33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

# Drug Design For Cardiovascular Disease: The Effect Of Solvation Energy On Rac1-Ligand Interactions

Norbert Maggi, *Student Member, IEEE*, Patrizio Arrigo, and Carmelina Ruggiero, *Member, IEEE*

***Abstract*—'OMICS' techniques have deeply changed the drug discovery process. The availability of many different potential druggable genes, generated by these new techniques, have exploited the complexity of new lead compounds screening. 'Virtual screening', based on the integration of different analytical tools on high performance hardware platforms, has speeded up the search for new chemical entities suitable for experimental validation. Docking is a key step in the screening process. The aim of this paper is the evaluation of binding differences due to solvation. We have compared two commonly used software, one of which takes into account solvation, on a set of small molecules (Morpholines, flavonoids and imidazoles) which are able to target the RAC1 protein - a cardiovascular target. We have evaluated the degree of agreement between the two different programs using a machine learning approach combined with statistical test. Our analysis, on a sample of small molecules, has pointed out that 35% of the molecules seem to be sensitive to solvation. This result, even though quite preliminary, stresses the need to combine different algorithms to obtain a more reliable filtered set of ligands.**

## I. Introduction

THE investigation of the molecular mechanism of the diseases (Molecular Medicine) is one of the major challenges in the post-genome era. From a pharmacological point of view the identification of genes, involved in a specific pathology, (markers) is an essential step. In addition, is it also important to elucidate the functional interactions among them. From a molecular point of view a disease is characterized by altered genomic and proteomic profiles [1]. Statistical analysis allows to determine the set of critical gene or proteins in a panel of samples. 'Omics' has heavily impacted the drug discovery process. The high throughput screening method, developed during the sequencing projects, has increased the capability to screen, in a parallel way, thousand potential targets. Despite the availability of all these new technologies, the productivity, as new potential drugs, of pharmaceutical companies is declined.

The identification of new chemical compounds suitable for high throughput pharmacological screening has become computationally intensive. The chemoinformatic screening is also limited by data accessibility because many compounds are under copyright and there not information about them. In the last decade the Computer Aided Drug Design (CADD) has increase its complexity including genetic information in the screening process (pharmacogenetics). In the past the role of genetic variability was not considered, now the effect of genetic variability on drug efficacy is a well-established knowledge. In many cases different proteins forms, associated to a slight variation at genomic level, can be targeted by a drug in a more or less efficient way. On the other site a drug can induce differential gene expressions. The evaluation of genomic modification, triggered by drugs, is the aim of pharmacogenomics. It is quite obvious that, for virtual screening, pharmacogenetic knowledge has a critical influence. The identification of new drug targets requires the integration of bioinformatics and chemoinformatics with experimental results. These data are the knowledge background for the QSAR (Quantitative Structure-activity Relationship) modelling. At higher level the final step of the discovery process is related to the capability to evaluate the efficacy by the ADME (Adsorbtion Diffusion Metabolism and Excretion) simulation. In order to improve the performance of target identification and design of new drugs, bioinformatics and chemoinformatics tools have to be integrated.

The use of data mining techniques related to gene expression level during FP6 Cardioworkbench project (www.cardioworkbench.eu) has highlighted the role of Rac1 protein in cardiovascular diseases [2]. In particular Rac1 protein is implicated in several events of atherosclerotic plaque development and represents a new potential pharmacological target for cardiovascular diseases. Rac1 is member of the Rho family. This protein family serve as conformational switches in a wide variety of signal transduction pathways. It is a key regulators of angiogenesis, modulating a diversity of cellular processes, including vascular permeability, extracellular matrix remodeling, migration, proliferation, morphogenesis, and survival [3]. Rac1 plays a central role in endothelial cell migration, tubulogenesis, adhesion, and permeability in response to vascular endothelial growth factor (VEGF) and sphingosine-1-phosphate (S1P), which is likely due to the inability of Rac1-deficient endothelial cells to form lamellipodial structures and focal adhesions, and to remodel their cell-cell contacts [4]. It is been demonstrate that the activation of Rac1, but not RhoA, in human aortic smooth muscle cells (SMCs) through the engagement of $\alpha 2\beta 1$ integrin by type I collagen induces the expression of matrix metalloproteinase 1 (MMP1) and MMP2, an event that may contribute to atherosclerotic plaque rupture [2].

Manuscript received April 15, 2011. Patrizio Arrigo is with the CNR Institute of Macromolecular Studies (ISMAC), Section of Genoa, Via De Marini 6, 16149 Genoa, Italy (E-mail: arrigo@ge.ismac.cnr.it). All other authors are with the Department of Communication, Computer and System Sciences, Nanobiotechnology and Medical Informatics Laboratory University of Genoa, Via all'Opera Pia, 13, 16145 Genoa, Italy (*corresponding author, phone: +390103532991; fax: +390103532154; E-mail: carmel@dist.unige.it).

