Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Structural bioinformatics

Optimal water networks in protein cavities with GAsol and 3D-RISM

Lucia Fusani[1,2], Ian Wall[1], David Palmer[2], Alvaro Cortes[1*].

[1]GSK Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK.

[2]Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, UK.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Water molecules in protein binding sites play essential roles in biological processes. The popular 3D-RISM prediction method can calculate the solvent density distribution within minutes, but is difficult to convert it into explicit water molecules.

Results: We present *GAsol*, a tool that is capable of finding the network of water molecules that best fits a particular 3D-RISM density distribution in a fast and accurate manner and that outperforms other available tools by finding the globally optimal solution thanks to its genetic algorithm.

Availability: https://github.com/accsc/GAsol. BSD 3-clauses license

Contact: alvaro.x.cortes@gsk.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The function of the water molecules in the binding sites of the proteins has become of considerable interest recently. It is well known that water plays a key role in ligand recognition and in stabilizing protein structures. In order to complement experimental techniques and to improve our understanding of active-site hydration, several computational approaches have been developed during the years (Bodnarchuk, 2016). Some of the most popular methods to locate water molecules in protein binding sites include WaterMap (Abel, et al., 2008), GIST (Nguyen, et al., 2012) and the Three-Dimensional Reference Interaction Site Model (3D-RISM) (Beglov and Roux, 1997), to name a few. During this time, it has become clear that the water predicting tools can have a significant impact on medicinal chemistry programs. One recent example includes the development of inhibitors of platelet-derived growth factor receptor β (Horbert, et al., 2014) Of particular relevance here, 3D-RISM (Kovalenko and Hirata, 1999) is a computational approach that calculates the distribution of solvent molecules around a solute and which has its roots in statistical mechanical integral equation theories (IET) of liquids. Most popular 3D-RISM implementations can calculate the solvent distribution around a rigid solute within minutes, using only the solute structure and the solvent composition as input. To address the difficulty to convert the continuous distribution function of 3D-RISM into explicit water molecules some algorithms have been developed, e.g. Placevent (Sindhikara, et al., 2012), but either they present some deficiencies regarding finding a truly global solution or they cannot be applied easily to a wide range of targets. Here we present *GAsol*, a tool that is capable of finding the network of water molecules that best fits a particular 3D-RISM density distribution in a fast and accurate manner.

2 Methods and application

GAsol addresses the search for the optimal network of water molecules from a global point of view, by using a genetic algorithm and a desirability function based on the 3D-RISM density (Fig. 1A) that try to avoid the local minima problem (Fig. 1B, Supp. Table S1-S3). The analysis can be carried out typically in a couple of minutes on modern workstations due to the built-in multiprocessor capabilities, and only requires a grid file in DX format. The resulting network is written to a PDB file that can be visualized with any standard molecular viewer.

2.1 Detecting potential water sites

The number of water sites to consider in the optimization process is a critical parameter of the algorithm. We have implemented a double filter procedure that, first, uses a minimum threshold value for the density distribution to consider a grid point as a potential water site (by default $g(r) \geq 5$) and second, a spatial constraint in the form of a sphere with user supplied centre and radius, to consider only grid points inside the defined region (e.g. binding site). To facilitate this process, the program allows users to specify a ligand of interest in PDB format to automatically set the centre of the region to the geometrical centre of the molecule.

2.2 Genetic algorithm

After selecting the potential water sites, the algorithm initializes a population of individuals of potential solutions to the problem (chromosomes). Each chromosome is made of multiple genes, as many as water sites are available. Each gene is set to a value of 1, meaning the site is occupied by a water molecule or to 0, meaning that the site is empty. The initial population is then evolved during a total of 10,000 generations. In each generation, the population is subjected to selection, crossover and mutation. The selection procedure chooses individuals in the current generation with a tournament scheme. In this tournament, three individuals are selected randomly, allowing repetition, and only the best solution is allowed to reproduce. In the crossover phase, these individuals are mated by combining their chromosomes defining two random crossover points. Finally, in the mutation step, random gene flips are introduced in the offspring with a low probability to add variability.

