Bioinformatics 2016;17:Article number 246 https://doi.org/10.1186/s12859-016-1100-z

33. Marini S, Xu Q, Yang Q. *In silico* protein-protein interaction prediction with sequence alignment and classifier stacking. Curr Protein Pept Sci 2011;12(7):614-20. https://doi.org/10.2174/1389203711109070614

34. Jia J, Xiao X, Liu B. Prediction of Protein-Protein Interactions with Physicochemical Descriptors and Wavelet Transform via Random Forests. JALA-J Lab Autom 2016;21(3):368-77. https://doi.org/10.1177/2211068215581487

35. Liu W, Guo Y, Luo J, et al. Prediction of kinase-specific phosphorylational interactions using random forest. Chemometr Intell Lab 2013;126:117-22. https://doi.org/10.1016/j.chemolab.2013.05.005

36. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002;47(4):409-43. https://doi.org/10.1002/prot.10115

37. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 2004;3(11):935-49. https://doi.org/10.1038/nrd1549

38. Veselinović JB, Kocić GM, Pavic A, et al. Selected 4-phenyl hydroxycoumarins: *in vitro* cytotoxicity, teratogenic effect on zebrafish (Danio rerio) embryos and molecular docking study. Chem Biol Interact 2015;231:167-74. https://doi.org/10.1016/j.cbi.2015.02.011

39. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Eng 2002; 41(15):2644-76. https://doi.org/10.1002/1521-3773(20020802)41:15<2644::AID-ANIE2644>3.0.CO;2-O

40. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Annu Rev Biophys Biolmol Struct 2003;32:335-73. https://doi.org/10.1146/annurev.biophys.32.110601.142532

41. Thomsen R, Christensen MH. MolDock: a new technique for high-accuracy molecular docking. J Med Chem 2006;49(11):3315-21. https://doi.org/10.1021/jm051197e

42. Ničković VP, Mitić NR, Krdžić BD, et al. Design and development of novel therapeutics for brucellosis treatment based on carbonic anhydrase inhibition. J Biomol Struct Dyn 2020;38(6):1848-57. https://doi.org/10.1080/07391102.2019.1619626

43. Ćirić Zdravković S, Pavlović M, Apostlović S, et al. Development and design of novel cardiovascular therapeutics based on Rho kinase inhibition - *In silico* approach. Comput Biol Chem 2019:79:55-62. https://doi.org/10.1016/j.compbiolchem.2019.01.007

44. Stoičkov V, Šarić S, Golubović M, et al. Development of non-peptide ACE inhibitors as novel and potent cardiovascular therapeutics: An *in silico* modelling approach. SAR QSAR Environ Res 2018;29(7):503-15. https://doi.org/10.1080/1062936X.2018.1485737

45. Zivkovic M, Zlatanovic M, Zlatanovic N, et al. Development of novel therapeutics for the treatment of glaucoma based on actin-binding kinases inhibition - *In silico* approach. New J Chem 2020;44:6923-31. https://doi.org/10.1039/C9NJ05967A

46. Zivkovic M, Zlatanovic M, Zlatanovic N, et al. The Application of the Combination of Monte Carlo Optimization Method based QSAR Modeling and Molecular Docking in Drug Design and Development. Mini-Rev Med Chem 2020;20(14):1389-402. https://doi.org/10.2174/1389557520666200212111428

Mreža protein–protein interakcija i protein-ligand doking – trenutno stanje i perspektive

Aleksandar Velesinović, Goran Nikolić

Univerzitet u Nišu, Medicinski fakultet, Departman za hemiju, Niš, Srbija

SAŽETAK

Tradiconalna istraživanja bazirana na *in vivo* i *in vitro* modelima dosledno se koriste za testiranje biohemijskih hipoteza. U poslednjoj deceniji sve se više razvijaju i računarske (*in silico*) metode za razvoj i testiranje hipoteza o biohemijiskim istraživanjima. *In silico* modeli imaju za cilj da analiziraju kvantitativne aspekte naučnih (velikih) podataka, koji se ili čuvaju u velikim bazama podataka ili generišu sofisticiranim alatima za modeliranje i simulaciju; da steknu osnovno razumevanje različitih biohemijskih procesa, koji se naročito odnose na velike biološke makromolekule, primenom računarskih metoda na velikim skupovima podataka i računanjem ponašanja bioloških sistema. Računarske metode, koje se koriste u biohemijskim istraživanjima, uključuju mapiranje interakcije proteina i DNK na čitavom genomu, bioinformatiku zasnovanu na proteomici i mapiranje mreža interakcija protein–protein sa visokim propusnim opsegom. Neke od široko korišćenih tehnika molekularnog modeliranja i simulacija su molekularna dinamika, Monte Carlo i Langevinova (stohastička, Brovnijeva) dinamika, kontinuirana elektrostatika, statistička termodinamika, tehnike modeliranja proteina, vezivanje proteina i liganda, izračunavanje afiniteta proteina i liganda i računarska simulacija procesa sakupljanja proteina i delovanje enzima. Ovaj rad predstavlja kratak pregled dve važne metode koje se koriste u studijama biohemije–predviđanje mreža interakcija protein–protein i vezivanje proteina–liganda.

Ključne reči: in silico, **protein–protein interakcije, protein-ligand doking, molekularno modelovanje**