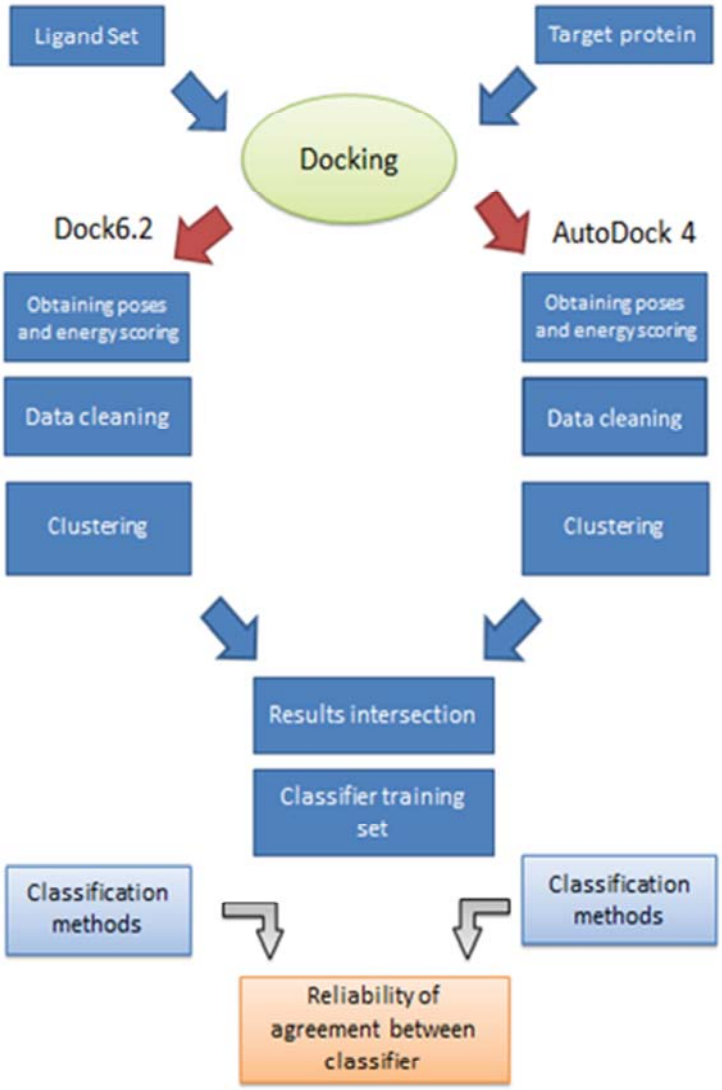


Fig. 1. Workflow for the identification of solvation sensitive molecules

We propose a new method to identify a set of putative ligands that could be more sensitive to the water solvation effect due to the solvation forces based on the disagreement of the result of two classifiers. The contribution of solvation energy is critical for functional aspect of macromolecular interaction, for instance solvation is essential for enzyme kinetic or for drug metabolism.

Solvation free energies in aqueous solution are important for structural and metabolic equilibria and kinetics. Unfavourable solvation free energies of associating nonpolar groups may lead to hydrophobic stabilization of their mutual complexes, although the electrostatic energy may not be so favourable. On the other hand, a favourable electrostatic energy does not always mean a high affinity, owing to the high solvation free energy penalty. Thus, the solvation free energy plays an important role in both protein-ligand association and protein folding [5].

## II. METHOD

In order to evidenciate the differences between two docking programs before their integration in a chemo-bioinformatic workflow, the use of different algorithms can help to refine a set of potential new lead compounds. Docking is one of the most commonly used techniques in Computer Aided Drug Design (CADD). Docking speed up the process of ligand selection. It allows to identify correct poses of a ligand in one specific region of a target protein. This approach is also used to estimate the strength of protein–ligand complexes. A massive screening of small molecules must be performed in order to obtain a reasonable small set of ligands suitable molecules for biological testing. A large variety of over 60

different docking programs with more of 30 scoring function have been proposed during the last two decades, for both commercial and academic use (DOCK [6], AutoDock [7], FlexX [8], Surflex [9], GOLD [10], ICM [11], Glide [12, 13], Cdocker [14], LigandFit [15], MCDock [16] and many others). Different strategies in the ligand placement are being used in each program, but all programs can be categorized into four broad categories: stochastic Monte Carlo, fragment-based, evolutionary-based, and shape complementary methods. A systematic search of all degrees of freedom in the ligand molecule is not used by any of these programs because of the enormous computational cost [17]. Two complementary components are the basis for each docking program. One is a method to explore the conformational space of the ligand and/or the protein target. The other is a scoring function to evaluate the proposed poses. During the docking procedure a large number of poses is generated, therefore fast and reliable function is required in order to estimate the strength of the interaction between the protein and the ligand. The best score should be assigned to the 'correct pose' by a scoring function that evaluates the stability of the target-ligand complexes.