2.3 Desirability function

Before the algorithm starts to generate solutions, the density distribution from the 3D-RISM calculation is transformed to a population function by using the equation $P(\vec{r}) = \rho_{bulk} V_{voxel} g(\vec{r})$ where ρ_{bulk} is the density of the bulk solvent, V_{voxel} is the volume of one voxel in the grid and g(r) is the density function. Following, for each water site detected in the first phase of the program, we calculate the minimum number of voxels required to account for one unit of the population. Each water site is then scored by dividing the final population value (which should be around 1.0) by the radius of the sphere calculated. This scoring method guarantees that water sites with more compact populations, and therefore more likely, are selected preferentially. To score individual solutions, we have introduced a desirability function with two subcomponents and one penalty term (Supp. Inf.). The first subcomponent accounts for the amount of population considered for a particular solution by summing all the individual values for each occupied water site and normalizing by the sum of the values for all water sites in the solution space (occupied or not). The second subcomponent tries to avoid double-counting the same part of the population multiple times in the case of proximal water sites. The function has a value of 1 by default except when two or more occupied water sites are at a distance of less than a threshold, which sets the value to 0. A penalty term has been introduced to improve the efficiency of the algorithm regarding the second subcomponent. As the desirability of the non-feasible solutions is always 0, the algorithm tends to waste several initial iterations since the random solutions usually contain several incompatible occupied water sites. The penalty term is defined then as the weighted ratio of the number of incompatible water sites and the total number of sites in the chromosome.

2.4 Evaluation datasets and results

To validate the tool we have selected a dataset of X-ray crystal ligand-proteins complexes with confirmed water networks that includes the HIV-1 protease (PDB 2ZYE), neuraminidase (PDB 1NNC), bovine pancreatic trypsin (PDB 5PTI) and a series of 184 BRD4 bromodomain 1 (BRD4-BD1) complexes to evaluate the robustness of the algorithm to small changes in the binding site (Supp.Inf. Table 1) in a highly conserved water network (Fig. 1C). As a metric, we have used the number of water molecules predicted within a distance of 2.0 A from the crystallographic position (Supp. Fig. S1). The tool can detect all the water molecules around the ligands in the HIV-1 protease, neuraminidase and in the bovine pancreatic trypsin systems. For the 184 BRD4-BD1 complexes, GAsol identifies correctly 94.3% (Supp. Fig. S1) of the key water molecules of the complexes with an improvement of the results of 90% if compared to a standard tool (Fig. 1D). Moreover, the number of false positive defined as the number of predicted water molecule not matching a crystallographic one is comparable between GAsol and placevent (Supp. Fig. S2).

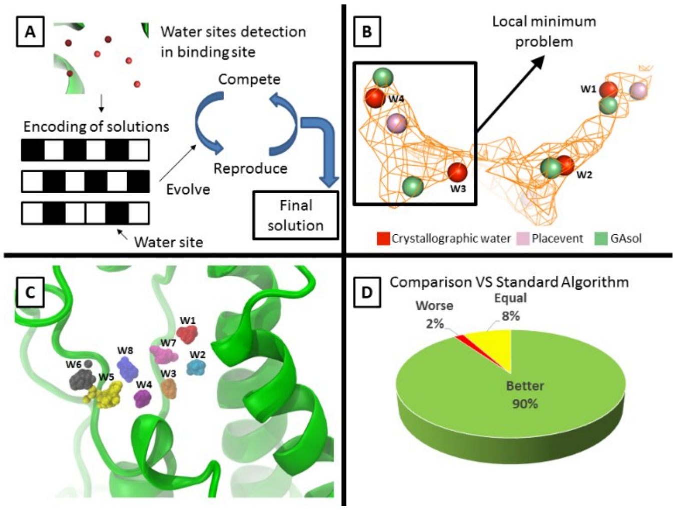

Figure 1 A) General overview of the GAsol algorithm. B) Example of misprediction of two water molecules (in pink) due to a local minima problem (PDB ID 5I80). The water predicted by GAsol are reported in green. C) Overlay of the highly-conserved water network in the 184 bromodomains BRD4-BD1. D) Results of the validation procedure on BRD4 crystals vs. placevent (standard algorithm).

References

Abel, R., *et al.* (2008) Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society,* 130, 2817-2831.

Beglov, D. and Roux, B. (1997) An integral equation to describe the solvation of polar molecules in liquid water. *The Journal of Physical Chemistry B,* 101, 7821-7826.

Bodnarchuk, M.S. (2016) Water, water, everywhere… It's time to stop and think. *Drug discovery today,* 21, 1139-1146.

Horbert, R., *et al.* (2014) Optimization of potent DFG-in inhibitors of platelet derived growth factor receptorβ (PDGF-Rβ) guided by water thermodynamics. *Journal of medicinal chemistry,* 58, 170-182.

Kovalenko, A. and Hirata, F. (1999) Potential of mean force between two molecular ions in a polar molecular solvent: a study by the three-dimensional reference interaction site model. *The Journal of Physical Chemistry B,* 103, 7942-7957.

Nguyen, C.N., *et al.* (2012) Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril. *The Journal of chemical physics,* 137, 044101.

Sindhikara, D.J., *et al.* (2012) Placevent: An algorithm for prediction of explicit solvent atom distribution—Application to HIV-1 protease and F-ATP synthase. *Journal of computational chemistry,* 33, 1536-1543.