The comparison of results obtained with different docking methods is difficult due to differences in docking algorithms and scoring functions procedures [18], so novel techniques (consensus scoring) have been applied [19] to integrate the information obtained using diverse software with the same targets and ligand libraries. In our analysis each docking algorithm acts as specialized sensor. We combined the two sensors in a data fusion system and compared the results from the two docking methods by application of a machine learning tool. Our analysis has been carried out on the PDB structure 2FJU. The sample set of ligands has been extract, taking the available information about inhibitors of RAC1 into account, from Zinc database [20]. The data set comprises 1488 small molecules belonging to three categories of compounds (Morpholines, Flavonoids and Imidazoles) that seem to have a particular activity respect to Rac1 protein. Solvation plays a critical role in ligand-target interaction. Not all the scoring functions allow to evaluate the influence of solvation on binding affinity. In order to verify the effect of the solvation energy on Rac1-ligand interactions we introduce the workflow in Figure 1. We used two docking program with different scoring function, Dock 6.2 and AutoDock 4. In particular with Dock 6.2 we do not take into account the solvation energy contribution, on the other hand we do it with Autodock4. This software embeds a semiempirical function to estimate solvation effect. This function is not the best available, but AutoDock is one of more common tools for docking [21]. Differences can depend on grid topology or selected scoring function. Dock 6.2 does not consider the solvation estimation in the default binding affinity calculation. The user must select specific options to include solvation in the binding affinity computation. We have obtained the energy score for both the programs, we proceed with a data cleaning operation to remove the outliers that should give problems in the analysis. In this procedure the interquartile range (IQR) [22] methods has been used.

After removal we have clustered the remaining data using a k-means algorithm. The obtained clusters have been validated with the Davies-Bouldin index (DBI), that is a metric for clustering algorithms evaluation [23]. DBI is expressed by the following equation:

$$DBI = \frac{1}{K}\sum_{k=1}^{K}\max_{k\neq l}\left\{\frac{S(U_k)+S(U_l)}{d(U_k,U_l)}\right\} \qquad (1)$$

Where $S(U_k)$ and $S(U_l)$ are the average distances between objects within clusters $U_k$ and $U_l$ respectively and $d(U_k, U_l)$ a between-cluster distance. For each class of ligands, ranked by using both algorithm, we have obtained a different $K$ value (data not show) resulting from optimization of DBI. After preliminary clustering we have selected, for each ligands class, the clusters that maximize the cardinality ($C^{Dock}$,$C^{Autodock}$), and we have used them to generate a new set of compound by intersection (Fig.2).

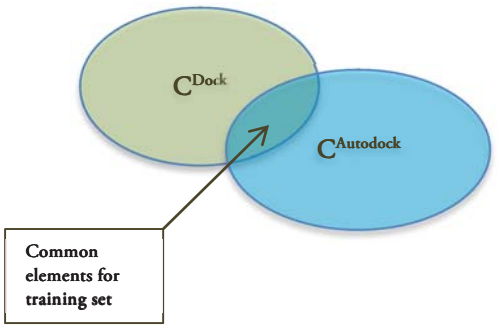


Fig. 2. Intersection of clusters for training set construction

This new data set is used as training set for a classification process and it is composed by 221 small molecules and in particular 38 Morpholines, 51 Flavonoids and 132 Imidazoles. This set is an unbalanced set because it reflects the query used to extract the ligand from the database. In fact we have intersected the sets for 'Rac1' and for each class of ligands. This situation has limited the bootstrap process. A small equally distributed set of ligands is too small for our purpose. Two Linear Discriminant Analysis (LDA) classifiers have been used to discriminate between the three classes of molecules using the obtained training set and 50 test set of about 30 molecules randomly selected from the whole set of molecules. The results of the two classification process have been used to assessing the reliability of agreement between the two classifiers using the Fleiss' Kappa method [24]. The kappa, $\kappa$, can be defined as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \qquad (2)$$

The factor $1-\overline{P}_e$ gives the degree of agreement that is attainable above chance, and $\overline{P}-\overline{P}_e$, gives the degree of agreement actually achieved above chance. The measure calculates the degree of agreement in classification over that which would be expected by chance and is scored as a number between 0 and 1.

## III. Results and Discussion

The work presented here used the data from two docking algorithms and especially one that considers the presence of solvation in the protein ligand interaction and another that does not. The purpose of this work was to identify a possible method to identify the possible difference in behaviour of the small molecules considered in relation to the relationship with the target RAC1. Applying the Fleiss' Kappa method we have obtained a substantial agreement between the two classifiers with a value of 0.6545. In the light of this result seems that about 35% of the considered small molecules are subjected to solvation effect. This set of molecules has been designed as test set for more refined MM analysis in order to estimate how chemical features of these ligands affects the interaction with specific residues of Rac1. Our analysis has been mainly focused to mine new chemical entities suitable for more refined computational analysis. Our resulting set of new molecules, influenced by solvation, is suitable for subsequent QSAR analysis taking this factor into account.

## References


[1] G. C. Terstappen and A. Reggiani, "In silico research in drug discovery," *Trends Pharmacol Sci,* vol. 22, pp. 23-6, Jan 2001.

[2] N. Ferri, *et al.*, "Virtual screening approach for the identification of new Rac1 inhibitors," *J Med Chem,* vol. 52, pp. 4087-90, Jul 23 2009.

[3] X. R. Bustelo, *et al.*, "GTP-binding proteins of the Rho/Rac family: regulation, effectors and functions in vivo," *Bioessays,* vol. 29, pp. 356-70, Apr 2007.

[4] W. Tan, *et al.*, "An essential role for Rac1 in endothelial cell function and vascular development," *FASEB J,* vol. 22, pp. 1829-38, Jun 2008.

[5] J. Wang, *et al.*, "Solvation Model Based on Weighted Solvent Accessible Surface Area," *The Journal of Physical Chemistry B,* vol. 105, pp. 5055-5067, 2001.

[6] T. J. Ewing, *et al.*, "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases," *J Comput Aided Mol Des,* vol. 15, pp. 411-28, May 2001.

[7] G. Morris, *et al.*, "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function," *J. Comput. Chem.,* vol. 19, pp. 1639-1662, 1998.

[8] M. Rarey, *et al.*, "A fast flexible docking method using an incremental construction algorithm," *J Mol Biol,* vol. 261, pp. 470-89, Aug 23 1996.

[9] A. N. Jain, "Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine," *J Med Chem,* vol. 46, pp. 499-511, Feb 13 2003.

[10] G. Jones, *et al.*, "Development and validation of a genetic algorithm for flexible docking," *J Mol Biol,* vol. 267, pp. 727-48, Apr 4 1997.

[11] R. Abagyan, *et al.*, "ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation," *Journal of Computational Chemistry,* vol. 15, pp. 488-506, 1994.

[12] R. A. Friesner, *et al.*, "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," *J Med Chem,* vol. 47, pp. 1739-49, Mar 25 2004.

[13] T. A. Halgren, *et al.*, "Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening," *J Med Chem,* vol. 47, pp. 1750-9, Mar 25 2004.

[14] G. Wu, *et al.*, "Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm," *J Comput Chem,* vol. 24, pp. 1549-62, Oct 2003.

[15] C. M. Venkatachalam, *et al.*, "LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites," *J Mol Graph Model,* vol. 21, pp. 289-307, Jan 2003.

[16] M. Liu and S. Wang, "MCDOCK: a Monte Carlo simulation approach to the molecular docking problem," *J Comput Aided Mol Des,* vol. 13, pp. 435-51, Sep 1999.

[17] N. Moitessier, *et al.*, "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go," *British Journal of Pharmacology,* vol. 153, pp. S7-S26-S7-S26, 2009.

[18] V. Mohan, *et al.*, "Docking: successes and challenges," *Curr Pharm Des,* vol. 11, pp. 323-33, 2005.

[19] D. Plewczynski, *et al.*, "VoteDock: Consensus docking method for prediction of protein-ligand interactions," *J Comput Chem,* vol. 32, pp. 568-81, Mar 2011.

[20] J. J. Irwin and B. K. Shoichet, "ZINC--a free database of commercially available compounds for virtual screening," *J Chem Inf Model,* vol. 45, pp. 177-82, Jan-Feb 2005.

[21] S. Y. Huang and X. Zou, "Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions," *J Chem Inf Model,* vol. 50, pp. 262-73, Feb 22 2010.

[22] J. W. Tukey, *Exploratory Data Analysis*: University Microfilms International, 1988.

[23] D. Davies and D. Bouldin, "A Cluster Separation Measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. PAMI-1, pp. 224-227, 1979.

[24] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin,* vol. 76, pp. 378-382, 1971.